



HAL
open science

Bandwidth extension tools for audio digital signals

Patrice Collen

► **To cite this version:**

Patrice Collen. Bandwidth extension tools for audio digital signals. domain_other. Télécom Paris-Tech, 2002. English. NNT: . pastel-00000512

HAL Id: pastel-00000512

<https://pastel.hal.science/pastel-00000512>

Submitted on 19 Jan 2004

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Ecole Nationale Supérieure des Télécommunications

THESE

pour l'obtention du titre de

DOCTEUR EN SCIENCES

Spécialité : Traitement du signal

TECHNIQUES D'ENRICHISSEMENT DE SPECTRE DES SIGNAUX AUDIONUMERIQUES

Patrice COLLEN

Soutenue le 14 novembre 2002 devant la commission d'examen

Xavier Rodet	IRCAM Paris	Président du Jury
Jean-Christophe Valière Jean Ménez	Université de Poitiers Université de Nice, Sophia-Antipolis	Rapporteur Rapporteur
Dominique Massaloux Gaël Richard Sylvain Marchand	France Télécom R&D ENST Paris Université de Bordeaux 1	Examinatrice Examinateur Examinateur
Nicolas Moreau Pierrick Philippe	ENST Paris France Télécom R&D	Directeur de thèse Directeur de thèse

RESUME

Techniques d'enrichissement de spectre des signaux audionumériques

Afin de limiter les dégradations liées au codage bas-débit des signaux audionumériques, la stratégie adoptée par la plupart des systèmes de compression de parole et de musique consiste à ne pas transmettre le contenu hautes-fréquences. C'est ainsi qu'aux environs des 20kbit/s, les codeurs de musique actuels ne restituent pas les sons avec leur qualité naturelle (leur bande passante étant limitée aux environs des 6kHz). Les sons ainsi codés/décodés deviennent ternes et perdent de leur qualité. On se propose d'étudier dans cette thèse de nouvelles techniques susceptibles de palier à cette perte des aigus. Les systèmes d'enrichissement de spectre permettent, avec très peu de données additionnelles, de rehausser la bande passante, et donc la qualité de ces signaux à bande-limitée. Le principe de ces techniques consiste à exploiter les informations comprises dans le spectre basse-fréquence afin de synthétiser le signal pleine-bande de qualité proche de celle de l'original.

Dans le cadre d'un contrat financé par France Télécom R&D, l'objectif de cette thèse est la réalisation d'un système d'enrichissement de spectre des signaux audionumériques (parole et musique). La technique PAT (Perceptual Audio Transposition) implémentée a fait l'objet de deux propositions de normalisation dans les instances DRM (Digital Radio Mondiale) et MPEG-4 (Moving Picture Experts Group).

Le document est structuré en 4 parties. La première partie s'attache à introduire les principes de l'extension de bande en se fondant sur les propriétés psychoacoustiques et les caractéristiques des signaux audio mis en jeu. Grâce à cette étude préalable, l'enrichissement des signaux sonores est réalisé en deux étapes : une étape d'extension de la structure fine du spectre et une étape d'ajustement de l'enveloppe, qui font l'objet des deux chapitres suivants. Ainsi, la seconde partie est consacrée aux techniques d'estimation, de transmission et d'ajustement d'enveloppe spectrale. Deux techniques particulières sont développées : L'une basée sur la prédiction linéaire et l'autre sur la modélisation d'enveloppe par facteurs d'échelle dans le domaine fréquentiel. Dans la troisième partie, les différentes solutions permettant d'étendre la structure fine spectrale sont abordées. L'étude s'est portée notamment sur les translations de spectre dans le domaine fréquentiel et sur les distorsions non-linéaires. Enfin, en quatrième partie, on présente un schéma complet d'enrichissement de spectre avant d'en évaluer ses performances dans le cadre de la normalisation MPEG-4.

Une toute nouvelle technique de compression des signaux audionumériques est ainsi introduite dans cette thèse. Celle-ci a montré un réel intérêt dans le domaine de la compression du son. Pour une qualité équivalente, la réduction de débit obtenue est de l'ordre de 25%.

Mot clés :

Codage, audio, compression du son, régénération des hautes fréquences, bas-débit, parole et musique, enveloppe spectrale, translation spectrale, distorsion non linéaire

ABSTRACT

Bandwidth extension tools for audio digital signals

To maintain a reasonable perceived quality and to reduce degradations, classical audio or speech source coding algorithms need to limit the audio bandwidth and to operate at low sampling rates. For a data rate of 20 kbit/s for instance, the bandwidth of audio signals doesn't exceed 6kHz with classical audio coders. The signals produced suffer from some quality degradation due to the lack of high energy components.

To overcome this problem, new methods for improving the quality of bandlimited signal are proposed in this document. With the use of a little more transmitted information, bandwidth extension tools allow the recovery of spectral highband components and thus enhance the quality of such bandlimited signals.

The method exploits signal redundancy in the spectral domain and uses lowband components to synthesise the fullband signal.

This thesis was completed through a contract with France Telecom R&D. The project aim was to design an effective and low bitrate bandwidth extension tool.

The PAT (Perceptual Audio Transposition) technology produced during these 3 years was proposed in DRM (Digital Radio Mondiale) and MPEG-4 (Moving Picture Experts Group) consortiums.

This document is split into four parts.

Based on psychoacoustics properties and characteristics of treated audio signals, the first part introduces bandwidth extension tools. According to these considerations, a first bandwidth extension scheme is introduced. The process of bandwidth extension can be divided in two independent tasks: the extension of the frequency components (high frequency regeneration) and the highband spectral envelope estimation.

The second part is dedicated to estimation, coding and adjustment of the spectral envelope. Two particular techniques are considered: The first method is based on linear prediction, and the second method consists of modelling the spectral envelope in the frequency domain.

The third part examines in detail several techniques for high frequency regeneration. In particular, spectral translations in the frequency domain and non-linear distortions are developed.

Finally, in the fourth part, a new bandwidth extension scheme is proposed. Subjective tests evaluate the performances of this technique in the context of MPEG-4 normalisation.

A new method for highband audio compression techniques is introduced in this document. For same quality, subjective tests demonstrate that the bitrate reduction is about 25%. This new bandwidth extension tool demonstrates high performance in audio coding.

Key words

Coding, audio compression, HFR Techniques (High Frequency Regeneration), low bitrate, speech and music, spectral envelope, spectral translation, non-linear distortion

REMERCIEMENTS

Mes remerciements et ma gratitude se portent tout d'abord vers Pierrick Philippe qui m'a encadré et guidé quotidiennement pendant ma thèse et a su m'orienter vers les axes les plus pertinents. Je le remercie pour ses compétences, son ouverture d'esprit et sa grande disponibilité. Toutes les recherches menées pendant ces trois années se sont appuyées sur son expertise qui a guidé nombre de mes choix et conclusions.

Pierrick a de plus contribué fortement à l'implémentation de la technique PAT, ainsi qu'à la rédaction et à l'amélioration de ce document de thèse.

Mes pensées se tournent ensuite tout naturellement vers Jean-Bernard Rault qui m'a co-encadré pendant toute une année et qui a également contribué à l'implémentation du codeur final. Toujours à l'écoute, il a su répondre à mes interrogations, et ce, avec grand intérêt.

Je suis particulièrement reconnaissant du souci de Pierrick et de Jean-Bernard de m'avoir permis de mener mes recherches dans les meilleures conditions, et ce toujours dans la bonne humeur et dans la jovialité. Ils ont tous deux été de formidables encadrants, tant du point de vue scientifique qu'humain, et ce fut un réel bonheur de travailler avec eux pendant ces trois trop courtes années.

Je remercie Nicolas Moreau, mon directeur de thèse, de m'avoir fait confiance dans la réalisation de ce travail.

Bien sûr, je remercie chaleureusement les membres du jury de l'intérêt qu'ils ont clairement manifesté pour ce travail, et des remarques et corrections qu'ils ont apportées à ce document:

Xavier Rodet m'a fait un grand honneur en acceptant de présider le jury de cette thèse.

J'exprime toute ma gratitude à Jean Ménez et à Jean-christophe Valière pour avoir accepté d'être les rapporteurs de ce mémoire.

Enfin, je tiens à exprimer ma reconnaissance à Dominique Massaloux, Gaël Richard et Sylvain Marchand pour leur contribution et la pertinence de leurs remarques.

Je remercie Vincent Marcaté, chef du laboratoire HDM et Henri Sanson, chef de l'URD SIM (Codage et indexation multimedia) de m'avoir accueillis dans le laboratoire HDM/DIH.

Je remercie tous ceux qui ont collaboré à la réalisation de la technique d'extension de bande pendant ces trois années de thèse. En particulier Balazs Kovesi pour son aide précieuse sur la quantification des paramètres LPC. Je n'oublie évidemment pas toutes les personnes de l'équipe audio, à savoir Cathy Colonnès, Jean-christophe Rault, Pierre Urcun, Mathieu Carré, Guillaume Fayemendy et Mathieu Lagrange, qui ont toujours été présents pour participer aux nombreux tests d'écoutes.

Merci à Sarah Hudson pour ses compétences linguistiques.

Je finirai en remerciant toutes les personnes que j'ai pu rencontrer durant ces trois années de thèse à France Télécom R&D de Lannion et de Rennes.

TABLE DES MATIÈRES

CHAPITRE 1

INTRODUCTION AUX TECHNIQUES D'ENRICHISSEMENT DE SPECTRE 21

1.1.	INTRODUCTION.....	22
1.2.	EXEMPLE DE MISE EN ŒUVRE	23
1.3.	CADRE ET CONTEXTE NORMATIF DE LA THESE.....	24
1.4.	PRINCIPES DES METHODES D'EXTENSION DE BANDE.....	25
1.5.	OBJECTIFS DE LA THESE	27
1.6.	PLAN DU DOCUMENT.....	28
1.7.	BIBLIOGRAPHIE DU CHAPITRE 1	29

CHAPITRE 2

PROPRIETES ET PERCEPTION DES SIGNAUX DE PAROLE ET DE MUSIQUE..... 31

2.1.	INTRODUCTION.....	32
2.2.	PROPRIETES DE L'OREILLE HUMAINE.....	33
2.2.1.	<i>Sensibilité de l'oreille humaine</i>	<i>33</i>
2.2.2.	<i>Bandes critiques</i>	<i>33</i>
2.2.3.	<i>Phénomènes psychoacoustiques de masquage.....</i>	<i>34</i>
2.2.4.	<i>Dissonance et JND (Just Noticeable Difference).....</i>	<i>35</i>
2.3.	PROPRIETES DES SIGNAUX AUDIONUMERIQUES.....	37
2.3.1.	<i>Signaux de parole</i>	<i>37</i>
2.3.2.	<i>Signaux musicaux</i>	<i>38</i>
2.3.3.	<i>Classes de signaux et propriétés haute-fréquence</i>	<i>39</i>
2.4.	CONCLUSION DU CHAPITRE.....	46
2.5.	BIBLIOGRAPHIE DU CHAPITRE 2.....	47

CHAPITRE 3

MODELISATION DE L'ENVELOPPE SPECTRALE 49

3.1.	INTRODUCTION.....	50
3.1.1.	<i>Remarques préalables sur la notion d'enveloppe.....</i>	<i>51</i>
3.2.	TECHNIQUES D'EXTRAPOLATION D'ENVELOPPE SPECTRALE.....	52
3.2.1.	<i>Principe</i>	<i>52</i>
3.2.2.	<i>Etat de l'art.....</i>	<i>52</i>
3.2.3.	<i>Extrapolation d'enveloppe par dictionnaires de formes d'ondes.....</i>	<i>52</i>
3.2.4.	<i>Application aux signaux de musique</i>	<i>53</i>
3.2.5.	<i>Conclusions</i>	<i>54</i>
3.3.	MODELISATION D'ENVELOPPE PAR PREDICTION LINEAIRE	55
3.3.1.	<i>Principe de la prédiction linéaire.....</i>	<i>55</i>
3.3.2.	<i>Représentation des coefficients de prédiction</i>	<i>59</i>
3.3.3.	<i>Ajustement d'enveloppe</i>	<i>60</i>
3.3.4.	<i>Applications à l'enrichissement de spectre.....</i>	<i>60</i>
3.3.5.	<i>Mise en forme spectrale du signal haute-fréquence.....</i>	<i>63</i>
3.3.6.	<i>Conclusion sur la modélisation d'enveloppe par prédiction linéaire.....</i>	<i>65</i>
3.4.	ESTIMATION D'ENVELOPPE DANS LE DOMAINE FREQUENTIEL	66
3.4.1.	<i>Principe</i>	<i>66</i>
3.4.2.	<i>Avantages de l'estimation d'enveloppe dans le domaine fréquentiel.....</i>	<i>67</i>
3.4.3.	<i>Ajustements en fonction de la nature des signaux.....</i>	<i>67</i>

3.4.4.	<i>Quantification et coût de transmission des facteurs d'échelle</i>	68
3.4.5.	<i>Conclusion</i>	69
3.5.	CONCLUSION DU CHAPITRE	70
3.6.	BIBLIOGRAPHIE DU CHAPITRE 3	71

CHAPITRE 4

TECHNIQUES D'EXTENSION DE LA STRUCTURE FINE SPECTRALE73

4.1.	INTRODUCTION	74
4.2.	TECHNIQUE D'EXTRAPOLATION DE SPECTRE SANS TRANSMISSION D'INFORMATION	75
4.2.1.	<i>Principe</i>	75
4.2.2.	<i>Résultats</i>	75
4.3.	TECHNIQUES PARAMETRIQUES	76
4.3.1.	<i>Introduction</i>	76
4.3.2.	<i>Synthèse des hautes fréquences sur les signaux de parole</i>	76
4.3.3.	<i>Synthèse des hautes fréquences sur les signaux musicaux</i>	77
4.3.4.	<i>Problèmes liés à l'approche paramétrique sur les signaux musicaux</i>	78
4.3.5.	<i>Conclusion sur l'approche paramétrique</i>	78
4.4.	TECHNIQUES DE TRANSLATIONS SPECTRALES.....	79
4.4.1.	<i>Translations spectrales réalisées dans le domaine temporel</i>	79
4.4.2.	<i>Translations spectrales mettant en oeuvre des transformées temps/fréquences</i>	81
4.5.	ENRICHISSEMENT DE SPECTRE PAR DISTORSION NON-LINEAIRE	86
4.5.1.	<i>Introduction</i>	86
4.5.2.	<i>Etude des distorsions non-linéaires sur les signaux transitoires</i>	87
4.5.3.	<i>Etat de l'art</i>	88
4.5.4.	<i>Etude des non-linéarités sur les signaux de parole</i>	89
4.5.5.	<i>Etude sur les signaux plus complexes</i>	94
4.5.6.	<i>Conclusions sur les distorsions non linéaires</i>	96
4.6.	CONCLUSION DU CHAPITRE	98
4.7.	BIBLIOGRAPHIE DU CHAPITRE 4.....	99

CHAPITRE 5

TECHNIQUES COMPLETES D'ELARGISSEMENT DE BANDE.....101

5.1.	INTRODUCTION	102
5.2.	TECHNIQUE PAT (PERCEPTUAL AUDIO TRANSPOSITION)	103
5.2.1.	<i>Introduction</i>	103
5.2.2.	<i>Détection de signaux transitoires</i>	105
5.2.3.	<i>Détection d'harmonicité</i>	108
5.2.4.	<i>Module d'extension de la structure fine</i>	110
5.2.5.	<i>Module d'estimation et d'ajustement d'enveloppe</i>	116
5.3.	TECHNIQUE SBR (SPECTRAL BAND REPLICATION)	128
5.3.1.	<i>Principe</i>	128
5.3.2.	<i>Conclusions</i>	129
5.4.	RESULTATS DES TESTS MPEG-4	131
5.4.1.	<i>Technique PAT associée au codeur de parole ITU G-729</i>	131
5.4.2.	<i>Technique PAT adaptée au codeur de musique MPEG-4 AAC</i>	132
5.4.3.	<i>Conclusion</i>	133
5.5.	CONCLUSION DU CHAPITRE	134
5.6.	BIBLIOGRAPHIE DU CHAPITRE 5	135

CHAPITRE 6	
CONCLUSIONS ET PERSPECTIVES	137
6.1. BIBLIOGRAPHIE DU CHAPITRE.....	141
CHAPITRE 7	
ANNEXES	143
TRANSFORMEES TEMPS/FREQUENCE	144
CONVERSION LPC/LSP.....	146
CONVERSION LSP-LPC	148
SYNTAXE DU TRAIN BINAIRE PAT	149

TABLE DES ILLUSTRATIONS

Figure 1.1 : Principe de l'enrichissement de spectre.....	23
Figure 1.2 : Système d'enrichissement de spectre	23
Figure 1.3 : Spectre d'un signal harmonique de 720 Hz de fréquence fondamentale.....	25
Figure 1.4 : Reconstruction du spectre par transmission d'enveloppe.....	26
Figure 2.1 : Seuil d'audition absolu	33
Figure 2.2 : Bandes critiques	34
Figure 2.3 : Masquage fréquentiel et courbes de masquage associées	35
Figure 2.4 : Différents types de masquages temporels.....	35
Figure 2.5 : Distinction parole voisée / non-voisée	38
Figure 2.6 : Snoise.....	40
Figure 2.7 : $S_{\text{mono_harm}}$	41
Figure 2.8 : $S_{\text{Multi_harm}}$	42
Figure 2.9 : S_{Inharm}	43
Figure 2.10 : S_{Trans}	44
Figure 3.1 : Estimation, transmission et ajustement d'enveloppe spectrale.....	50
Figure 3.2 : Signal de parole et enveloppes associées.....	51
Figure 3.3 : Extrapolation d'enveloppe par dictionnaire de forme d'onde.....	52
Figure 3.4 : Spectrogramme d'un signal composé d'un piano et de cymbale	53
Figure 3.5 : Filtre d'enveloppe pour deux trames de 20 ms à $t=0.8$ et $t=0.9s$	53
Figure 3.6 : Système d'entrée/sortie	55
Figure 3.7 : Filtre blanchisseur et filtre de synthèse.....	56
Figure 3.8 : Quantification SVQ (Split Vector Quantization).....	59
Figure 3.9 : Estimation d'enveloppe sur $S_{\text{Mono_harm}}$	61
Figure 3.10 : Prédiction linéaire sur un signal passe-bande	62
Figure 3.11 : Nécessité de blanchir le signal d'excitation avant la remise en forme spectrale.....	64
Figure 3.12 : Ajustement d'enveloppe par facteurs d'échelle dans le domaine fréquentiel	66
Figure 3.13 : Ajustement d'enveloppe par MDCT d'un signal transitoire.....	68
Figure 4.1 : Extension de la structure fine dans la méthode complète d'enrichissement de spectre.....	74
Figure 4.2 : Extrapolation linéaire dans le domaine fréquentiel.....	75
Figure 4.3 : Modèle paramétrique de production de la parole.....	76
Figure 4.4 : Extension de bande basée sur une approche paramétrique	77
Figure 4.5 : Translation par repliement spectral.....	79
Figure 4.6 : Extension de bande par repliement spectral.....	80
Figure 4.7 : Modulation numérique.....	80
Figure 4.8 : Translation de spectre	81
Figure 4.9 : Translations de spectre par banc de filtres.....	81
Figure 4.10 : Translations d'un peigne harmonique	82
Figure 4.11 : Translation par MDCT d'un signal transitoire	83
Figure 4.12 : Translation par MDCT d'un signal harmonique	84
Figure 4.13 : Effets des non-linéarités sur des signaux transitoires	87
Figure 4.14 : Codage RELP.....	88
Figure 4.15 : Fonction non linéaire de type W.....	89
Figure 4.16 : Comportement de la fonction $\ln(I+x)$ sur un sinus	90
Figure 4.17 : Synthèse de bruit haute-fréquence par non-linéarité.....	91
Figure 4.18 : Densités de probabilité des bruits synthétisés par non-linéarité	91
Figure 4.19 : Effets des non-linéarités sur un peigne de sinus	92
Figure 4.20 : Effets des non-linéarités sur un signal harmonique bruité.....	93
Figure 4.21 : Phénomènes d'intermodulation	94
Figure 4.22 : Limitation des phénomènes d'intermodulation	95
Figure 4.23 : Effets des non-linéarités sur deux peignes de sinus	96

Figure 5.1 : Diagramme de fonctionnement du codeur/décodeur PAT	103
Figure 5.2 : S_{Trans}	105
Figure 5.3 : Détection d'attaque sur un signal de clochette.....	106
Figure 5.4 : Agencement des fenêtres lors d'un changement de taille de fenêtre	107
Figure 5.5 : Estimation de l'harmonicité d'un signal.....	109
Figure 5.6 : Translations spectrales avec et sans retournement de spectre	110
Figure 5.7 : Spectre d'une trame de parole chantée voisée (Suzanne Vega).....	111
Figure 5.8 : Rupture d'harmonicité	112
Figure 5.9 : Perception des ruptures d'inharmonicité.....	112
Figure 5.10 : Corrélations hautes-basses fréquences sur un signal de parole et de musique	114
Figure 5.11 : Module d'extension de la structure fine.....	114
Figure 5.12 : Diagramme de fonctionnement du codeur par prédiction linéaire	117
Figure 5.13 : Diagramme de fonctionnement du décodeur.....	119
Figure 5.14 : Fonctionnement du codeur par facteurs d'échelle en sous-bandes	121
Figure 5.15 : Débit variable du PAT sur la séquence es01.wav	122
Figure 5.16 : Fonctionnement du décodeur par facteurs d'échelle en sous-bandes.....	123
Figure 5.17 : Spectre MDCT et DFT d'un sinus à 3333 Hz.....	124
Figure 5.18 : Signaux résultants de l'ajustement d'énergie par MDCT et par DFT	125
Figure 5.19 : Comparaison entre les deux techniques de modélisation d'enveloppe spectrale.....	126
Figure 5.20 : Diagramme de fonctionnement du codeur SBR.....	128
Figure 5.21 : Diagramme de fonctionnement du décodeur SBR	129
Figure 5.22 : Débit variable du SBR sur la séquence es01.wav	130
Figure 5.23 : Tests comparatifs entre le CELP (24 kbit/s) et le G-729+PAT (13,8 kbit/s)	131
Figure 5.24 : Tests Mono	132
Figure 5.25 : Tests Stéréo	133

TABLE DES TABLEAUX

Tableau 1.1 : Bande passante du codeur AAC	22
Tableau 1.2 : Calendrier des différentes phases de normalisation.....	24
Tableau 5.1 : Débit associé aux différents modes de fonctionnement	118
Tableau 5.2 : Découpage fréquentiel en fonction de la fréquence de coupure.....	122
Tableau 5.3 : Echelle de tests	131
Tableau 5.4 : Systèmes testés	132
Tableau 5.5 : Séquences testées.....	132

LISTE DES ABREVIATIONS

Traitement du signal

AM	Amplitude Modulation
AR	(Filtre) AutoRégressif
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
FM	Frequency Modulation
LAR	Log-Area Ratio
LSF	Line Spectral Frequency
MDCT	Modified Discrete Cosine Transform
SNR	Signal To Noise Ratio
SVQ	Split Vector Quantization
TDAC	Time Domain Aliasing Cancellation
VQ	Vector Quantization

Algorithmes de codage

AAC	Advanced Audio Coding
CELP	Code-Excited Linear Prediction
HILN	Harmonic and Individual Lines plus Noise
HVXC	Harmonic Vector eXcitation Coding
LPC	Linear Predictive Coding
MBE	Multi-band excited
PAT	Perceptual Audio Transposition
RELp	Residual Excited Linear Prediction
SBR	Spectral Band Replication
STC	Sinusoidal Transform Coding

Organisations de standardisation

DRM	Digital Radio Mondiale
ISO	International Standard Organisation
ITU	International Telecommunications Union
MPEG	Moving Picture Experts Group

Divers

DCR	Degradation Category Rating
SPL	Sound Pressure Level

LEXIQUE

DRM	Consortium "Digital Radio Mondiale" (DRM) dédié à la normalisation de la radio numérique en bande AM (Modulation d'Amplitude)
MPEG	Moving Picture Experts Group, groupe de travail de l'ISO (International Standard Organization) chargé du développement de normes internationales consacrées aux domaines audiovisuels

CHAPITRE 1

INTRODUCTION AUX TECHNIQUES D'ENRICHISSEMENT DE SPECTRE

Plan du chapitre

1.1.	INTRODUCTION.....	22
1.2.	EXEMPLE DE MISE EN ŒUVRE	23
1.3.	CADRE ET CONTEXTE NORMATIF DE LA THESE.....	24
1.4.	PRINCIPES DES METHODES D'EXTENSION DE BANDE.....	25
1.5.	OBJECTIFS DE LA THESE.....	27
1.6.	PLAN DU DOCUMENT.....	28
1.7.	BIBLIOGRAPHIE DU CHAPITRE 1	29

1.1. Introduction

L'oreille humaine perçoit les sons dans une bande de fréquence comprise entre 20 Hz et 20 kHz, mais transmettre des sons ayant une telle bande est parfois irréalisable vue la nature des moyens de transmission. Les signaux en bande téléphonique ont par exemple une bande passante limitée en basse-fréquence à 300 Hz et qui n'excède pas 3.4 kHz en haute-fréquence. Cette perte des aigus ôte tout son côté naturel à la voix et dégrade la qualité audionumérique des conversations téléphoniques.

En matière de codage audio-numérique, la transmission et le stockage des signaux pleine-bande sont également très coûteux en terme de débit. Afin de limiter les distorsions à bas débit, la stratégie consiste alors généralement à ne transmettre que la partie basse-fréquence des signaux, c'est-à-dire à supprimer les aigus. Les sons ainsi codés/décodés deviennent ternes et perdent de leur qualité.

Prenons l'exemple du codeur AAC (Advanced Audio Coding) de la norme ISO MPEG-4¹, décrit dans [BOS 97]. La stratégie de codage consiste à coder et à transmettre les informations dans le domaine fréquentiel. Ainsi, selon le débit alloué et selon la qualité de restitution souhaitée, le codeur/décodeur AAC génère des signaux de bande-passante variable.

La Tableau 1.1 donne cette bande-passante en fonction du débit requis pour coder des signaux monophoniques.

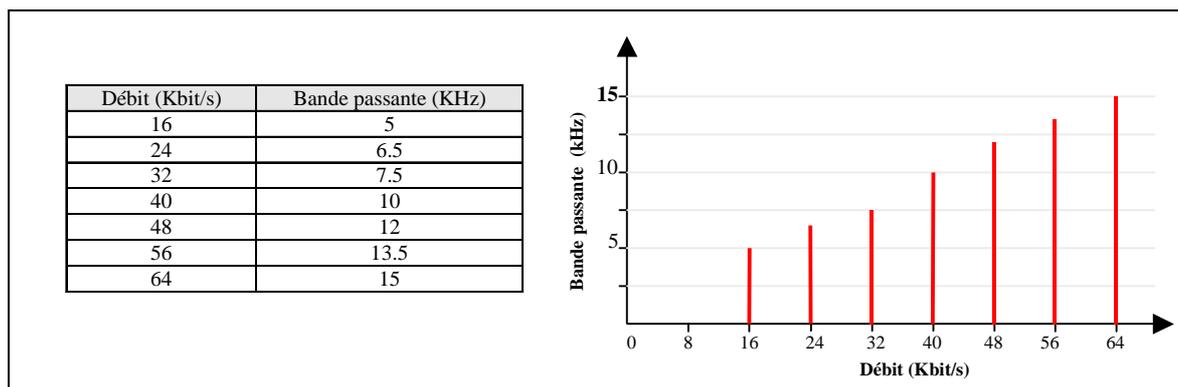


Tableau 1.1 : Bande passante du codeur AAC

C'est ainsi qu'à bas-débit, aux environs des 20 kbit/s, la bande passante n'excède pas 6 kHz pour une qualité de restitution acceptable.

Pour coder un signal monophonique en bande FM ([0-16kHz]) de façon transparente², le codeur AAC requiert un débit de 64 kbit/s. A ce débit, le codeur utilise en moyenne 28 kbit/s, soit plus de 40% du débit, pour transmettre la bande comprise entre 6 et 15 kHz.

Toutes ces considérations ont amené les chercheurs à développer de nouvelles stratégies de codage susceptibles d'étendre la bande passante des signaux à bande limitée sans, ou avec très peu de données auxiliaires. L'idée forte consiste à utiliser les informations comprises dans le spectre basse-fréquence afin de synthétiser le signal pleine-bande de qualité proche de celle du signal original. Ce principe est illustré sur la Figure 1.1.

¹ Cf lexique

² Un codage est dit transparent lorsque des auditeurs experts n'entendent pas de différence entre le signal reconstruit et le signal original

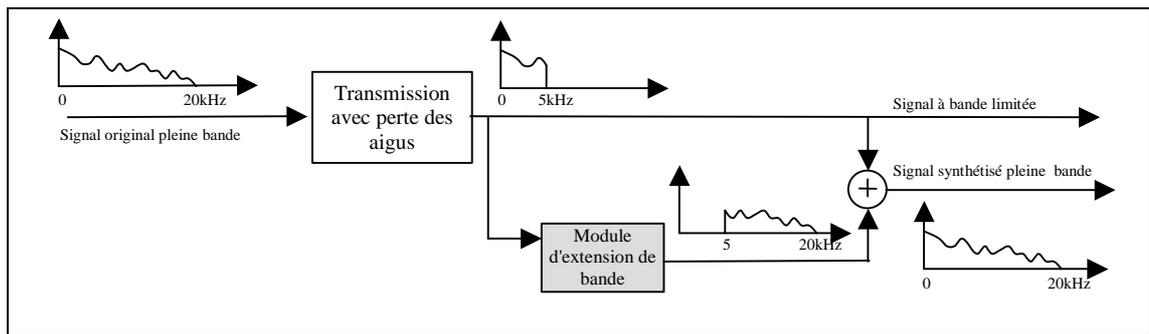


Figure 1.1 : Principe de l'enrichissement de spectre

Les applications de ces techniques d'enrichissement de spectre sont multiples, puisque pour un même débit transmis la qualité perçue sera améliorée, le son devenant "plus clair" car plus riche en aigus; le corollaire étant que le débit peut être diminué pour une qualité équivalente. Les applications sont donc toute transmission de signaux sur réseaux, qu'ils soient hertziens ou câblés, et ceci quelle que soit la nature de ces signaux : sons musicaux, parole ou autre. Ces techniques touchent ainsi à de nombreux domaines d'application, tels que la téléphonie fixe, la téléphonie mobile, le codage bas débit sur Internet, le stockage de fichiers sons, et toute diffusion audio en général.

1.2. Exemple de mise en œuvre

Les techniques d'enrichissement de spectre étudiées dans ce document proposent une nouvelle approche basée sur l'utilisation de deux modules indépendants :

- Le premier, le module cœur, est un codeur/décodeur audio "classique", qui peut être de parole ou musical.
- Le second, le module d'enrichissement de spectre, qui vise à reconstituer les aigus non transmis par ce premier.

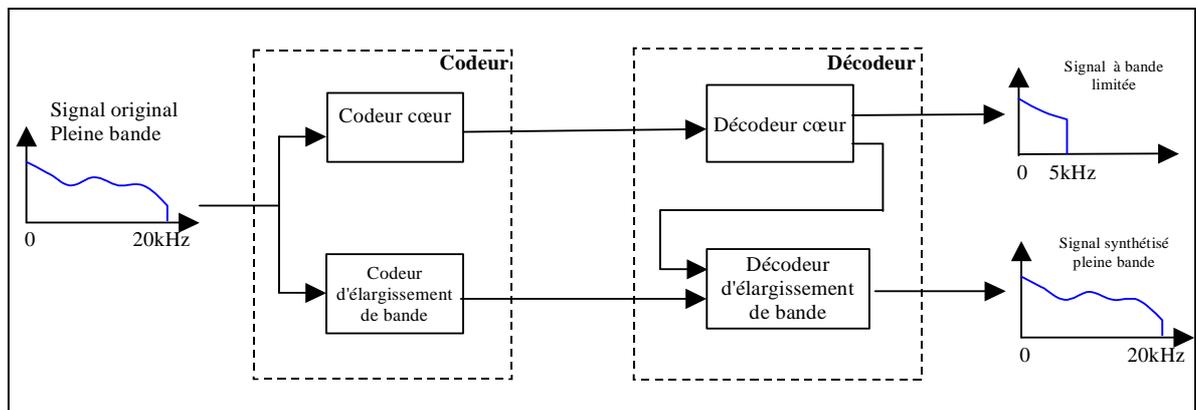


Figure 1.2 : Système d'enrichissement de spectre

Le codeur/décodeur cœur utilisé peut être de nature multiple. Il est possible en particulier de considérer des signaux non codés, qui ont subi une limitation de bande par sous-échantillonnage par exemple.

Il peut également s'agir d'un codeur par transformée de type ISO/MPEG-1 ([ISO 92]), MPEG-2 ([ISO 94]) ou MPEG-4 ([ISO 01]) ou de type CELP ([ATA 82]) ou même paramétrique (codeur paramétrique HILN [PUR 00]).

1.3. Cadre et contexte normatif de la thèse

Les techniques d'enrichissement de spectre ont suscité de nombreuses recherches depuis une vingtaine d'années dans le domaine du codage de parole. Les signaux de parole étant basés sur un modèle de production bien connu, il semble en effet assez aisé de synthétiser les hautes fréquences de tels signaux à partir de la connaissance seule du signal à bande-limitée. Mais comme nous le verrons dans la suite de ce document, les techniques développées offrent malgré tout des résultats peu concluants. L'extrapolation de la bande haute des signaux de parole à partir de la bande basse semble délicate sans transmission de données auxiliaires.

On comprend dès lors qu'un tel système applicable à tout type de signal audio, qu'il soit de parole ou de musique, devient des plus complexe à mettre en oeuvre. Au vu de la diversité des sons, il semble en effet difficile de trouver un modèle générique capable d'extrapoler la bande passante tout en conservant une qualité d'écoute proche de celle du signal original. La littérature concernant ce sujet reste aujourd'hui relativement pauvre bien que quelques techniques commencent à voir le jour.

La première technique d'extension de bande applicable sur les signaux audio générique, nommée SBR (Spectral Band Replication), est apparue en 1999 dans le cadre de la normalisation dans le projet DRM³. La technique SBR fait aujourd'hui référence dans le domaine. Elle est notamment utilisée dans le MP3-Pro [EKS 02] et en cours de normalisation dans MPEG-4.

Afin de répondre aux échéances du processus de normalisation dans DRM, une technique alternative et concurrente, le PAT (Perceptual Audio Transposition), a été développée à partir de 1999 dans les laboratoires de France Télécom R&D.

Les différentes versions du PAT ont ainsi fait l'objet de deux propositions face au SBR :

- Dans le projet DRM en mars 2000
- Dans le projet MPEG-4 en juillet 2000 [ISO 00a]

Le calendrier de cette thèse, présenté Tableau 1.2, a ainsi été régi par les différentes phases de normalisation de ces deux projets durant ces trois années.

Normalisation dans DRM		
Phase de normalisation	Date	Document
Etudes	Mars 2000	[DRM 00a]
Propositions	Avril 2000	[DRM 00b]
Résultats des tests	Décembre 2000	[DRM 00c] et [DRM 00d]
Norme	Septembre 2001	[DRM 01]
Normalisation dans MPEG-4		
Phase de normalisation	Date	Document
Etudes	Janvier 01	[ISO 00b] et [ISO 01a]
Propositions	Juillet 2001	[ISO 01b] et [ISO 01c]
Résultats des tests	Décembre 2001	[ISO 01d]
Norme	Mars 2002	[ISO 02]

Tableau 1.2 : Calendrier des différentes phases de normalisation

Notons qu'une phase de normalisation se déroule en trois étapes dans MPEG-4:

- L'étape du "Call for evidence" (phase d'étude) qui vise à montrer les avantages apportés, en terme de débit et/ou de qualité, par l'introduction d'une nouvelle technique par rapport aux techniques déjà présentes dans la norme.

³ Cf lexique

- L'étape du "Call for proposal" (phase de proposition) pendant laquelle différentes alternatives concurrentes sont proposées.
- L'étape de test qui permet de statuer sur la meilleure technique à retenir pour la constitution du standard.

Ces trois étapes s'échelonnent sur plusieurs années et aboutissent progressivement à l'établissement d'une norme.

Dans le cadre de la normalisation dans les projets DRM et MPEG-4, deux types de codeurs ont été testés et implémentés sur la technique d'extension de bande développée pendant ces trois années de thèse. Le codeur de parole ITU G729 [UIT 96] et le codeur MPEG-4 AAC.

1.4. Principes des méthodes d'extension de bande

Un des éléments importants du son en général est la notion de timbre. Le timbre est la caractéristique permettant de qualifier la "richesse" d'un son. Il est spécifique de chaque instrument et reste l'élément le plus difficile à appréhender.

Le premier élément permettant de définir le timbre d'un instrument est la présence d'harmoniques. En substance, tout son périodique entretenu est composé d'une fréquence fondamentale, et d'une série d'harmoniques dont les fréquences sont multiples de cette fondamentale (Figure 1.3).

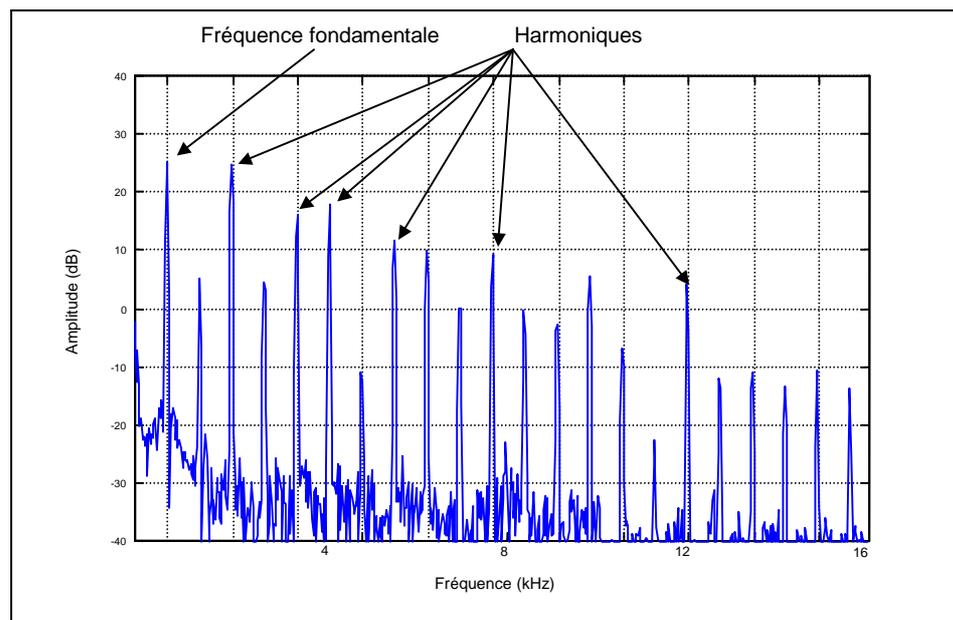


Figure 1.3 : Spectre d'un signal harmonique de 720 Hz de fréquence fondamentale

Ce qui différencie deux sons périodiques entretenus de même fréquence fondamentale et de même énergie est l'enveloppe spectrale, c'est-à-dire principalement l'amplitude des harmoniques sur tout le spectre.

Le deuxième élément permettant de définir le timbre d'un son est l'enveloppe temporelle du signal, c'est à dire l'évolution de l'énergie du signal au cours du temps. On peut repérer typiquement dans l'enveloppe temporelle trois phases :

- L'attaque qui correspond à l'apparition du son
- Le maintien, période pendant laquelle le signal peut être considéré comme stationnaire (partie tenue)
- La fin qui correspond à une décroissance de l'énergie du signal

On comprend donc que pour reconstituer les aigus d'un signal limité en bande, il est nécessaire de :

- Reconstituer à chaque instant le contenu hautes-fréquences du signal. Sur le signal représenté Figure 1.3 par exemple, on s'attachera à reconstituer les harmoniques hautes-fréquences, c'est à-dire les positionner à la bonne fréquence et leur donner la bonne amplitude.
- Préserver les caractéristiques temporelles du signal

Ce principe est valide sur des signaux "simples" comme illustrés ci-dessus mais aussi sur des sons plus complexes (mélange de sons).

Le principe de l'extension de bande d'un signal périodique stationnaire est illustré sur la Figure 1.4 sur laquelle on donne une version à bande limitée du spectre représenté Figure 1.3. Une des approches classiques, utilisée notamment par les codeurs d'extension de bande sur les signaux de parole, est de décomposer le problème en deux sous-problèmes :

- L'estimation et la transmission éventuelle de l'enveloppe spectrale haute-fréquence (l'enveloppe étant soit calculée et transmise par le codeur, soit estimée directement par le décodeur).
- L'extension de la structure fine du spectre haute-fréquence à partir des basses fréquences (synthèse des harmoniques et du niveau de bruit)

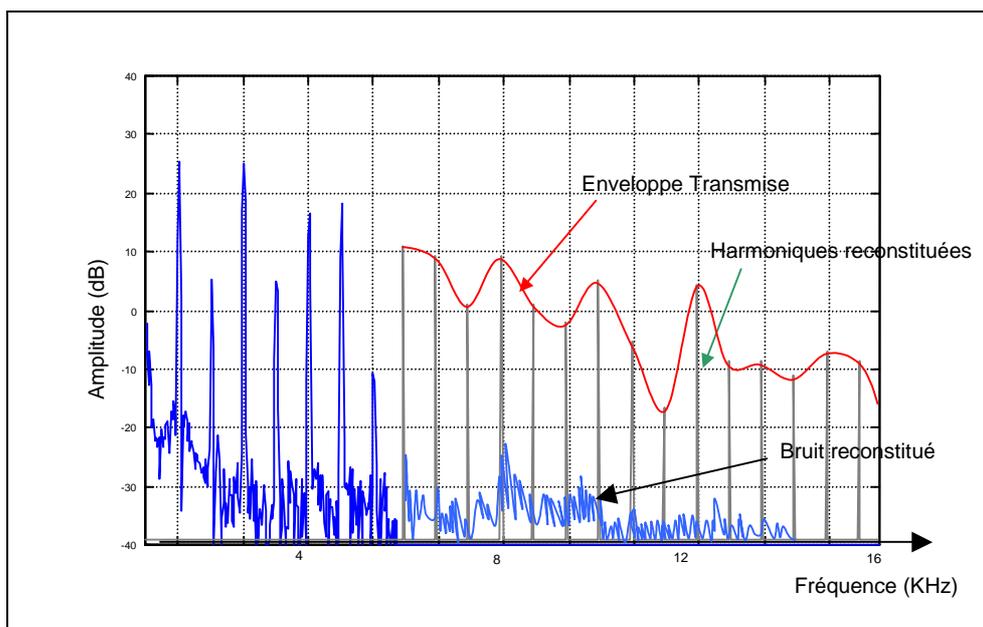


Figure 1.4 : Reconstruction du spectre par transmission d'enveloppe

Basé sur ce principe général, toute la problématique du système revient donc à trouver la meilleure enveloppe à transmettre ainsi que la meilleure façon d'étendre la structure fine du spectre de telle sorte que les hautes fréquences synthétisées soient perceptivement de même nature que celles de l'original.

1.5. Objectifs de la thèse

L'objectif de cette thèse est le développement d'un outil générique d'amélioration de la qualité des codeurs audio numériques à bas et moyen débit basé sur l'extension de bande. Contrairement aux techniques d'enrichissement de spectre déjà existantes, la technique développée devra fonctionner sur tout type de signaux audio numériques : signaux de parole et signaux musicaux, monophoniques ou stéréophoniques. Par le terme générique, on sous-entend que la technique est applicable quel que soit le type de codage utilisé par le codeur cœur (codeur par prédiction linéaire, codeur par transformée, codeur paramétrique...). Selon le type de codage et le débit alloué au codeur cœur, la fréquence de coupure du signal de base est susceptible de varier; un codeur de parole de type CELP en bande téléphonique codera par exemple les signaux jusqu'à une fréquence de 3,4 kHz, alors qu'un signal codé par un codeur par transformée aux environs des 30 kbit/s aura une bande d'environ 7 kHz (Tableau 1.1). La technique d'enrichissement de spectre développée devra donc s'adapter à ces différentes fréquences de coupure, non seulement selon le type de codeur cœur utilisé, mais également selon la stratégie de codage utilisée par celui-ci au cours du temps (fréquence de coupure variable au cours du temps).

On s'attachera à réaliser un système d'enrichissement spectral indépendant du codeur/décodeur cœur. Cet aspect est intéressant pour des notions de compatibilité et de hiérarchisation de trains binaires : Le système d'élargissement de spectre étant optionnel, ne pas l'utiliser permet d'obtenir néanmoins un premier niveau de restitution sonore (Figure 1.2).

Notons que, contrairement aux techniques d'enrichissement de spectre existantes (techniques dédiées à la parole), on a ici la possibilité d'analyser le signal à la source et la possibilité de transmettre des informations au décodeur pour la synthèse des hautes-fréquences. Le débit cible du codeur/décodeur d'extension de bande est fixé aux alentours de 2 kbit/s par voie. Face aux techniques classiques de transmissions des hautes fréquences (28 kbit/s pour le codeur AAC par exemple, comme illustré Tableau 1.1), on voit dès lors que cette approche offrirait, à qualité proche, un gain en débit considérable.

La qualité de restitution est essentielle. On tachera d'obtenir la meilleure qualité de codage possible sous les contraintes de débit imparties. Notons qu'avec les techniques de codages mises en œuvre, les mesures objectives sont inutilisables. La qualité est de ce fait totalement subjective. On utilise tout au long de cette thèse les méthodologies de tests subjectifs normalisées par l'ITU (tests subjectifs de type MUSHRA [ITU 00] et de type DCR [ITU 94]).

La technique développée étant en particulier dédiée à des applications audio numériques à décodeur embarqué, on s'attachera enfin à développer un décodeur de complexité raisonnable.

1.6. Plan du document

Le document est structuré en cinq chapitres.

Après ce premier chapitre introductif, nous donnons au *second chapitre* les principales propriétés psychoacoustiques utilisées en codage audionumérique, ainsi que les propriétés des signaux musicaux. Nous mettrons notamment l'accent sur la perception des hautes fréquences et sur les similarités entre les différentes parties du spectre. Nous évoquerons également les principales caractéristiques des signaux de parole et de musique et tenterons de regrouper la pluralité de ces sons en différentes classes de signaux de même nature.

Le *troisième chapitre* est consacré aux techniques d'estimation, de transmission et d'ajustement d'enveloppe spectrale appliquée à l'enrichissement de spectre. On distingue dans ce chapitre les techniques d'extrapolation d'enveloppe qui consistent à retrouver l'enveloppe haute-fréquence à partir de la connaissance seule de l'enveloppe basse-fréquence et les techniques d'estimation et de transmission d'enveloppe.

Le *quatrième chapitre* expose les techniques d'extension de la structure fine du spectre. On compare dans ce chapitre les différentes méthodes susceptibles d'étendre le spectre des signaux audio-numériques à bande-limitée. Sont développées en particulier les techniques basées d'une part sur l'extrapolation de spectre, et d'autre part sur les translations spectrales et sur les distorsions non-linéaires.

Nous détaillerons enfin au *chapitre 5* deux techniques complètes d'enrichissement de spectre bas débit:

- La technique SBR (Spectral Band Replication) développée depuis 1998 par la société Germano-suédoise Coding Technologies et en cours de normalisation dans MPEG-4
- La technique PAT (Perceptual Audio Transposition), solution alternative développée dans les laboratoires de France Télécom R&D durant ces trois années de thèse.

1.7. Bibliographie du chapitre 1

- [ATA 82] B.S. ATAL & J. REMDE
A new model of LPC excitation for producing natural sounding speech at low bit rates
Proc. Int. Conf. Acoust, Speech, Signal Processing, pp. 614-617, 1982
- [BOS 97] M. BOSI, et al.
ISO/IEC MPEG-2 Advanced Audio Coding
Journal Audio Engineering Society, pp. 789-813, Octobre 1997
- [EKS 02] P. EKSTRAND, A. EHRET, M. LUTZKY & T. ZIEGLER
Enhancing mp3 with SBR: Features and Capabilities of the new mp3PRO Algorithm.
112th AES Convention, Munich, Mai 2002
- [ISO 92] Information technology – Coding of moving pictures and associated audio for digital storage
Media at up to about 1.5Mbits/s, Part 3 : Audio" (MPEG-1)
ISO/IEC JTC1/SC29/WG11 MPEG, IS11172-3, 1992
- [ISO 94] Information technology – Generic Coding of moving pictures and associated Audio, Part 3 :
Audio" (MPEG-2)
ISO/IEC JTC1/SC29/WG11 MPEG, IS13818-3, 1994
- [ISO 01] Information technology – Generic Coding of Audio Visual Objects, Part 3 : Audio (MPEG-4)
ISO/IEC JTC1/SC29/WG11 FDIS 14496, edition 2001
- [ITU 94] International Telecommunication Union (ITU)
Methods for the subjective assessment of small impairments in audio systems including
multichannel sound systems
Recommendation BS.1116, 1994
- [ITU 00] International Telecommunication Union (ITU)
Multi stimulus test with hidden reference and anchor (MUSHRA) - EBU method for subjective
listening tests of intermediate audio quality
Recommendation ITU [10-11Q/62], Janvier 2000
- [PUR 00] H. PURNHAGEN & N. MEINE
HILN : The MPEG-4 Parametric Audio Coding Tools
Proceedings ISCAS, Mai 2000
- [UIT 96] Codage de la parole à 8kbit/s par prédiction linéaire avec excitation par séquences codées à
structure algébrique conjuguée – G.729
Union internationale des télécommunications, Recommandation UIT-T G729, Mars 1996

Documents associés à la normalisation dans DRM

- [DRM 00a] Digital Radio Mondiale DRM TC SC 056
Subjective tests on the Perceptual Audio Transposition system (PAT) : an alternative to the
SBR system.
P. COLLEN, P. PHILIPPE, et J.C. RAULT , Erlangen, Mars 2000
- [DRM 00b] Digital Radio Mondiale DRM TC SC 061
Test plan for DRM Audio Bandwidth widening tools (SBR and PAT)
Berlin, Avril 2000
- [DRM 00c] Digital Radio Mondiale DRM TC SC 071
Complexity of PAT decoder, France Télécom R&D
P. COLLEN, P. PHILIPPE, et J.C. RAULT , Décembre 2000
- [DRM 00d] Digital Radio Mondiale DRM TC SC 077

DRM Source Coding Group, Report on Subjective Listening Tests of AAC-based
T. BUCHOLZ (T-Nova), T. MLASKO & F. HOFMANN (Bosch), A. MURPHY (BBC),
Berlin, Décembre 2000

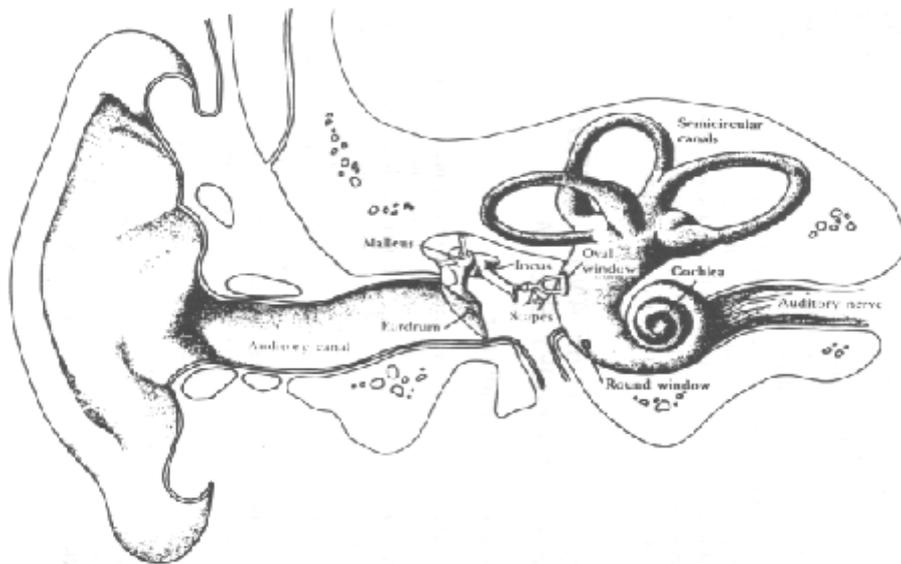
[DRM 01] Digital Radio Mondiale (DRM) : ETSI TS 101 980 V1.1.1
System Specification
Berlin, Septembre 2001

Documents associés à la normalisation dans MPEG-4

- [ISO 00a] ISO/IEC JTC1/SC29/WG11 MPEG00/M6141
Request for Action w.r.t. the “Call for evidence justifying the testing of audio coding
technology” made in Geneva
P. PHILIPPE, Beijing, China, Juillet 2000
- [ISO 00b] ISO/IEC JTC1/SC29/WG11 MPEG00/M6437
Report of the AHG on Audio Call for Evidence
S. R. Quackenbush, La Baule, France, Octobre 2000
- [ISO 01a] ISO/IEC JTC1/SC29/WG11 MPEG01/N3793
Report on Audio call for Evidence
Pise, Italie, Janvier 2001
- [ISO 01b] ISO/IEC JTC1/SC29/WG11 MPEG01/N3794
Call for Proposals for New Tools for Audio Coding
Pise, Italie, Janvier 2001
- [ISO 01c] ISO/IEC JTC1/SC29/WG11 MPEG01/N4225
Report on Proposals for New Tools for Audio Coding
Sydney, Juillet 2001
- [ISO 01d] ISO/IEC JTC1/SC29/WG11 MPEG01/N4378
Report of MPEG-4 Audio Bandwidth Extension RM0 Test
Pattaya, Décembre 2001
- [ISO 02] ISO/IEC JTC1/SC29/WG11 MPEG02/N4611
WD Text for Backward Compatible Bandwidth Extension for General Audio Coding
Jeju, Korea, Mars 2002

CHAPITRE 2

PROPRIETES ET PERCEPTION DES SIGNAUX DE PAROLE ET DE MUSIQUE



Plan du chapitre

1.1.	INTRODUCTION.....	32
1.2.	PROPRIETES DE L'OREILLE HUMAINE.....	33
1.3.	PROPRIETES DES SIGNAUX AUDIONUMERIQUES.....	37
1.4.	CONCLUSION DU CHAPITRE.....	46
1.5.	BIBLIOGRAPHIE DU CHAPITRE 2.....	47

2.1. Introduction

La psychoacoustique est la branche de l'acoustique qui étudie la perception des sons par le système auditif humain. Le modèle associé aide à saisir la manière dont nous percevons les sons.

L'oreille humaine est un récepteur complexe de comportement non-linéaire. L'étude de ses caractéristiques permet d'exploiter les propriétés intrinsèques de l'oreille afin de réaliser des systèmes de compression avec perte d'informations. Cette compression est supposée transparente lorsque les exigences du modèle psychoacoustiques sont satisfaites.

On voit dès lors toute l'importance du modèle psychoacoustique qui permet, tout en accompagnant les caractéristiques fondamentales de l'oreille, de guider l'injection de "dégradations" dans le signal audio sans en entacher sa qualité, offrant ainsi des informations essentielles pour la réduction de débit.

La première partie de ce chapitre présente la plupart des propriétés psychoacoustiques utilisées en codage audionumérique. La sensibilité de l'oreille humaine dans les différentes bandes de fréquences, la notion de bandes critiques, les phénomènes psychoacoustiques de masquage et enfin la notion de dissonance seront développés dans cette première partie.

La connaissance des propriétés spectrales et temporelles des signaux est essentielle pour notre application et nous tâchons, dans la seconde partie, de caractériser l'ensemble des signaux audio numériques. Après une première distinction entre les signaux de parole et les signaux musicaux en général, nous tâchons de regrouper la pluralité de ces sons en cinq classes de signaux de même nature.

2.2. Propriétés de l'oreille humaine

Nous reprenons ici les principaux résultats présentés par Zwicker dans [ZWI 81]. Toutes les propriétés psychoacoustiques présentées ici seront reprises par la suite et justifieront du choix des techniques mises en œuvre dans le codeur/décodeur d'enrichissement de spectre.

2.2.1. Sensibilité de l'oreille humaine

On considère que l'oreille humaine est capable de discerner les sons compris entre 20 Hz et 20 kHz. L'oreille n'est toutefois sensible à un son pur, dans une ambiance parfaitement silencieuse, que si sa puissance est supérieure au seuil d'audition absolu (absolute threshold of hearing). Ce seuil est exprimé en dB SPL (Sound Pressure Level) et approché par l'équation :

$$T_q(f) = 3.64\left(\frac{f}{1000}\right)^{-0.8} - 6.5e^{-0.6\left(\frac{f}{1000}-3.3\right)^2} + 0.001\left(\frac{f}{1000}\right)^4 \quad (\text{dB SPL}), \text{ avec } f \text{ en Hz} \quad (2.1)$$

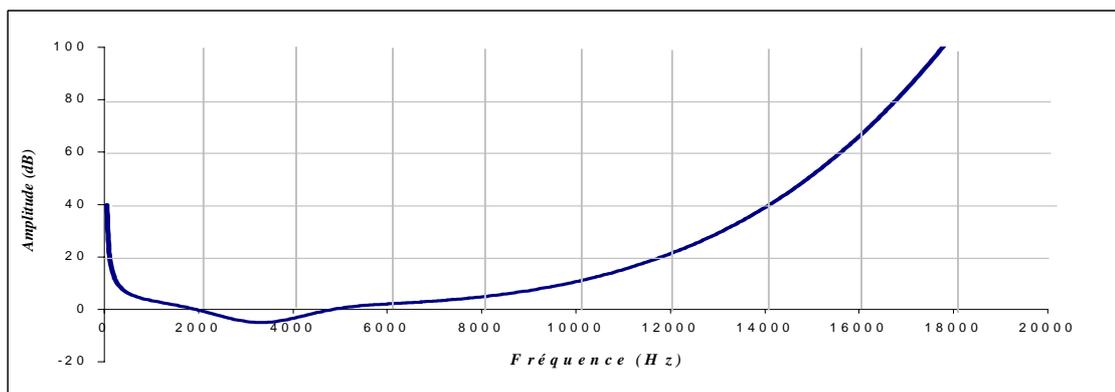


Figure 2.1 : Seuil d'audition absolu

L'oreille possède ainsi un maximum de sensibilité pour des fréquences comprises entre 2 et 5 kHz. Vers les plus hautes fréquences en revanche, la courbe croît exponentiellement et l'oreille n'y perçoit plus que les signaux d'énergie relativement élevée. Au-delà de 16 kHz, seuls les sons d'énergie supérieure à 60 dB SPL sont perçus. L'oreille est pratiquement insensible aux signaux de fréquence supérieure à 20 kHz.

2.2.2. Bandes critiques

La notion de bande critique est d'une importance primordiale en psychoacoustique. La puissance perçue par l'oreille dans une bande critique est égale à la somme de toutes les puissances des composantes dans cette bande de fréquence. Si cette somme est supérieure au seuil d'audition absolu alors le signal compris dans la bande considérée est audible, sinon il est masqué et donc inaudible.

Les bandes critiques jouent également un rôle dans la perception de l'intensité et de la hauteur. La force sonore perçue est ainsi indépendante de la largeur de bande du signal, tant que celle-ci est inférieure à la largeur de la bande critique concernée ([ZWI 81])

On peut modéliser l'oreille sous la forme d'un banc de filtres s'échelonnant le long du domaine audible. La zone entre 15 Hz et 16 kHz peut par exemple être divisée en 24 bandes dont la largeur varie en fonction de la fréquence ([FLE 40]).

Numéro de la bande	Largeur de la bande (Hz)	Fréquence inférieure (Hz)	Fréquence supérieure (Hz)	Numéro de la bande	Largeur de la bande (Hz)	Fréquence inférieure (Hz)	Fréquence supérieure (Hz)
1	80	20	100	13	280	1720	2000
2	100	100	200	14	320	2000	2320
3	100	200	300	15	380	2320	2700
4	100	300	400	16	450	2700	3150
5	110	400	510	17	550	3150	3700
6	120	510	630	18	700	3700	4400
7	140	630	770	19	900	4400	5300
8	150	770	920	20	1100	5300	6400
9	160	920	1080	21	1300	6400	7700
10	190	1080	1270	22	1800	7700	9500
11	210	1270	1480	23	2500	9500	12000
12	240	1480	1720	24	4000	12000	16000

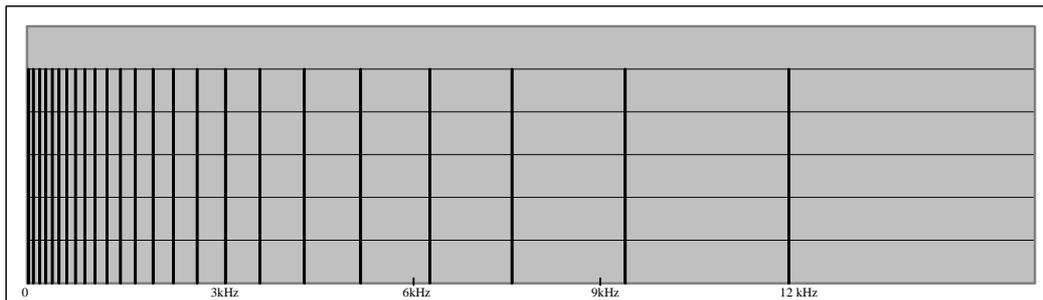


Figure 2.2 : Bandes critiques

Afin de prendre en compte la perception de l'oreille humaine et les phénomènes psychoacoustiques, une nouvelle échelle fréquentielle a été introduite, l'échelle des Barks, dans laquelle, un accroissement de 1 Bark correspond à une augmentation en fréquence de 1 bande critique. La relation Bark/Hertz est quasi-linéaire jusqu'à 500Hz; Au-delà, elle est quasi-logarithmique. Une relation analytique l'approchant est fournie dans [ZWI 80] :

$$v(f) = 13 \arctg(0.00076 \cdot f) + 3.5 \cdot \arctg\left[\left(\frac{f}{7500}\right)^2\right] (\text{Bark}), \text{ avec } f \text{ en kHz} \quad (2.2)$$

2.2.3. Phénomènes psychoacoustiques de masquage

Un certain nombre de critères perceptifs auditifs permet de réduire considérablement le volume des données à stocker ou à transmettre sans dégradation subjective de la qualité. On distingue deux types de masquage que nous développons dans ce paragraphe : le phénomène de masquage fréquentiel et celui de masquage temporel.

Dans le cadre du codage audio, le phénomène le plus exploité est le masquage fréquentiel. Les courbes de masquage déduites du modèle psychoacoustique offrent le moyen de répartir le bruit de codage en dessous du seuil d'audition sans altérer la perception du son.

2.2.3.1. Masquage fréquentiel

Notre oreille est équipée de récepteurs sélectifs en fréquence, traitant des zones fréquentielles dont la largeur est proportionnelle à la largeur de la bande critique. Le masquage fréquentiel simultané existe lorsque les sons sont présents simultanément, un son pouvant ainsi être rendu inaudible par la présence d'un second signal dit *son masquant*.

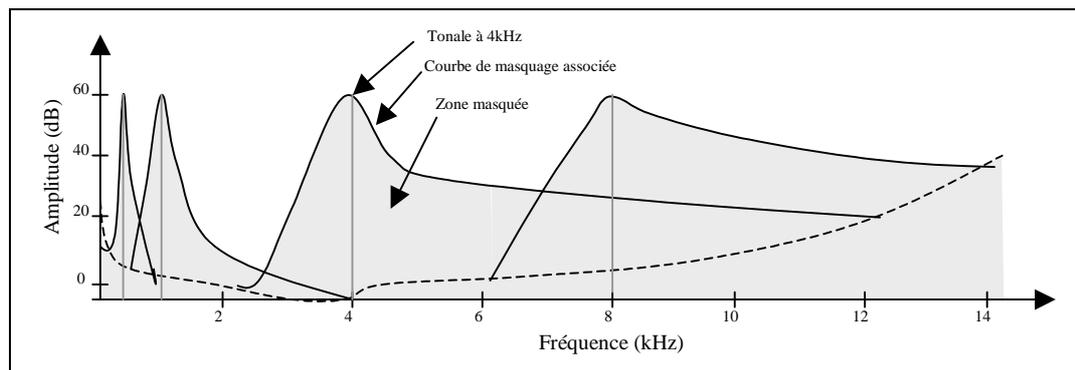


Figure 2.3 : Masquage fréquentiel et courbes de masquage associées

La courbe de masquage, déduite du modèle psychoacoustique, est utilisée dans la phase de codage et permet de répartir efficacement le bruit de quantification sans dégradation audible (mise en forme spectrale du bruit de quantification). Cette méthode minimise en moyenne le nombre de bits alloué pour la quantification en adaptant localement la répartition des bits en fonction des caractéristiques de l'oreille et du son à transmettre.

La Figure 2.3 donne les courbes de masquages associées à quatre tonales de fréquences respectives 330, 1000, 4000 et 8000 Hz. Les sons situés dans la zone grisée située en dessous de la courbe de masquage déduite sont masqués.

2.2.3.2. Masquage temporel

Quand l'oreille a été stimulée par un son pur, après cessation du son, il y a une perte de sensibilité autour de cette fréquence : environ 10 dB de perte qui disparaissent au bout de quelques centaines de millisecondes. Les phénomènes de masquage apparaissent dans le domaine temporel lors de fortes variations du signal, les signaux transitoires créant des zones de pré et de post masquage importantes. Ainsi l'oreille ne perçoit pas les sons faibles précédant ou suivant immédiatement un son de forte intensité. La durée effective du masquage temporel antérieur est brève, de l'ordre de 5ms, contrairement au masquage postérieur qui persiste plus de 100ms.

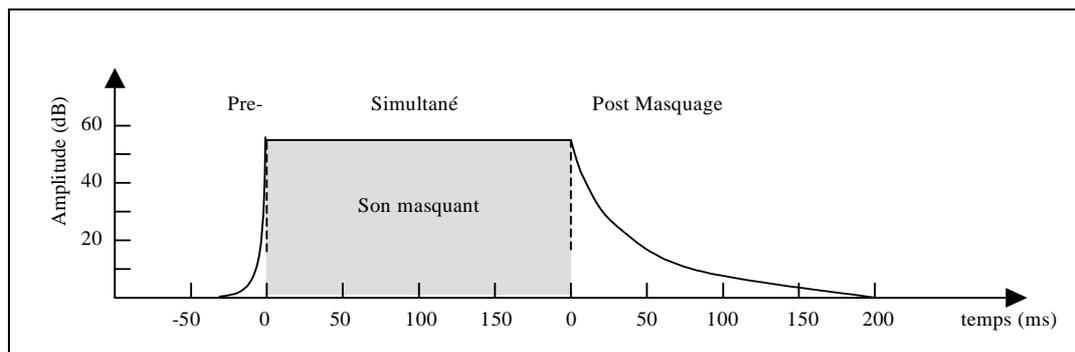


Figure 2.4 : Différents types de masquages temporels

Le phénomène de masquage temporel est difficile à modéliser et donc peu utilisé en codage audio. Il est toutefois exploité dans les codeurs par transformée (codeur AAC par exemple) pour le traitement des signaux transitoires. Lors de l'apparition de tels signaux, la sélection de fenêtres d'analyse plus courtes est utilisée afin de réduire les phénomènes d'étalement du bruit non masqués.

2.2.4. Dissonance et JND (Just Noticeable Difference)

Nous mettons l'accent dans ce paragraphe sur trois notions importantes de la perception auditive. Ces notions nous guideront dans le choix des techniques mises en œuvre dans la suite de ce document.

2.2.4.1. Phénomène de dissonance

Un critère de dissonance est présenté dans [PLO 65], et montre que deux tonales sont considérées comme dissonantes si leur différence de fréquence est approximativement contenue dans 5 à 50% de la largeur de la bande critique dans laquelle elles sont situées. Pour référence, la largeur de bande critique pour une fréquence donnée peut être approchée par :

$$cb(f) = 25 + 75(1 + 1.4(\frac{f}{1000})^2)^{0.69} \quad (2.3)$$

avec f et cb en Hz.

C'est ainsi que, présentées simultanément, deux tonales de fréquences respectives 400 Hz et 411 Hz (différence de 10% de la largeur de la bande critique), entrent en dissonance et génèrent des artefacts perceptivement gênants.

2.2.4.2. JND fréquentiel

Le JND fréquentiel est la limite fréquentielle à partir de laquelle l'oreille ne perçoit plus la différence de hauteur entre deux signaux de fréquence pure, présentés séparément, de fréquence voisine [LEI 96]. Ce seuil de discrimination des hauteurs se situe aux alentours de 1%. C'est ainsi que, présentées séparément, deux tonales à 400 et 404Hz provoquent la même sensation de hauteur (une seule et unique fréquence perçue).

Basée sur cette notion de JND fréquentiel, on constate dès lors que la sensation de hauteur devient approximative en haute-fréquence.

2.2.4.3. JND temporel

L'échelle logarithmique des amplitudes est adaptée à notre perception des hauteurs. Le JND temporel est la limite énergétique à partir de laquelle l'oreille ne perçoit plus la différence d'intensité entre deux sons d'énergie proche [LEI 96]; il se situe aux environs de 1dB; ce qui signifie qu'en terme de perception auditive, 1 dB est la différence minimale audible de la pression acoustique sous les meilleures conditions. En d'autres termes, l'oreille ne perçoit pas les variations de puissance d'un signal de moins de 1 dB au cours du temps.

2.3. Propriétés des signaux audionumériques

Ce paragraphe décrit les caractéristiques spectrales et temporelles des signaux audionumériques. Cette caractérisation des signaux est importante dans le cadre de cette étude afin de bien comprendre et d'adapter les techniques à mettre en œuvre pour étendre efficacement le spectre de tous les signaux audionumériques. Nous distinguons en un premier temps les signaux de parole de ceux de musique avant d'introduire cinq classes de signaux représentatifs des sons les plus fréquents.

2.3.1. Signaux de parole

Les ondes acoustiques produites par le système vocal peuvent être représentées efficacement par un modèle source-filtre. Pour les sons voisés (voyelles notamment), la source est modélisée par un train périodique résultant d'une vibration des cordes vocales. Pour les sons non-voisés (fricatives f, s, ...), la source est modélisée par un bruit blanc, les sons fricatifs résultant de l'écoulement de l'air dans une constriction étroite située en un point du conduit vocal, en particulier au niveau des lèvres et des dents [KLE 95].

Le filtre représente le comportement du conduit vocal.

Notons enfin que les sons plosifs (ou occlusifs) sont produit par une occlusion momentanée du conduit vocal en un point donné, suivie par une ouverture brusque. Ces sons peuvent être voisés (b,d...) ou non-voisés (p,t...).

Ce modèle est très répandu en codage et permet de réduire fortement les informations; citons pour exemple le codeur de parole CELP [SCH 85] qui fonctionne à partir d'un débit de 6 kbit/s.

Le signal vocal peut être considéré comme quasi-stationnaire sur des intervalles de temps de l'ordre de 20 ms. La plupart des codeurs de parole analysent ainsi le son par tranches de 15 à 20 ms.

On repère typiquement 4 formants (maximums de l'enveloppe spectrale représentée Figure 2.5) dans le spectre du signal de parole en bande téléphonique (300Hz-3,4kHz), ce nombre passe à 7 en bande élargie (50Hz-7.4kHz).

La localisation des trois premiers formants est essentielle pour caractériser le spectre vocal [JBI 99]; les formants d'ordres supérieurs ont une influence plus limitée dans la compréhension du son mais sont toutefois primordiaux pour un rendu sonore de qualité.

Le signal de parole présente enfin une atténuation spectrale d'environ 12 à 15 dB par octave, limitant ainsi la bande la plus énergétique à environ 3 kHz. Au-delà de 8 kHz, le signal de parole est atténué et ne contribue que peu à l'intelligibilité; la suppression de cette bande ne nuit pas de ce fait à la qualité de restitution dans la plupart des cas.

2.3.1.1. Distinctions signaux voisés/non-voisés

Comme nous l'avons vu précédemment, les signaux de parole peuvent se décomposer en deux principales catégories de signaux, les sons voisés et les sons non voisés. La Figure 2.5 superpose les spectres des deux types de sons et met en évidence la structure formantique, c'est-à-dire l'enveloppe spectrale, qui joue un rôle primordial dans la perception des signaux de parole. La précision de cette enveloppe au cours du temps contribue en effet à plus d'intelligibilité et de naturel dans la voix.

Les sons voisés résultent de l'excitation du conduit vocal par des impulsions quasi-périodiques de pression liées aux vibrations des cordes vocales. Le degré de périodicité et la continuité temporelle entre les différents segments jouent un rôle primordial ([SPA 94]) dans la qualité de restitution. Les séquences voisées possèdent une fréquence fondamentale et des harmoniques multiples de celle-ci. La

période associée à la fréquence fondamentale (pitch) est comprise entre 60 et 600 Hz selon le locuteur. La valeur moyenne du pitch est de l'ordre de 150 Hz pour un locuteur masculin et de 250 Hz pour un locuteur féminin.

L'énergie des signaux voisés est concentrée en basse-fréquence et décroît rapidement en haute-fréquence. Le niveau de bruit présent dans les sons voisés est fonction du degré de voisement mais il est en général faible en basse-fréquence et la bande [0-2kHz] se compose essentiellement de composantes tonales de forte énergie. Le niveau de bruit augmente ensuite en haute-fréquence et au-delà des 4 kHz, les harmoniques sont "noyées" dans le bruit (niveau d'énergie des harmoniques comparable à celui du bruit).

Le flux d'air produit par les poumons peut également être restreint à sortir par une petite ouverture au niveau du conduit vocal, causant des turbulences, c'est-à-dire du bruit et par conséquent de la parole non-voisée. Contrairement aux sons voisés, les sons non-voisés ne présentent pas de structure périodique. Ils peuvent être modélisés par un bruit blanc filtré par le conduit vocal. La structure fine du spectre est de ce fait sensiblement la même sur tout le spectre.

Notons que contrairement aux sons voisés, leur énergie est plus concentrée dans les hautes fréquences.

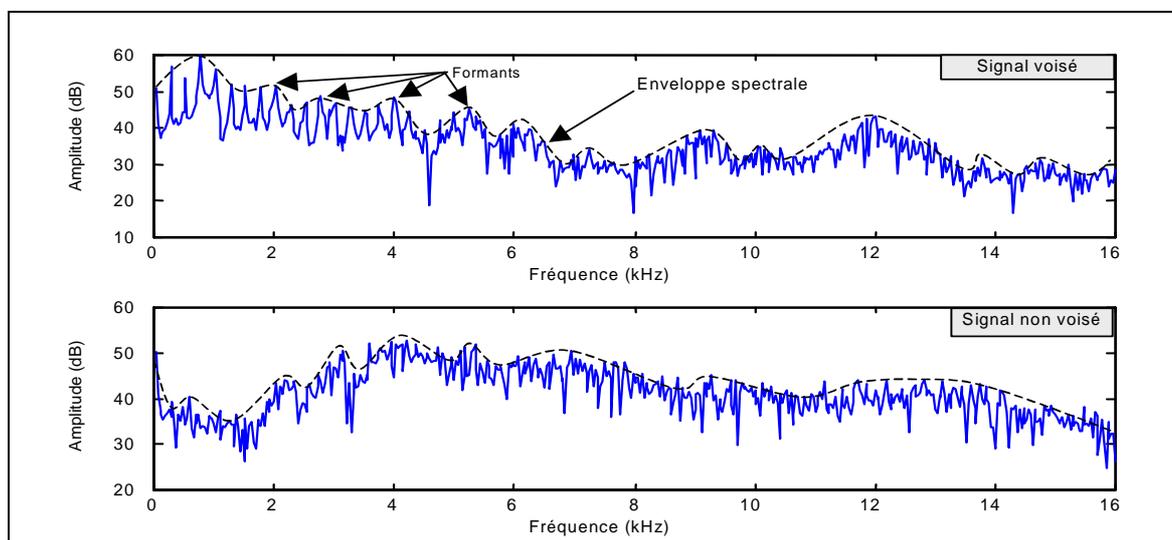


Figure 2.5 : Distinction parole voisée / non-voisée (signaux stationnaires)

2.3.1.2. Remarque

La catégorie des signaux de parole traitée dans ce paragraphe englobe les signaux de parole mono-locuteurs. Les signaux de parole multi-locuteurs, qui résultent d'un mélange de plusieurs sons mono-locuteur, peuvent être composés de plusieurs fréquences fondamentales et ne vérifient plus le modèle de production source-filtre défini en 2.3.1.

2.3.2. Signaux musicaux

Les instruments de musique seuls peuvent également être représentés par un modèle source filtre adapté. Notons que le signal de parole est ce sens assimilé à un instrument de musique particulier.

La catégorie des signaux musicaux dits « génériques » résulte du mélange de tous ces sons (signaux de parole et instruments de musique). Ces signaux peuvent revêtir des formes extrêmement variées, tant du point de vue temporel que spectral, et sont de ce fait difficiles à caractériser. Il n'existe en particulier pas de modèle de production simple et universel permettant de représenter efficacement la pluralité des signaux musicaux. Ils possèdent généralement une structure harmonique forte et peuvent présenter une dynamique en puissance importante.

Les signaux représentés au paragraphe suivant illustrent cette diversité temporelle et fréquentielle.

2.3.3. Classes de signaux et propriétés haute-fréquence

Après avoir vu les propriétés des signaux de parole et de musique, nous présentons ici cinq classes de signaux représentatives de la pluralité des sons musicaux. On tente pour chacune d'elles de décrire les différentes techniques à mettre en œuvre afin de synthétiser les hautes fréquences à partir des signaux à bande-limitée. On suppose dans ce paragraphe que les signaux sont filtrés (filtrage passe-bas) aux environs de 5 kHz. A partir de la bande [0-5kHz], on cherche à synthétiser les hautes fréquences comprises entre 5 et 16 kHz.

Les 5 signaux synthétiques décrits ci après sont échantillonnés à 32 kHz. Pour chacun d'entre eux, on donne :

- Le spectre d'une trame de 32 ms (1024 échantillons) choisie arbitrairement et représentative du signal.
- La correspondance avec les signaux réels, c'est-à-dire les instruments pouvant produire de tels sons.
- On associe également à chacun des signaux un modèle mathématique du signal basé sur une décomposition paramétrique en composantes "simples". Ce modèle simple, décrit en détail au paragraphe ci-dessous, permet d'analyser de façon analytique la corrélation entre les différentes bandes spectrales.

2.3.3.1. Modèle paramétrique associé

Les signaux audio-numériques $s(t)$ peuvent être modélisés par une combinaison linéaire des quatre composantes élémentaires suivantes :

$$Harm(t), InHarm(t); Noise(t); Trans(t)$$

où :

$Harm(t)$ représente la partie harmonique du signal et est composé d'une ou de plusieurs séries harmoniques $Harm_i(t)$. Chacune de ces séries harmoniques est déterminée par sa fréquence fondamentale f_i , l'amplitude $A_{i,n}$ et la phase $\Phi_{i,n}$ de chacun des partiels et s'écrit :

$$Harm_i(t) = \sum_{n=1}^{N_i} A_{i,n} \cos(2\pi n f_i t + \phi_{i,n}) \quad (2.4)$$

N_i déterminant le nombre de partiels composant chacune des séries harmoniques.

$InHarm(t)$ représente la partie inharmonique du signal et est composée de tonales isolées non harmoniquement liées (sans relation entre elles). Chacune des tonales est déterminée par son amplitude, sa fréquence et sa phase :

$$InHarm(t) = \sum_{m=0}^M B_m \cos(2\pi f_m t + \phi_m) \quad (2.5)$$

$Trans(t)$ représente les composantes transitoires du signal (attaques de castagnettes, percussion..). Cette catégorie de signaux est composée de bruit dans la majorité des cas et est caractérisée par de fortes variations d'énergie sur des temps très courts (quelques millisecondes).

$Noise(t)$ représente le signal résiduel (composantes non paramétrables par les trois composantes décrites ci-dessus) et est souvent associé à un bruit blanc mis en forme par une enveloppe fréquentielle et temporelle de la forme suivante :

$$Noise(t) = e(t)[h(t, \tau) * u(t)] \quad (2.6)$$

$Noise(t)$ est modélisé par filtrage d'un bruit blanc $u(t)$ par un filtre tout pôle $h(t, \tau)$ et multiplié par une fonction d'enveloppe temporelle $e(t)$.

Dans ce modèle, on considère les signaux $Harm(t)$ et $InHarm(t)$ stationnaires pendant la durée d'une trame d'analyse (32 ms). En d'autres termes, l'amplitude, la fréquence et la phase des composantes sinusoïdales sont supposées constantes pendant cette durée.

2.3.3.2. Signal S_{Noise}

Le signal S_{noise} , dont le spectre est représenté Figure 2.6, est constitué d'un bruit blanc mis en forme par une enveloppe spectrale. Cette enveloppe, qui décrit la structure formantique du signal, est composée de 9 formants sur l'exemple Figure 2.6.

S_{Noise} est associé au modèle paramétrique $Noise(t)$ décrit ci-dessus et représente en particulier la catégorie des signaux de parole non-voisée.

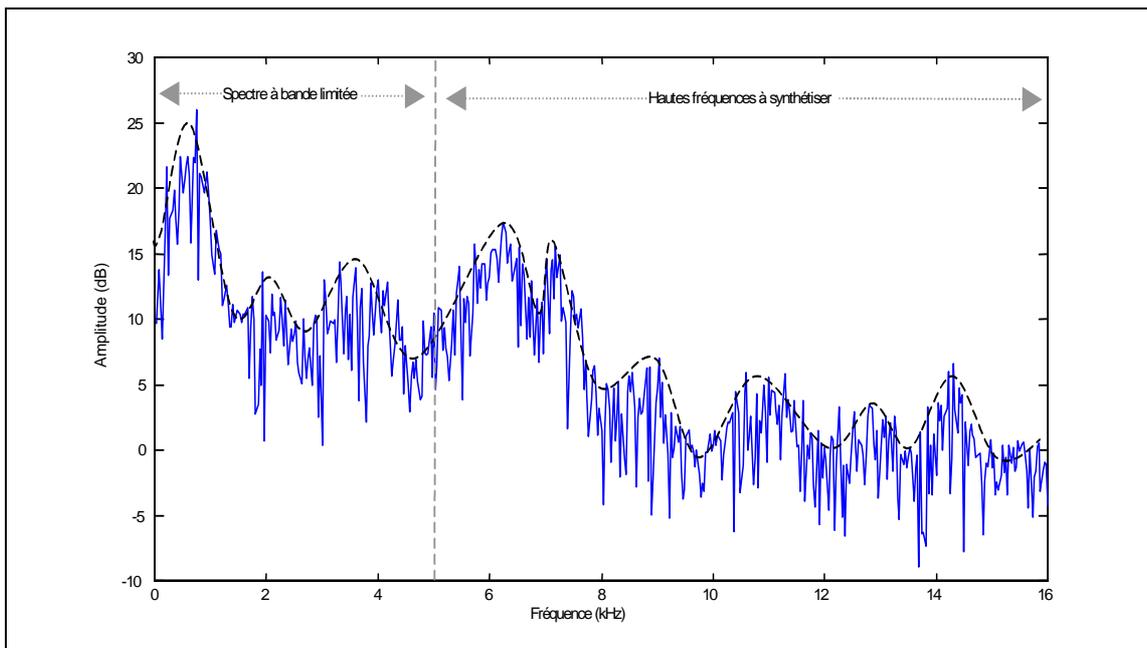


Figure 2.6 : Snoise

L'extension de bande de ce type de signaux peut être réalisée efficacement par injection de bruit en haute-fréquence et par une mise en forme du bruit synthétisé par une enveloppe spectrale adéquate.

La structure fine spectrale, composée de bruit, étant la même en basse et en haute-fréquence, l'extension de bande de ce type de signaux peut également se faire par translation du contenu basse-fréquence vers les hautes fréquences et par un ajustement d'énergie du signal translaté par une enveloppe adéquate.

2.3.3.3. Signal $S_{\text{Mono-harm}}$

Le signal $S_{\text{mono_harm}}$, dont le spectre est représenté Figure 2.7, est constitué de bruit et d'une série harmonique (peigne de sinus) de fréquence fondamentale $f_0 = 440\text{Hz}$ et de ses $N=35$ partiels multiples de f_0 , le tout mis en forme par une enveloppe spectrale.

Il est associé au modèle paramétrique

$$S_{\text{Mono_harm}}(t) = \text{Harm}_0(t) + \text{Noise}(t) = \sum_{n=1}^N A_{0,n} \cos(2\pi n f_0 t + \phi_{0,n}) + \text{Noise}(t) \quad (2.7)$$

$S_{\text{Mono_harm}}$ représente la gamme des sons de parole mono-locuteur voisée et des signaux musicaux entretenus tels que l'accordéon, la guitare, le violon (instruments de musique harmoniques ou faiblement inharmoniques⁴).

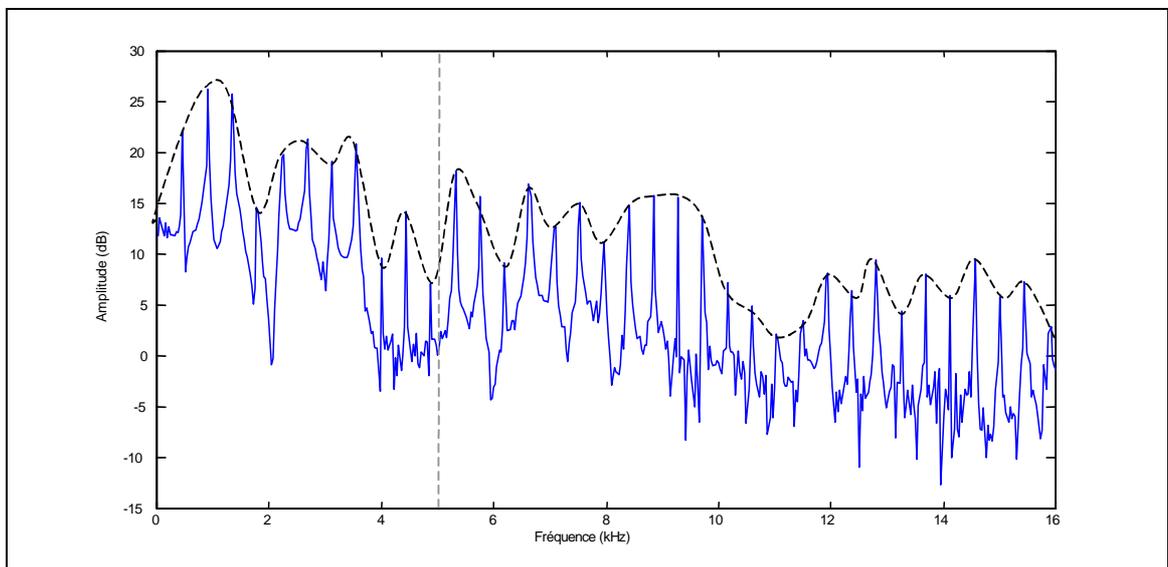


Figure 2.7 : $S_{\text{mono_harm}}$

Deux approches permettent d'étendre efficacement la bande passante de tels signaux:

- L'approche paramétrique qui vise à traiter séparément les composantes harmoniques et bruitées du signal. La partie harmonique du signal peut être synthétisée par extraction de la fréquence fondamentale f_0 et par extension de la série harmonique tronquée (génération des partiels supérieurs), l'amplitude des partiels synthétisés devant ensuite être ajustée, soit individuellement, soit par une enveloppe spectrale adéquate. Le bruit peut être synthétisé selon la méthode décrite au paragraphe précédent (synthèse de bruit et ajustement d'énergie par une enveloppe)
- Partant de la constatation que la structure fine est la même sur tout le spectre, la seconde approche consiste à translater le spectre basse-fréquence vers les hautes fréquences, en prenant soin de respecter l'écart (pitch) entre les raies spectrales et le rapport tonales à bruit en haute-fréquence. Le signal synthétisé devra ensuite être ajusté en énergie par une enveloppe spectrale adéquate.

⁴ Un critère d'inharmonicité est défini équation (2.9)

2.3.3.4. Signal $S_{\text{Multi_harm}}$

Le signal $S_{\text{Multi_harm}}$ est constitué de bruit et de plusieurs séries harmoniques de fréquences fondamentales différentes. Celui représenté Figure 2.8 est constitué de 2 séries harmoniques de fréquence fondamentales respectives $f_0=440$ Hz et $f_1=750$ Hz. Il est associé au modèle paramétrique :

$$S_{\text{Multi_harm}}(t) = \sum_{n=1}^{N_0} A_{0,n} \cos(2\pi n f_0 t + \phi_{0,n}) + \sum_{m=1}^{N_1} A_{1,m} \cos(2\pi m f_1 t + \phi_{1,m}) + \text{Noise}(t) \quad (2.8)$$

$S_{\text{Multi_harm}}$ est représentatif de la plupart des signaux musicaux. Il correspond à un mélange d'instruments harmoniques et/ou de parole.

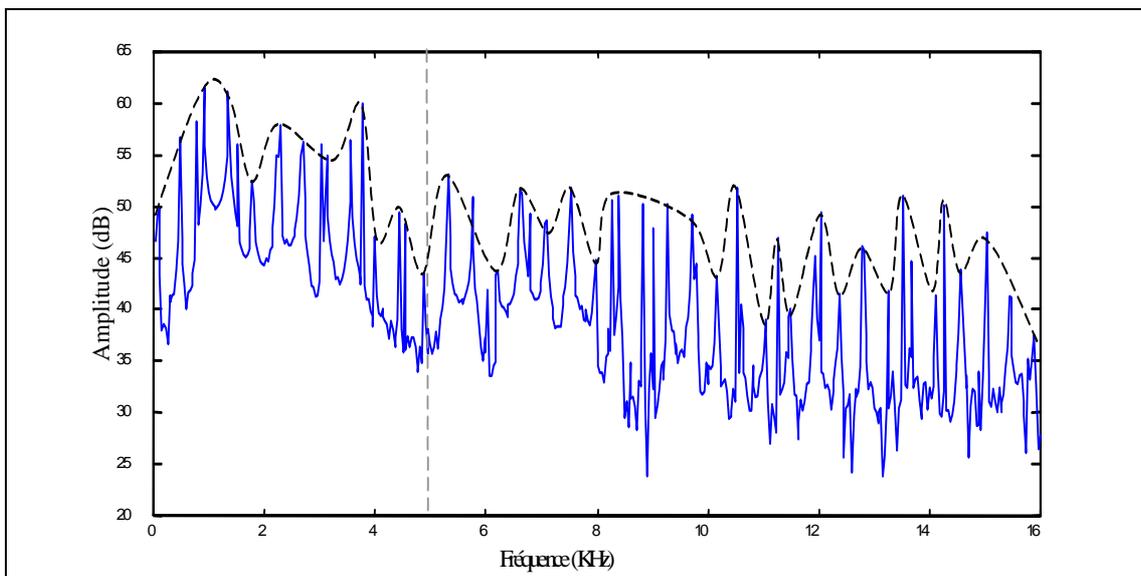


Figure 2.8 : $S_{\text{Multi_harm}}$

L'extraction des composantes élémentaires (bruit, fréquences fondamentales...) est plus délicate à réaliser sur ce type de signaux et l'approche paramétrique devient de ce fait complexe à réaliser.

Notons toutefois que, là encore, la structure fine du spectre, composée des deux peignes mélangés et de bruit, est la même sur toute la bande. L'extension de bande de ce type de signaux peut donc se faire efficacement par translation du contenu basse-fréquence vers les hautes fréquences et par un ajustement d'énergie du signal translaté par une enveloppe adéquate.

2.3.3.5. Signal S_{Inharm}

Le Signal S_{Inharm} , dont le spectre est représenté Figure 2.9, est composé de tonales réparties sur l'axe des fréquences sans relation harmonique entre elles. Il est associé au modèle paramétrique $S_{\text{Inharm}}(t)$ décrit équation (2.5).

Ce signal représente la catégorie des instruments inharmoniques, tels que le carillon, la cloche ou le triangle.

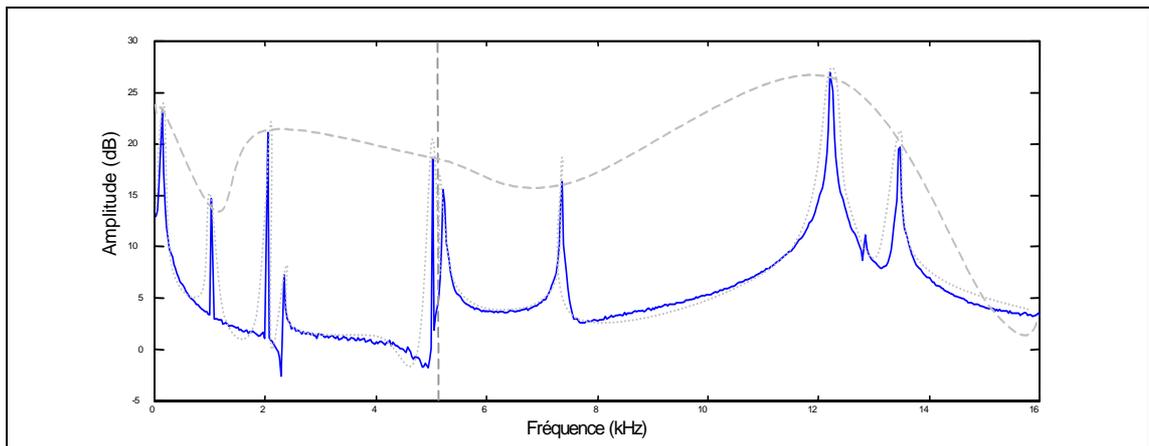


Figure 2.9 : S_{Inharm}

Les tonales en haute-fréquence n'étant pas liées à celles contenues en basse-fréquence, l'extension de la structure fine des signaux inharmoniques est irréalisable à partir de la connaissance seule du signal à bande limitée. Des données auxiliaires (position des tonales hautes-fréquences) sont dès lors incontournables.

De plus la notion d'enveloppe spectrale devient difficile à définir pour ce type de signaux, car généralement très peu de tonales sont mises en jeu. Selon le type d'application souhaitée, l'enveloppe peut en effet prendre différentes formes (courbes grisées Figure 2.9). Ce problème sera développé au chapitre suivant.

Citons enfin d'autres types d'inharmonicité, telle que celles générées par le piano par exemple. Le spectre du signal de piano est composé d'une fréquence fondamentale f_0 et de partiels non multiples de cette fondamentale situés aux fréquences :

$$f_n = nf_0 [1 + B(n^2 - 1)]^{0.5} \quad (2.9)$$

où B représente le facteur d'inharmonicité ([KLA 99]).

2.3.3.6. Signal S_{Trans}

La Figure 2.10 décrit le comportement temporel et fréquentiel (spectrogramme) du signal S_{Trans} qui correspond à la succession de trois signaux transitoires :

- La première partie du signal comprise entre 0 et 0.3s est une attaque de castagnettes. Cette attaque correspond à une brusque variation d'énergie temporelle, répartie sur toute la bande de fréquence. La structure fine du spectre est sensiblement la même sur cette bande.
- La seconde partie, comprise entre 0.3 et 0.45s correspond à une attaque d'un signal harmonique (clavecin sur l'exemple). Cette attaque correspond à une brusque variation d'énergie temporelle, répartie sur toute la bande de fréquence. Elle est facilement modélisable par un peigne d'harmonique (modèle paramétrique $Harm(t)$) mis en forme par une enveloppe temporelle variant fortement en un laps de temps très court. La structure fine du spectre est la même sur toute la bande.
- La dernière partie du signal comprise entre 0.45 et 0.7s est constituée d'une attaque de clochette. Les hautes fréquences ne sont pas harmoniquement liées aux basses fréquences sur cette attaque.

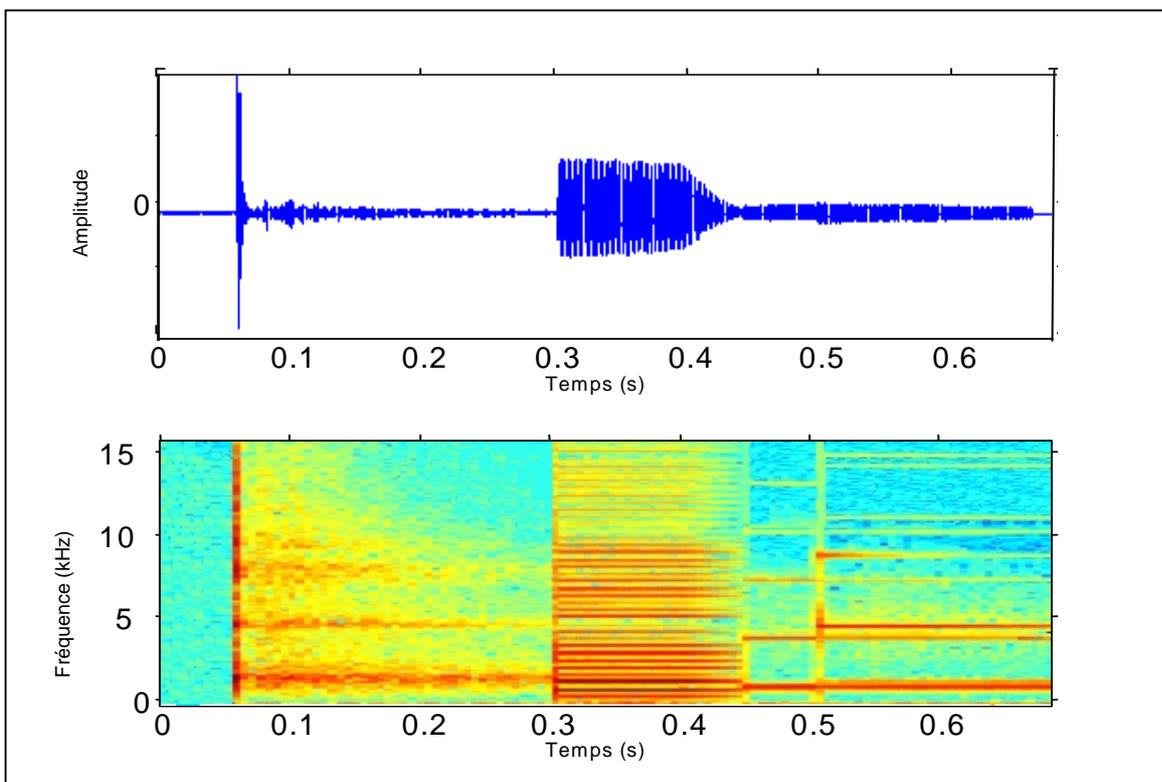


Figure 2.10 : S_{Trans}

La génération des contenus haute-fréquence sur les signaux transitoires peut donc se faire en utilisant les techniques présentées précédemment sur les signaux stationnaires. Une bonne précision temporelle sera toutefois nécessaire afin de ne pas perdre les localisations temporelles des attaques.

2.3.3.7. Conclusion

Cette série de signaux reflète la diversité, tant sur le plan temporel que fréquentiel, des signaux musicaux. Pour la majorité d'entre eux, la structure fine du spectre présente en basse-fréquence se retrouve en haute-fréquence.

Pour les signaux bruités (parole non-voisée) et les signaux purement harmoniques (instruments de musique), la structure fine est la même sur tout le spectre. On peut alors étendre efficacement la bande de ces signaux en dupliquant le spectre basse-fréquence vers les hautes fréquences, et en ajustant les hautes fréquences synthétisées par une enveloppe adéquate.

Pour les signaux plus complexes composés de bruit(s) et de peigne(s) harmonique(s), et sur les signaux de parole voisée notamment, la structure fine est également sensiblement la même sur tout le spectre, exception faite du rapport tonal sur bruit. L'enrichissement de spectre de tels signaux par duplication de spectre requiert alors des ajustements afin de contrôler ce rapport sur les hautes fréquences synthétisées.

Sur les signaux inharmoniques enfin, les tonales en haute-fréquence ne sont pas harmoniquement liées à celles présentes en basse-fréquence. Sur ce type de signaux, il semble donc difficile de trouver un modèle générique susceptible d'étendre efficacement la bande passante. Notons toutefois que cette catégorie de signaux reste peu fréquente puisque les signaux sonores sont en général composés soit de parole, soit de signaux musicaux complexes (mélange de $S_{\text{Multi_harm}}$ et de S_{Trans}).

Basé sur toutes ces constatations, il apparaît clairement qu'un procédé simple et générique pour étendre la bande passante de la plupart des signaux musicaux consiste en la succession de ces deux opérations:

- La translation spectrale de la structure fine du spectre
- L'ajustement d'énergie haute-fréquence par une enveloppe spectrale

2.4. Conclusion du chapitre

Les propriétés psychoacoustiques introduites dans ce chapitre offrent des informations essentielles pour les applications d'enrichissement de bande. Concernant la sensibilité de l'oreille en fonction de la fréquence, nous avons vu que la restitution des hautes fréquences requiert moins de précision que celles présentes en basse-fréquence; la sensation de hauteur devient par exemple approximative en haute-fréquence.

Concernant les corrélations entre les différentes parties du spectre, il apparaît que pour la majorité des signaux, les hautes et les basses fréquences ont un contenu spectral sensiblement de même nature.

Toute la stratégie des techniques d'enrichissement de spectre est basée sur ces deux constatations.

Contrairement aux techniques de codage "classique" (Codeur de parole par forme d'onde de type CELP, approche par transformée de type AAC, approche paramétrique de type HILN...), on ne tente pas ici de coder avec précision les hautes fréquences. L'idée novatrice des méthodes d'enrichissement de spectre consiste à extraire les informations contenues en basse-fréquence afin de les exploiter par duplication en haute-fréquence. Pour réaliser cette opération efficacement, on prendra soin :

- De limiter les phénomènes de dissonances, notamment lors de l'extension de bande des signaux harmoniques;
- De respecter le rapport tonal sur bruit en haute-fréquence;
- Sur les signaux transitoires, de ne pas générer de bruit avant les attaques afin de ne pas créer d'artefacts audibles (pré masquage temporel d'une durée de 5 ms);
- De respecter l'enveloppe du signal en haute-fréquence. Sur les signaux de parole notamment, on a vu que les sept premiers formants jouaient un rôle essentiel dans la qualité de restitution;

Concernant ce dernier point, on développe au chapitre suivant les différentes techniques d'estimation et de transmission d'enveloppe spectrale.

2.5. Bibliographie du chapitre 2

- [FLE 40] H. FLETCHER
Auditory patterns
Review of Modern Physics, 1940
- [JBI 99] JBIRA A.
Codage hiérarchique à faible retard
Thèse de l'école nationale supérieure des télécommunications, Février 1999
- [KLA 99] A. KLAPURI
Wide-band Pitch Estimation for Natural Sound Sources with Inharmonicities
106th Audio Engineering Society Convention, Munich, Allemagne, Mai 1999
- [KLE 95] W.B. KLEIJN & K.K. PALIWAL
Speech coding and synthesis
Elsevier, 1995
- [LEI 96] E. LEIPP
Acoustique et musique
Masson, quatrième édition, 1996
- [MAK 75] J. MAKHOUL
Linear prediction: A tutorial review
Proceeding of the IEEE, vol. 63, n°4, pp 561-580, Avril 1975
- [PLO 65] R. PLOMP & W. LEVELT
Tonal consonance and critical bandwidth
Journal of the Acoustical Society of America, vol. 38, 1965
- [SCH 85] M. R. SCROEDER, B.S. ATAL
Coded Excited Linear Prediction (CELP), High Quality Speech at Very Low Bit Rates
ICASSP, pp. 937-940, 1985
- [SPA 94] A.S. SPANIAS
Speech Coding, a tutorial review
Proceedings of the IEEE, Octobre 1994
- [ZWI 80] E. ZWICKER, E. TERHARDT
Analytical expressions for critical-band rate
Journal of the Acoustical Society of America 68(5), pp 1523-1525, Novembre 1980
- [ZWI 81] E. ZWICKER, E. FELDTKELLER
Psychoacoustique, l'oreille récepteur d'informations
Edition MASSON, 1981

CHAPITRE 3

MODELISATION DE L'ENVELOPPE SPECTRALE

Plan du chapitre

1.1.	INTRODUCTION.....	50
1.2.	TECHNIQUES D'EXTRAPOLATION D'ENVELOPPE SPECTRALE.....	52
1.3.	MODELISATION D'ENVELOPPE PAR PREDICTION LINEAIRE	55
1.4.	ESTIMATION D'ENVELOPPE DANS LE DOMAINE FREQUENTIEL	66
1.5.	CONCLUSION DU CHAPITRE.....	70
1.6.	BIBLIOGRAPHIE DU CHAPITRE 3.....	71

3.1. Introduction

L'enveloppe spectrale est essentielle dans les techniques d'enrichissement de spectre puisque, une fois la structure fine du spectre haute-fréquence étendue, c'est elle qui permet d'ajuster l'énergie du spectre en haute-fréquence (Figure 3.1) et de reconstituer le timbre et la couleur du "signal". Une bonne fidélité dans la restitution de cette enveloppe est incontournable pour reproduire les hautes fréquences avec fidélité.

Nous supposons ici que la structure fine du spectre haute-fréquence est déjà synthétisée (à partir du spectre basse-fréquence). Il ne reste donc plus qu'à remettre en forme l'enveloppe du signal synthétisé afin d'approcher au mieux l'allure du signal original.

Notons que dans le cadre de cette étude, l'estimation de l'enveloppe haute-fréquence peut être réalisée:

- Soit directement sur le signal à bande limitée sans connaissance du signal original pleine-bande (cas illustré Figure 3.1 sans transmission de paramètres). Ces techniques d'extrapolation ne nécessitent aucune transmission de paramètres du codeur au décodeur et s'apparente alors à un post-traitement. Après un bref état de l'art des méthodes existantes, nous étudierons, au paragraphe 3.2 leur efficacité sur les signaux génériques.
- Soit sur le signal original pleine-bande. La technique requiert alors une estimation d'enveloppe haute-fréquence au codeur sur le signal original et une transmission (à bas débit) des paramètres décrivant cette enveloppe (cas illustré Figure 3.1 avec transmission de paramètres). Nous développons au paragraphe 3.3 deux techniques complètes, l'une basée sur la prédiction linéaire, et l'autre basée sur une représentation en facteurs d'échelle dans le domaine fréquentiel.

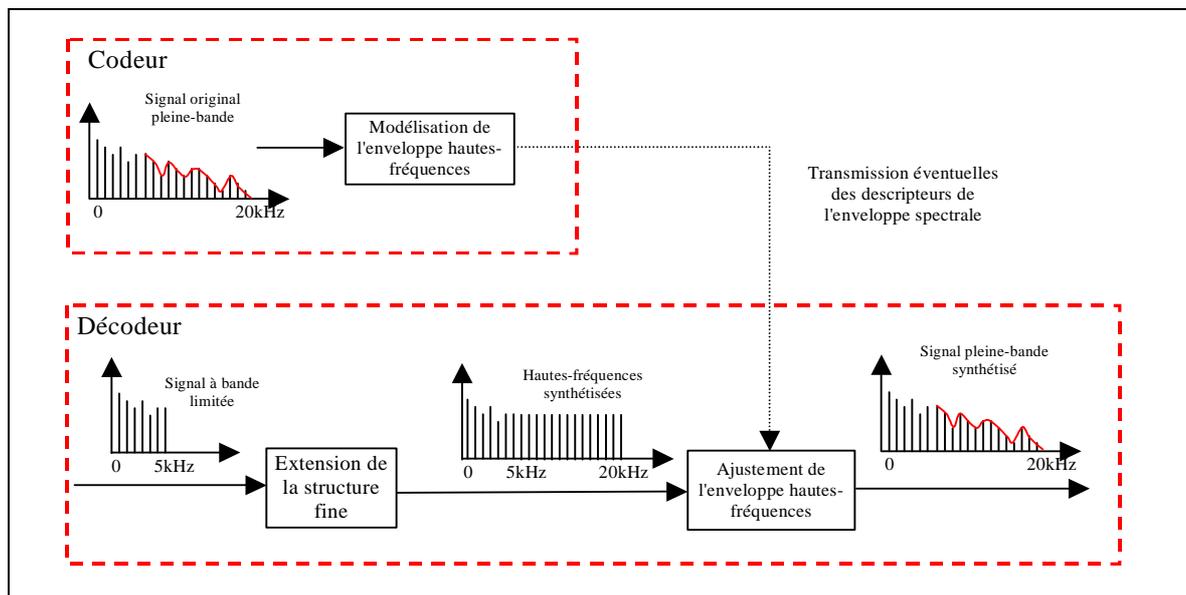


Figure 3.1 : Estimation, transmission et ajustement d'enveloppe spectrale

3.1.1. Remarques préalables sur la notion d'enveloppe

L'enveloppe que nous cherchons à modéliser ne doit pas trop varier en fréquence mais doit plutôt donner l'allure générale de la distribution de l'énergie du spectre. Pour les signaux de parole, et pour les signaux relativement bruités en général, les variations moyennes de l'énergie du spectre sont assez lentes et il est assez aisé d'en modéliser l'enveloppe. En revanche, sur des signaux à fortes composantes tonales, la notion d'enveloppe devient plus difficile à cerner, comme illustré sur la Figure 3.2 qui superpose le spectre d'un signal de parole chantée (Suzanne Vega), échantillonné à 24 kHz et différentes estimations d'enveloppes (problème également abordé sur le signal S_{Inharm} Figure 2.9).

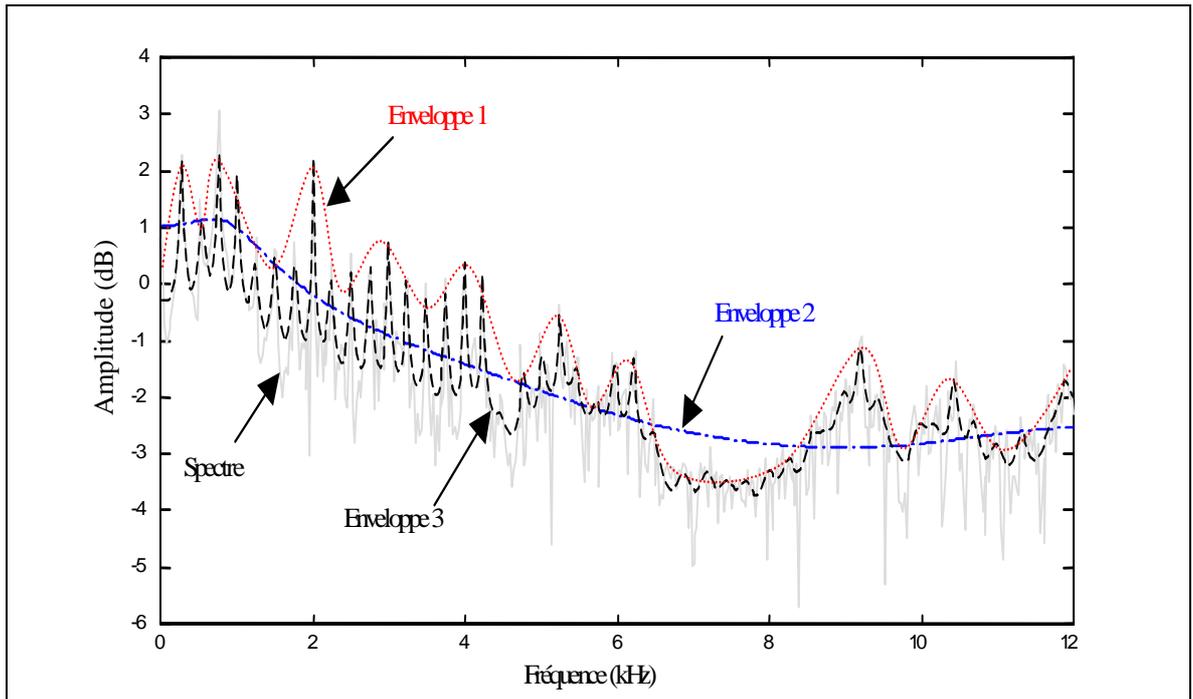


Figure 3.2 : Signal de parole et enveloppes associées

L'enveloppe 1 passe par les maximums du spectre et modélise correctement la structure formantique du signal. C'est vers cette enveloppe que nous tacherons de tendre dans la suite du document.

L'enveloppe 2 modélise plus "vaguement" l'énergie du spectre. Cette enveloppe n'épousant pas suffisamment la structure formantique du signal, la "couleur" du son n'est pas respectée.

L'enveloppe 3, plus précise, tend à "accrocher" la structure fine du spectre, et notamment les tonales présentes dans le signal.

3.2. Techniques d'extrapolation d'enveloppe spectrale

3.2.1. Principe

Ces techniques sont basées sur l'hypothèse que les formes d'enveloppes spectrales basses et hautes fréquences sont statistiquement "liées". L'idée consiste à extrapoler l'enveloppe haute-fréquence à partir de la connaissance seule de la forme de l'enveloppe basse-fréquence. Dans le domaine du codage audio, cette technique a l'avantage de ne requérir aucune transmission de donnée auxiliaire et s'apparente alors à un simple post-traitement. Elle comporte un intérêt pour certaines applications téléphoniques où aucun débit additionnel n'est possible.

3.2.2. Etat de l'art

L'extrapolation d'enveloppe a fait l'objet de nombreuses études sur la parole. Citons l'extrapolation d'enveloppe basée sur des dépendances statistiques entre les enveloppes basses et haute-fréquence [CHE 94] et [EPP 00], les techniques d'extrapolation par interpolation linéaire [MIE 00]. Nous présentons ci-dessous une méthode classique d'extrapolation d'enveloppe, celle par dictionnaire de formes d'ondes présentée dans [ABE 95].

3.2.3. Extrapolation d'enveloppe par dictionnaires de formes d'ondes

Le principe est de construire un dictionnaire générique de formes d'ondes des basses et des hautes fréquences. Ce dictionnaire est construit à partir d'une séquence d'apprentissage constituée de plusieurs heures de signaux de parole. A chacune des formes d'onde basse-fréquence, on associe une forme d'onde haute-fréquence.

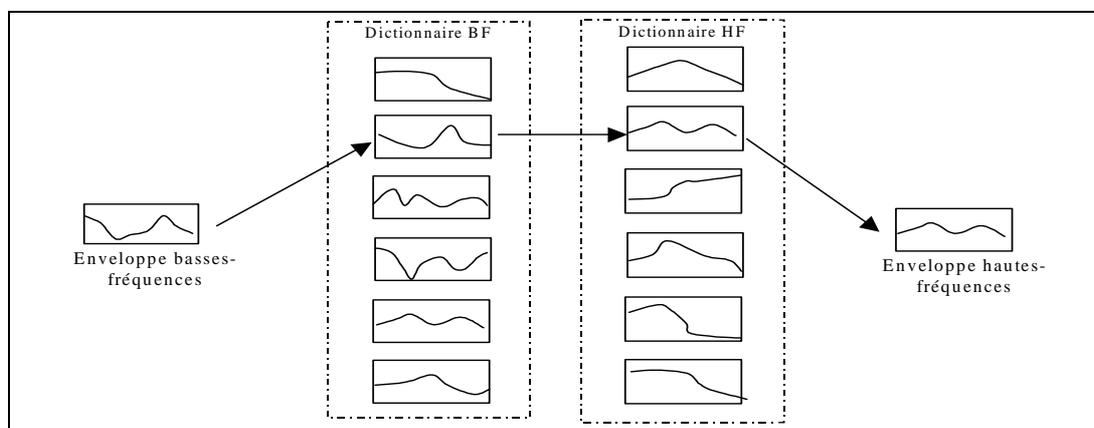


Figure 3.3 : Extrapolation d'enveloppe par dictionnaire de forme d'onde

La recherche de la forme d'onde haute-fréquence peut se faire :

- Soit directement en choisissant la meilleure approximation de l'enveloppe basse-fréquence dans le premier dictionnaire et en prenant l'enveloppe haute-fréquence associée.
- Soit en cherchant les n formes d'ondes les plus proches de la forme d'onde d'entrée et en combinant les n formes d'ondes associées pour trouver la forme d'onde en sortie. ("Codebook mapping with interpolation")

Notons enfin que selon la nature du signal de parole (voisé / non-voisé), on peut associer différents dictionnaires adaptés. ("Codebook mapping with split codebooks"). Cette adaptation se justifie par le fait que, sur les signaux de parole, l'énergie en haute-fréquence a tendance à décroître pour les sons

voisés contrairement aux sons non-voisés pour lesquels l'énergie est relativement importante dans les hautes fréquences ([YOS 94])

3.2.4. Application aux signaux de musique

Le signal représenté Figure 3.4 résulte de la superposition d'un signal de piano et d'attaques de cymbales). Le signal basse-fréquence, composé essentiellement du piano, est stationnaire dans le temps et la forme du spectre n'évolue que peu au cours du temps contrairement à l'énergie du signal haute-fréquence qui varie fortement selon la présence ou non des attaques de cymbales.

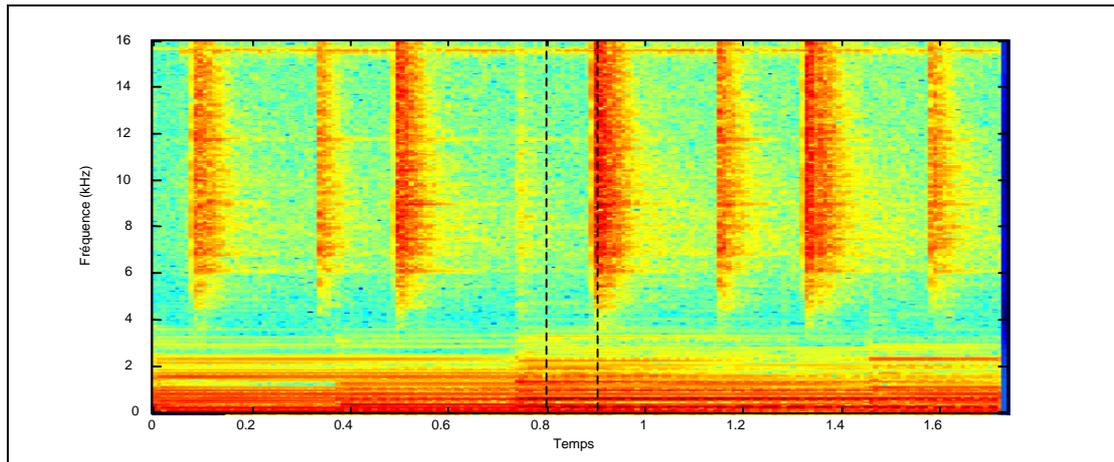


Figure 3.4 : Spectrogramme d'un signal composé d'un piano et de cymbale

La sélection de deux trames aux instants $t=0.8$ et $t=0.9$ s représentées Figure 3.5 met en évidence le problème de l'extrapolation de bande sur ce type de signal. Dans les fréquences basses inférieures à 4 kHz, les spectres sont en effet très corrélés, de même que les enveloppes associées. En revanche, dans les hautes fréquences, le comportement énergétique varie fortement selon la présence ou non du signal de cymbale. Les enveloppes associées sont de ce fait totalement décorrélées.

Les techniques d'extrapolation exposées précédemment deviennent dès lors totalement inefficaces puisque à une même enveloppe basse-fréquence peut correspondre une multitude d'enveloppes haute-fréquence. La fidélité à l'original ne peut être assurée « en aveugle ».

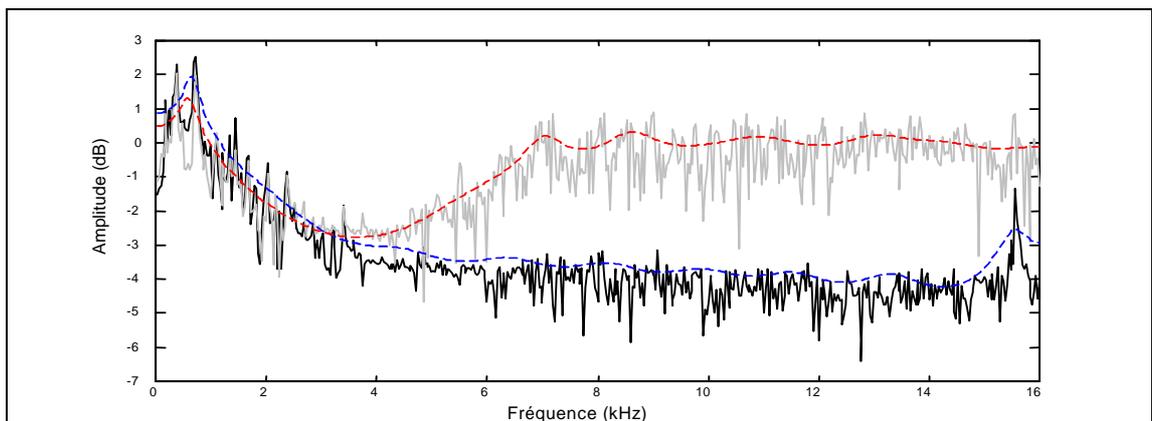


Figure 3.5 : Filtre d'enveloppe pour deux trames de 20 ms à $t=0.8$ et $t=0.9$ s

3.2.5. Conclusions

Les techniques d'extrapolation d'enveloppe spectrale ont l'avantage de ne requérir aucune transmission d'information mais donnent toutefois des résultats peu concluant sur la parole. La structure formantique haute-fréquence des signaux de parole est difficilement prédictible à partir de l'enveloppe basse-fréquence en bande téléphonique. [VAL 00]. L'extrapolation d'enveloppe permet d'étendre le spectre essentiellement au voisinage de la fréquence de coupure mais devient inefficace pour une largeur de bande plus élevée.

Sur les signaux musicaux, cette extrapolation est souvent irréalisable au vu de la diversité des enveloppes.

Des informations auxiliaires s'avèrent donc nécessaires pour transmettre l'enveloppe haute-fréquence dans le cas de sons génériques. Le paragraphe suivant étudie deux techniques de représentation et de transmission de cette information.

3.3. Modélisation d'enveloppe par prédiction linéaire

La prédiction linéaire, un outil fréquemment utilisé en codage de parole, a suscité de nombreuses études depuis une trentaine d'années. Particulièrement adaptée à la modélisation d'enveloppe spectrale des signaux de parole, nous en étudions dans ce chapitre les principes théoriques avant de nous intéresser à ses avantages et inconvénients dans le cadre de l'enrichissement de spectre des signaux musicaux.

Seront détaillés en particulier, la modélisation de l'enveloppe haute-fréquence sur le signal original, la quantification et la transmission (à bas débit) de ses descripteurs et enfin l'ajustement d'énergie des signaux par l'enveloppe transmise.

3.3.1. Principe de la prédiction linéaire

3.3.1.1. Introduction

Sur le système d'entrée/sortie modélisé Figure 3.6, le signal de sortie S_n s'écrit comme une combinaison linéaire des échantillons du signal de sortie observés aux P instants précédents, et des échantillons du signal d'entrée U_n observés à l'instant présent et aux Q instants précédents (Modèle ARMA, AutoRegressive Moving Average).

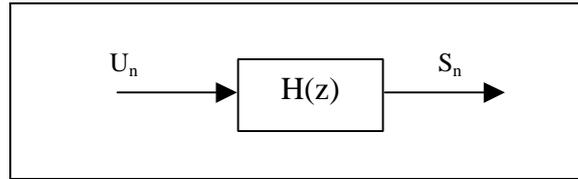


Figure 3.6 : Système d'entrée/sortie

$$S_n = \sum_{k=1}^P a_k \cdot S_{n-k} + G \cdot \sum_{l=0}^Q b_l \cdot U_{n-l}, \quad b_0 = 1 \quad (3.1)$$

où G et le couple $\{a_k\}, \{b_l\}$ représentent respectivement le gain et les coefficients du filtre H .

Dans le domaine fréquentiel, en désignant par $H(z)$ la fonction de transfert du système, l'équation (3.1) s'écrit :

$$H(z) = \frac{S(z)}{U(z)} = G \cdot \frac{1 + \sum_{l=1}^Q b_l \cdot z^{-l}}{1 - \sum_{k=1}^P a_k \cdot z^{-k}} \quad (3.2)$$

où

$$S(z) = \sum_{n=-\infty}^{\infty} S_n \cdot z^{-n} \quad (3.3)$$

Les racines du numérateur et du dénominateur sont respectivement les zéros et les pôles du modèle.

Dans le cas où $b_l = 0$, quel que soit $1 \leq l \leq Q$, $H(z)$ se réduit à un modèle tout-pôles, appelé modèle autorégressif. L'équation (3.1) devient dans ce cas :

$$S_n = \sum_{k=1}^P a_k \cdot S_{n-k} + G U_n \quad (3.4)$$

3.3.1.2. Modèle autorégressif

On tente d'approcher S_n avec les échantillons observés aux instants précédents.

Considérant la prédiction linéaire d'ordre P du signal S_n , le signal de prédiction, à l'instant n , s'écrit comme une combinaison linéaire des P échantillons passés :

$$\tilde{S}_n = \sum_{k=1}^P a_k \cdot S_{n-k} \tag{3.5}$$

Les coefficients $\{a_k\}$ sont appelés coefficients de prédiction.

La différence entre le signal S_n et sa valeur prédite \tilde{S}_n constitue l'erreur de prédiction, ou résidu :

$$e_n = GU_n = S_n - \tilde{S}_n = S_n - \sum_{k=1}^P a_k \cdot S_{n-k} \tag{3.6}$$

Le filtre, décrit en (3.6), est dit blanchisseur car l'opération de prédiction linéaire a pour conséquence de décorrélérer les valeurs de l'erreur de prédiction [MAK 75]. Le filtre blanchissant a pour transformée en z :

$$A(z) = 1 - a_1 z^{-1} + \dots + a_P z^{-P} = 1 - \sum_{k=1}^P a_k z^{-k} \tag{3.7}$$

En appliquant sur le résidu e_n le filtre inverse de fonction de transfert $\frac{1}{A(z)}$, on obtient le signal S_n .

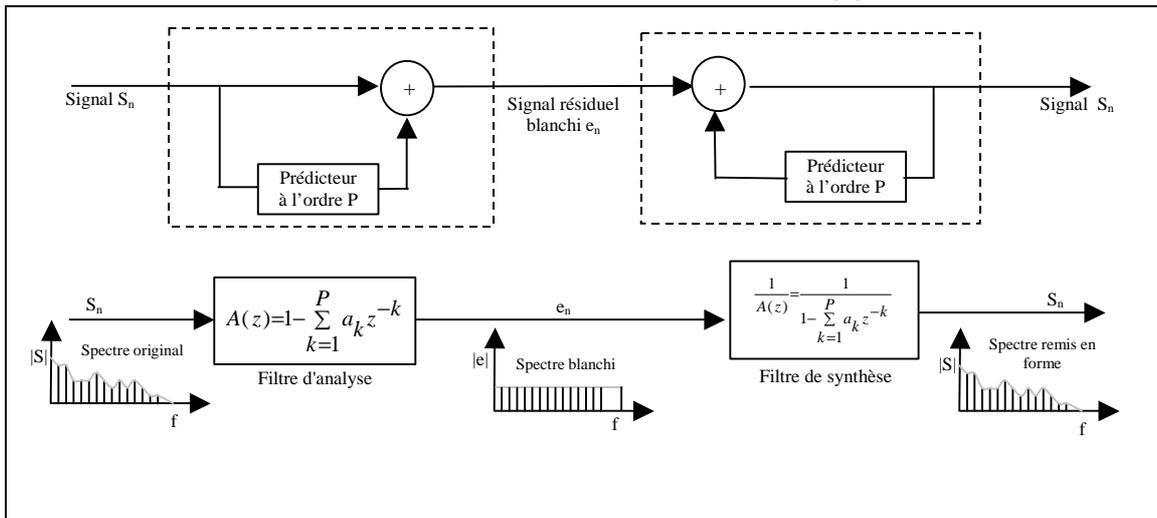


Figure 3.7 : Filtre blanchisseur et filtre de synthèse

3.3.1.3. Calcul des coefficients de prédiction

On distingue classiquement deux méthodes pour estimer les coefficients de prédiction :

- La méthode de l'autocorrélation
- La méthode de la covariance

3.3.1.3.1. Méthode de l'autocorrélation

Dans cette méthode, le signal S_n est multiplié par une fenêtre w_n . On obtient ainsi le signal fenêtré S'_n :

$$S'_n = w_n S_n \quad (3.8)$$

On minimise ensuite l'énergie du signal d'erreur E défini par :

$$E = \sum_{n=-\infty}^{\infty} e_n^2 = \sum_{n=-\infty}^{\infty} (S'_n - \sum_{k=1}^p a_k \cdot S'_{n-k})^2 \quad (3.9)$$

La recherche des coefficients $\{a_k\}$ se fait en minimisant E relativement aux coefficients $a_1 \dots a_p$.

En dérivant E par rapport aux coefficients $\{a_k\}$:

$$\frac{\partial E}{\partial a_k} = 0, \quad k = 1, \dots, p \quad (3.10)$$

On obtient les p équations suivantes :

$$\sum_{k=1}^p a_k \cdot \sum_{n=-\infty}^{\infty} S'_{n-i} S'_{n-k} = \sum_{n=-\infty}^{\infty} S'_n S'_{n-i}, \quad 1 \leq i \leq p \quad (3.11)$$

Dans les équations (3.11), on considère que les données sont nulles à l'extérieur de la fenêtre d'analyse w_n .

En définissant la fonction d'autocorrélation du signal fenêtré S'_n :

$$R(i) = \sum_{n=-\infty}^{\infty} S'_n S'_{n-i} = \sum_{n=i}^{N-1} S'_n S'_{n-i}, \quad 0 \leq i \leq p \quad (3.12)$$

où N représente la longueur de la fenêtre d'analyse, et en substituant les équations (3.12) aux équations (3.11), on obtient les équations suivantes :

$$\sum_{k=1}^p a_k R(i-k) = R(i), \quad 1 \leq i \leq p \quad (3.13)$$

Cette dernière équation peut s'écrire sous la forme matricielle :

$$\begin{bmatrix} R_0 & R_1 & \dots & R_{p-1} \\ R_1 & R_0 & \dots & R_{p-2} \\ \dots & \dots & \dots & \dots \\ R_{p-1} & R_{p-2} & \dots & R_0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_1 \\ \dots \\ a_p \end{bmatrix} = \begin{bmatrix} R_1 \\ R_2 \\ \dots \\ R_p \end{bmatrix} \quad (3.14)$$

Ce système matriciel se résout en tenant compte du fait que la matrice d'autocorrélation est une matrice de Toeplitz. Cette propriété permet de résoudre efficacement, c'est-à-dire sans inversion de la matrice $R(i)$, le système par l'algorithme de Levinson-Durbin décrit dans [MOR 95].

Cette propriété assure également que le filtre $A(z)$ est à phase minimale [HAY 96]. Dans le filtre de synthèse, $H(z) = \frac{1}{A(z)}$, les zéros de $A(z)$ deviennent les pôles de $H(z)$ et le fait que $A(z)$ soit à phase minimale garantit la stabilité du filtre de synthèse $H(z)$.

3.3.1.3.2. Méthode de la covariance

Dans la méthode de la covariance, on fenêtré le signal d'erreur (au contraire de la méthode de l'autocorrélation dans laquelle on fenêtré le signal S_n).

L'énergie E du signal s'écrit alors :

$$E = \sum_{n=-\infty}^{\infty} (e'_n)^2 = \sum_{n=-\infty}^{\infty} e_n^2 w_n \quad (3.15)$$

En dérivant E par rapport aux coefficients $\{a_k\}$, on obtient les p équations linéaires suivantes :

$$\sum_{k=1}^P \phi(i,k) \cdot a_k = \phi(i,0) \quad , \quad 1 \leq i \leq p \quad (3.16)$$

où $\phi(i,k)$ est la fonction de covariance du signal S_n définie par :

$$\phi(i,k) = \sum_{n=-\infty}^{\infty} w_n S_{n-i} S_{n-k} \quad (3.17)$$

Les équations (3.16) peuvent s'écrire sous la forme matricielle suivante :

$$\begin{bmatrix} \phi(1,1) & \phi(1,2) & \dots & \phi(1,P) \\ \phi(2,1) & \phi(2,2) & \dots & \phi(2,P) \\ \dots & \dots & \dots & \dots \\ \phi(P,1) & \phi(P,2) & \dots & \phi(P,P) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \dots \\ a_p \end{bmatrix} = \begin{bmatrix} \phi(1,0) \\ \phi(2,0) \\ \dots \\ \phi(P,0) \end{bmatrix} \quad (3.18)$$

Où $\phi(i) = \phi(i,0)$, $i = 1, 2, \dots, p$.

Cette matrice est symétrique mais les coefficients sur les diagonales ne sont pas égaux entre eux, à la différence de la matrice d'autocorrélation définie ci-dessus. La méthode de décomposition de Cholesky permet de résoudre ce système ([TAD 99]).

L'intérêt de la méthode de covariance est de ne faire aucune hypothèse concernant les données à l'extérieur de l'intervalle d'analyse [MOR 95], offrant ainsi une estimation spectrale plus fine ([CHO 98]). Cette méthode permet d'estimer plus précisément l'enveloppe spectrale, et de conserver une bonne précision temporelle car elle minimise l'erreur sur un intervalle fini. L'inconvénient est que contrairement à la méthode de l'autocorrélation, la stabilité du filtre tout-pôle n'est pas assurée.

3.3.2. Représentation des coefficients de prédiction

Les coefficients $\{a_k\}$, ou coefficients LPC (Linear Predictive Coding), offrent des caractéristiques qui ne facilitent pas le codage. Une faible erreur de quantification de l'un de ces paramètres entraîne de fortes variations dans le spectre restitué par l'ensemble du filtre et génère souvent des problèmes d'instabilité au filtre de synthèse. Les paramètres LSF (Line Spectral Frequencies), appelés encore paramètres LSP (Line Spectral Pair), possèdent en revanche des propriétés de quantification plus appropriées et permettent un meilleur contrôle de la stabilité. Notons que la conversion est réversible et sans perte. Elle est fournie en annexe B.

3.3.2.1. Quantification des coefficients LSP (Line Spectral Pair)

Les coefficients LSP sont liés à la position des pôles sur l'axe des fréquences et ont la propriété d'être rangés par ordre croissant; cet agencement des paramètres permet de prendre en compte des critères perceptifs et offre une propriété de codage efficace. La technique de quantification utilisée tout au long de cette thèse est une quantification vectorielle de type SVQ (Split Vector Quantization). Dans le cas d'une quantification d'un vecteur de P paramètres, la technique consiste à diviser le vecteur en L sous-vecteurs de dimension l_1, l_2, \dots, l_N avec $l_1+l_2+\dots+l_N = P$. Les sous-vecteurs sont ensuite quantifiés séparément.

L'exemple Figure 3.8 illustre une quantification sur 30 bits d'un vecteur de 18 LSP selon la technique SVQ. Le vecteur d'entrée de $P=18$ paramètres est subdivisé en $L=3$ sous-vecteurs de dimension respective $l_1=4, l_2=6$ et $l_3=8$.

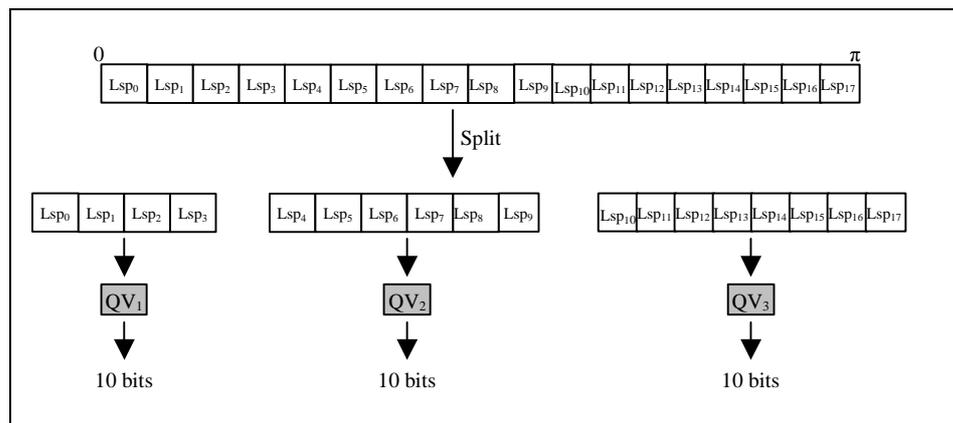


Figure 3.8 : Quantification SVQ (Split Vector Quantization)

Chacun de ces sous-vecteurs est associé à son plus proche représentant parmi 3 dictionnaires de LSP à 1024 entrées représentables sur 10 bits chacun. La génération de ces tables est réalisée par un algorithme classique de minimisation d'erreur, de type LBG [LIN 80], sur des bases d'apprentissages.

Cette sub-division non-uniforme permet de coder les LSP avec une précision non-uniforme sur l'axe des fréquences, prenant ainsi en compte les critères perceptifs (plus de précision dans les basses fréquences, là où l'oreille est plus sensible). Dans l'exemple ci-dessus, et pour un signal échantillonné à 32 kHz, les quatre premiers LSP compris entre 0 et $4\pi/18$ environ, soient entre 0 et 3.5 kHz se retrouvent ainsi quantifiés sur 2.5 bits chacun, alors que le dernier sous-vecteur correspondant à la bande haute-fréquence (entre $11\pi/18$ et π , soit entre 10 et 16 kHz environ) quantifie chaque LSP sur 1.25 bit.

De plus amples détails sur la quantification vectorielle des paramètres LSP sont fournis dans [KOV 97].

3.3.2.2. Lissage des coefficients LSP

De faibles variations de l'enveloppe spectrale entre deux trames consécutives peuvent entraîner une modification importante des coefficients LPC lors de l'analyse. A la synthèse, ces faibles variations peuvent engendrer des discontinuités temporelles lors de la synthèse du signal [KLE 95]. L'interpolation des filtres permet de résoudre ces problèmes de discontinuité. La qualité de restitution des signaux s'en trouve ainsi fortement améliorée sans exiger d'information additionnelle.

La technique consiste à interpoler linéairement les coefficients LSP calculés sur une trame de durée T de façon à les appliquer, à la synthèse, sur des sous-trames de durée plus faible. Ainsi pour deux trames d'analyse consécutives de 20ms, on interpolera avantagement les coefficients LSP afin d'appliquer les filtres de synthèse correspondants sur des trames de durée plus courte (typiquement de $T/8 = 2.5$ ms).

3.3.3. Ajustement d'enveloppe

Une fois les coefficients LSP transmis au décodeur, ils sont convertis en coefficients LPC (méthode de conversion décrite en annexe B). L'ajustement d'enveloppe est réalisé par filtrage du signal d'excitation (signal résiduel e_n spectralement blanchi décrit Figure 3.7) par le filtre de synthèse LPC déduit de ces coefficients, et de transformée en z :

$$\frac{1}{A(z)} = \frac{1}{1 - \sum_{k=1}^P a_k z^{-k}} \quad (3.19)$$

3.3.4. Applications à l'enrichissement de spectre

Nous développons dans ce paragraphe les ajustements nécessaires en terme de longueur de trame d'analyse et d'ordre de filtre afin d'adapter l'estimation d'enveloppe sur les signaux musicaux.

La représentation de l'enveloppe haute-fréquence par prédiction linéaire requiert également quelques ajustements puisqu'il ne s'agit plus ici de modéliser l'enveloppe du spectre complet mais seulement une partie de celui-ci.

3.3.4.1. Longueur de trame

La stationnarité des signaux de parole est de 10 à 20 ms et il est tout à fait standard de choisir des trames d'analyse de l'ordre de 20 ms ([MOR 95]), les paramètres LPC étant remis à jour toutes les 10 à 20 ms.

Concernant les signaux musicaux, la partie tenue (phase de maintien vue au paragraphe 1.4) possède une stationnarité d'une durée plus grande. L'énergie et l'enveloppe spectrale de ces signaux varient alors faiblement durant de longues périodes.

En revanche, pour les parties percussives des signaux (phases d'attaque), l'énergie varie en des temps très courts, de l'ordre de quelques millisecondes. Sur de tels signaux, l'enveloppe étant amenée à varier fortement, une prédiction linéaire sur une longue trame d'analyse donne des résultats erronés. Afin de bien modéliser le comportement temporel et fréquentiel des signaux transitoires, une solution consiste à diviser la trame en plusieurs sous-trames d'analyse de durée plus courte (de l'ordre de 5 ms) et à estimer l'enveloppe spectrale pour chacune de ces sous-trames. Cette technique modélise plus efficacement l'enveloppe des signaux transitoires mais requiert en contrepartie un surcoût d'information (transmission de plusieurs jeux de coefficients décrivant les différentes enveloppes spectrales) et donc un accroissement du débit.

3.3.4.2. Ordre de prédiction

L'ordre de prédiction linéaire à utiliser sur les signaux de parole est fonction du nombre de formants présents dans la bande passante du signal. Pour un signal échantillonné à la fréquence f_s , Smith ([SMI 98]) approche l'ordre de prédiction par la formule :

$$\frac{f_s}{1000} \leq P \leq \frac{f_s}{1000} + 4 \quad (3.20)$$

Ainsi pour une fréquence d'échantillonnage de 16 kHz, un ordre LPC compris entre 16 et 20 est approprié pour des signaux de parole.

Le Table 3.1 donne l'ordre de prédiction utilisé et le débit requis par le CELP, et ce pour les configurations les plus courantes ([NIS 99] et [SPA 94]).

Type de codeur	Bande passante	Ordre du filtre	Débit
CELP bande étroite	3.5 kHz	10	21 bits / 20 ms
CELP bande élargie	7 kHz	16	32 bits / 20ms

Table 3.1 : Ordre de prédiction utilisé en codage de parole

Pour les signaux musicaux stationnaires et bruités (signaux de parole en général et signaux résultant du mélange de plusieurs instruments et/ou de parole), un ordre de 20 modélisera efficacement l'enveloppe haute-fréquence (bande comprise entre 6 et 15 kHz environ).

Pour les signaux musicaux composés de fortes composantes tonales en revanche, l'estimation d'enveloppe par prédiction linéaire demande quelques ajustement. Prenons l'exemple de $S_{\text{Mono_harm}}$, signal harmonique de fréquence fondamentale 440 Hz et échantillonné à 16 kHz, dont le spectre est repris Figure 3.9. Pour une telle fréquence d'échantillonnage, Smith conseille un ordre de prédiction d'environ 18 (formule (3.20)).

On superpose au spectre représenté Figure 3.9 les filtres d'enveloppe aux ordres 10 et 18.

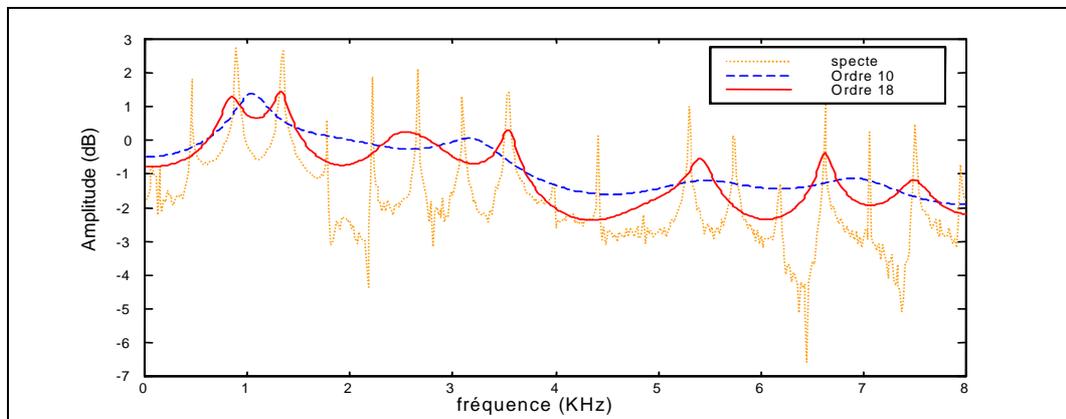


Figure 3.9 : Estimation d'enveloppe sur $S_{\text{Mono_harm}}$

Une prédiction linéaire à l'ordre 10 modélise correctement l'énergie moyenne du signal. En revanche, une prédiction à l'ordre 18 a tendance à "accrocher" certaines tonales et l'enveloppe retombe entre les pics. Ce phénomène est d'autant plus accentué lorsque le nombre de tonales dans le signal est proche de la moitié de l'ordre de prédiction du filtre LPC.

La prédiction linéaire étant basée sur un critère de minimisation énergétique, elle ne peut différencier les pics des formants des tonales. En augmentant l'ordre de prédiction P , l'erreur définie en (3.9) décroît et la puissance spectrale LPC définie dans [MAK 75]

$$\hat{P}(\omega) = |H(e^{j\omega})|^2 = \frac{G^2}{|A(e^{j\omega})|^2} \quad (3.21)$$

où G représente le gain, décrit de plus en plus précisément la structure fine du spectre (spectral matching).

Dans le cadre de notre étude, ce phénomène "d'accrochage" des harmoniques génère des problèmes lors de la mise en forme du signal par le filtrage inverse. Ce problème sera plus amplement développé au chapitre 5.2.5.2 lors de l'étude du comportement de la technique complète d'extension de bande sur les signaux composés de fortes tonales.

Un ajustement de l'ordre de prédiction en fonction de la nature du signal est donc nécessaire. Pour les signaux bruités et/ou composés de nombreuses tonales, un ordre de prédiction élevé modélisera efficacement la structure formantique. En revanche pour les signaux composés de quelques tonales, on limitera l'ordre du filtre afin d'éviter de modéliser la structure fine du spectre.

3.3.4.3. Modélisation de l'enveloppe haute-fréquence

Dans le contexte de l'enrichissement de spectre, on cherche à modéliser l'enveloppe d'une partie du spectre seulement et à transmettre les descripteurs correspondants à faible débit. Cette modélisation particulière requiert alors quelques ajustements.

On représente, respectivement de haut en bas Figure 3.10, le spectre du signal S_{noise} échantillonné à 32 kHz et ce même signal mais filtré entre 5 et 13 kHz (filtre passe-bande). On réalise sur ces deux signaux une prédiction linéaire à l'ordre 20. Les lignes verticales représentées Figure 3.10 indiquent la position des coefficients LSP calculés à partir des coefficients LPC.

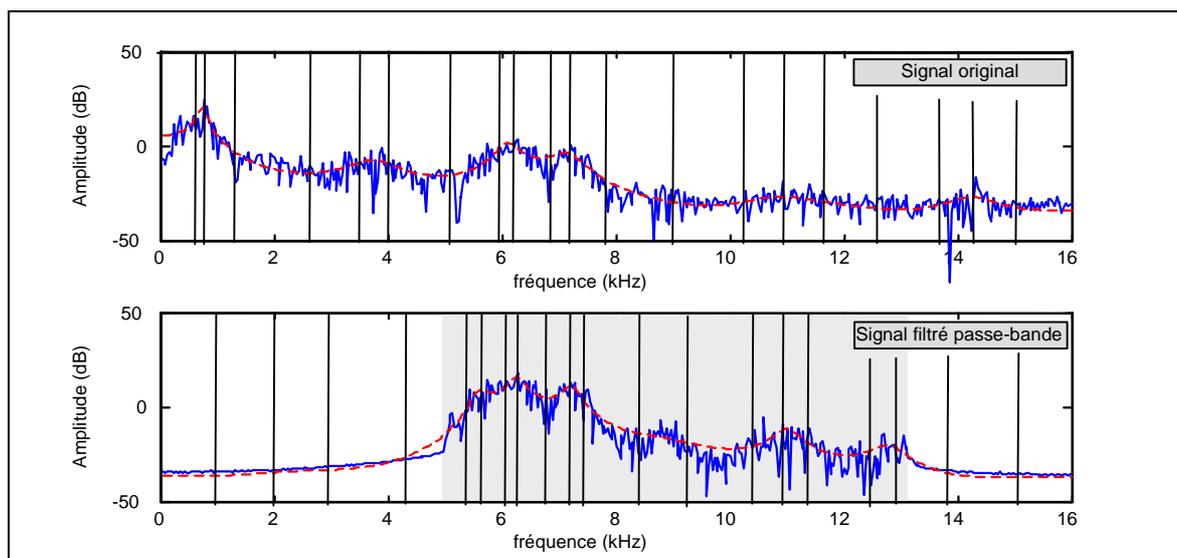


Figure 3.10 : Prédiction linéaire sur un signal passe-bande

On cherche sur cet exemple à modéliser l'enveloppe (courbe en pointillée de la partie grisée) du spectre haute-fréquence de bande passante [5-13kHz].

Le problème de l'estimation d'enveloppe sur un tel signal est que les coefficients LSP, au lieu d'être concentrés sur la bande passante du signal, se retrouvent répartis sur tout l'axe des fréquences. On se retrouve ainsi avec 6 coefficients LSP dans les bandes de fréquence d'énergie nulle, c'est-à-dire en dehors de la bande [5-13kHz] visée. L'estimation, et surtout la transmission de l'enveloppe devient dès lors sous-optimale.

Il est important de noter qu'il est délicat de ne pas transmettre ces pôles "indésirables" (pôles en dehors de la bande de fréquence à modéliser) car la suppression ou la modification de ceux-ci influent sur tout le filtre.

Deux techniques ont été développées afin d'optimiser l'estimation du filtre d'enveloppe de tels signaux à bande limitée

3.3.4.3.1. Prédiction linéaire sélective

Etant donné $P(\omega)$, la densité spectrale de puissance du signal S_n ,

$$P(\omega) = |S_n(e^{j\omega})|^2 \quad (3.22)$$

la prédiction linéaire sélective ([MAK 75] & [MAR 76]) permet de modéliser une partie du spectre $P(\omega)$ comprise entre $\omega_1 \leq \omega \leq \omega_2$ par un modèle tout pôle.

Afin de modéliser $P(\omega)$ dans la région $\omega_1 \leq \omega \leq \omega_2$ par $\hat{P}(\omega)$, on réalise un changement de variable faisant ainsi coïncider la région $\omega_1 \leq \omega \leq \omega_2$ à la bande de fréquence complète comprise entre 0 et π :

$$\omega' = \frac{\pi(\omega - \omega_1)}{\omega_2 - \omega_1} \quad (3.23)$$

Les coefficients de prédiction sélective sont calculés en résolvant les équations normales définies en (3.13):

$$\sum_{k=1}^P a_k R(i-k) = R(i), \quad 1 \leq i \leq P \quad (3.24)$$

Avec

$$R(k) = \frac{1}{\omega_2 - \omega_1} \int_{\omega_1}^{\omega_2} P(\omega) \cos(k\omega) d\omega \quad (3.25)$$

La prédiction linéaire sélective est utilisée en codage de parole dans le but de modéliser les différentes parties spectrales à des ordres de prédiction différents (ordre de prédiction élevé en basse-fréquence et modélisation plus grossière en haute-fréquence). La technique revient en fait à translater la bande de signal à traiter en bande de base (technique décrite ci-dessous).

3.3.4.3.2. Translation du signal en bande de base

La modélisation d'enveloppe par prédiction linéaire étant sous-optimale sur des signaux non pleine-bande, l'idée consiste à translater la partie spectrale à analyser en bande de base et à sous-échantillonner le signal résultant, de façon à obtenir un signal pleine-bande.

Reprenons l'exemple du signal représenté Figure 3.10. On cherche ici à modéliser l'enveloppe du signal compris entre 5 et 13 kHz. Pour ce faire, on translate la bande [5-13kHz] en [0-8kHz]. On sous-échantillonne ensuite le signal translaté à une fréquence de 16 kHz. On évite ainsi les problèmes rencontrés Figure 3.10 puisque la prédiction linéaire est réalisée uniquement sur la bande de signal visée.

3.3.5. Mise en forme spectrale du signal haute-fréquence

Après avoir étudié les ajustements requis lors de l'estimation d'enveloppe appliquée aux techniques d'enrichissement de spectre, nous nous intéressons maintenant à la remise en forme spectrale des signaux.

Dans les applications de codage audio classiques utilisant une méthode de modélisation d'enveloppe par prédiction linéaire, le signal résiduel, défini équation (3.6), est modélisé et codé par le codeur avant d'être transmis au décodeur où il sert d'excitation au filtre de synthèse d'enveloppe (défini équation (3.19)). Il est important de rappeler que ce signal résiduel est spectralement blanchi⁵.

⁵ Signal blanchi : on désigne par ce terme un signal ayant subi une démodulation d'enveloppe.

C'est ainsi que le blanchiment d'un signal bruité conduit à un bruit blanc, et que le blanchiment d'un signal harmonique conduit à un signal dont les harmoniques sont toutes de même amplitude.

Dans le cadre de notre application, le débit imposé ne permet pas de transmettre ce résidu. Ce dernier est en fait synthétisé à partir du signal basse-fréquence et peut de ce fait revêtir des formes d'enveloppe des plus variées. On parlera alors de signaux à spectre coloré, par opposition au spectre blanc défini plus haut.

Afin de réaliser une remise en forme spectrale efficace, c'est-à-dire afin d'obtenir une enveloppe spectrale proche de celle de l'original, il est donc nécessaire de blanchir le signal synthétisé au décodeur, avant de le remettre en forme par l'enveloppe transmise.

Ce principe est illustré Figure 3.11.

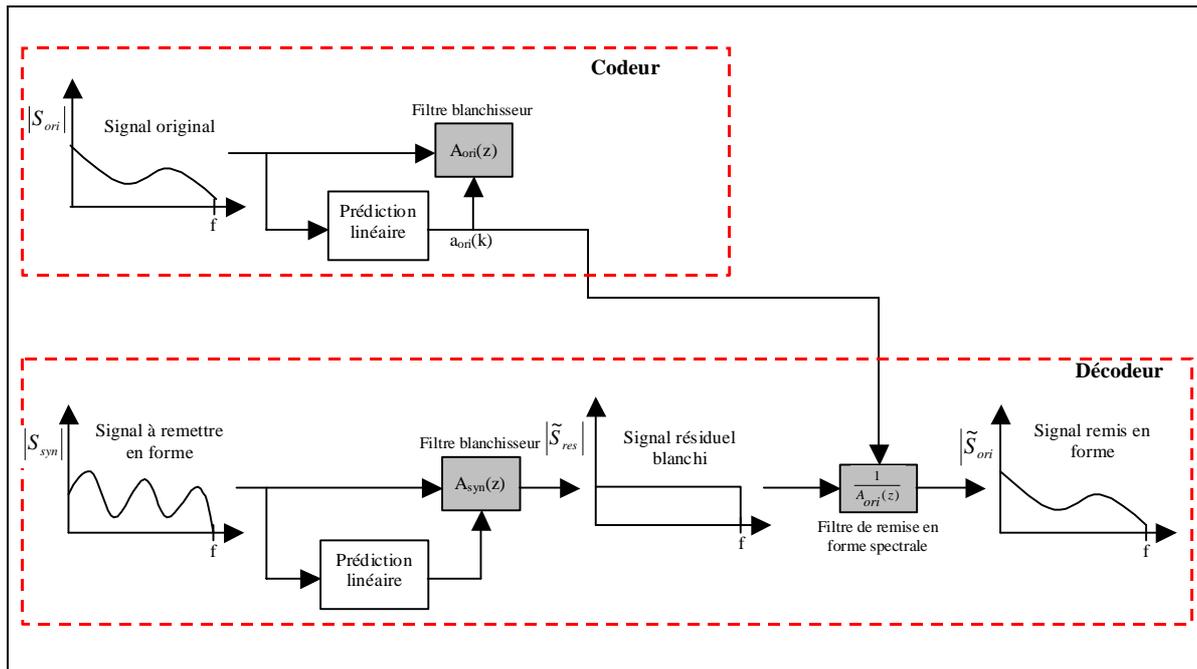


Figure 3.11 : Nécessité de blanchir le signal d'excitation avant la remise en forme spectrale

Dans le but de simplifier le schéma, on réalise ici une estimation et un ajustement d'enveloppe d'un signal en bande de base. Il est à noter que dans une technique d'enrichissement de spectre complète, l'estimation de l'enveloppe sur le signal original et l'ajustement d'enveloppe sur le signal synthétisé sont réalisés sur la partie haute-fréquence.

Le principe d'estimation et d'ajustement d'enveloppe est le suivant :

Au codeur, une prédiction linéaire sur le signal original fournit les coefficients du filtre $a_{ori}(k)$.

Au décodeur, on réalise une seconde prédiction linéaire sur le signal à remettre en forme (de forme spectrale quelconque différente de celle du signal original). On obtient ainsi les coefficients $a_{syn}(k)$.

En injectant le signal S_{syn} dans le filtre d'analyse de fonction de transfert $A_{syn}(z)$, défini par les coefficients $a_{syn}(k)$, on réalise un blanchiment spectral du signal S_{syn} et on obtient en sortie du filtre le signal résiduel spectralement blanc \tilde{S}_{res} .

En injectant ensuite ce signal \tilde{S}_{res} dans le filtre de synthèse $\frac{1}{A_{ori}(z)}$, calculés à partir des coefficients $a_{ori}(k)$ transmis par le codeur, on remet spectralement en forme le signal résiduel \tilde{S}_{res} . En sortie du filtre de synthèse $\frac{1}{A_{ori}(z)}$, on obtient ainsi le signal \tilde{S}_{ori} , d'enveloppe spectrale proche de celle du signal original.

Afin d'assurer une bonne fidélité dans la reconstruction de l'enveloppe spectrale, deux opérations de prédictions linéaires sont ainsi nécessaires :

- Une première prédiction linéaire, réalisée sur le signal original, dans le but de déterminer le filtre modélisant l'enveloppe spectrale haute-fréquence du signal original.
- Une seconde prédiction linéaire, réalisée sur le signal basses-fréquence à remettre en forme au décodeur, dans le but de blanchir préalablement ce signal qui sera remis en forme par le premier filtre.

3.3.6. Conclusion sur la modélisation d'enveloppe par prédiction linéaire

La prédiction linéaire est un outil largement répandu et utilisé en codage de parole. Cette technique a fait et fait encore aujourd'hui l'objet de nombreuses études. Les techniques d'estimation des filtres, de quantification des paramètres sont fiables, précises et robustes.

Les considérations développées dans ce chapitre nous amèneront à l'élaboration d'un premier schéma de codage susceptible de modéliser, de transmettre et d'ajuster efficacement l'enveloppe spectrale haute-fréquence des signaux génériques.

Pour la modélisation de l'enveloppe haute-fréquence du signal original :

- On prendra soin de translater la partie à modéliser en bande de base afin de réaliser une estimation d'enveloppe efficace.
- On utilisera un ordre de prédiction variable afin de prendre en compte la pluralité des formes d'ondes que peuvent revêtir les signaux musicaux. Sur les signaux bruités et les signaux de parole en particulier, un ordre de 20 modélisera efficacement la structure formantique de la bande haute-fréquence comprise entre 5 et 15 kHz environ. Sur les signaux composés de fortes composantes tonales en revanche, et sur les signaux harmoniques en particulier, on limitera l'ordre de prédiction afin de ne pas "accrocher" les tonales.
- Sur les signaux musicaux stationnaires et sur les signaux de parole en général, on choisira des fenêtres d'analyse LPC de l'ordre de 20 millisecondes. Sur les signaux transitoires en revanche, il sera nécessaire d'adapter la technique de modélisation et d'ajustement d'enveloppe afin de préserver une résolution temporelle suffisante (de l'ordre de 5 ms).

Concernant la transmission des coefficients du filtre d'enveloppe,

- Un passage des coefficients de prédiction a_k dans le domaine LSP offrira une meilleure stabilité et des propriétés de lissage intéressantes lors de l'ajustement d'enveloppe par le filtre de synthèse.
- En terme de coût de transmission, la conversion des coefficients du filtre en paramètres LSP et une quantification vectorielle de ceux-ci permettront de réduire fortement le débit requis pour transmettre les informations.

Concernant la mise en forme spectrale du signal haute-fréquence synthétisé :

- On prendra soin de blanchir le signal synthétisé avant de le remettre en forme par le filtre d'enveloppe transmis. Ce blanchiment préalable du signal est incontournable afin d'obtenir une forme d'enveloppe synthétisée proche de celle du signal original.

3.4. Estimation d'enveloppe dans le domaine fréquentiel

Nous étudions dans ce paragraphe une seconde méthode de modélisation d'enveloppe et de remise en forme spectrale. Contrairement à la technique de prédiction linéaire décrite précédemment qui consiste à modéliser l'enveloppe des signaux par un filtre AR, le principe consiste ici à réaliser une transformée temps/fréquences sur les signaux afin de modéliser et d'ajuster l'énergie du spectre directement dans le domaine fréquentiel.

Le document [SCH 97] passe en revue différentes méthodes développées dans la littérature (techniques basées sur le cepstre, sur le cepstre discret, techniques géométriques et technique HRMP (High Resolution Matching Pursuit)). Toutes ces méthodes sont dédiées à des applications d'analyse et de synthèse sonore; elles donnent des résultats précis mais requièrent en contrepartie un nombre de descripteurs élevés et donc un débit non négligeable dans un cadre de transmission.

En haute-fréquence, l'oreille est moins sensible aux variations fines de l'énergie du spectre et nous développons dans ce chapitre une méthode de modélisation d'enveloppe plus adaptée aux applications d'enrichissement de spectre à bas débit, la modélisation par facteurs d'échelle. L'estimation d'enveloppe et l'ajustement d'énergie sont réalisés directement dans le domaine transformé à l'aide de facteurs d'échelle correspondant aux énergies du spectre en sous-bandes.

3.4.1. Principe

Prenons l'exemple de l'ajustement d'énergie haute-fréquence du signal S_{Noise} décrit au paragraphe 2.3.3.2. Afin d'illustrer la technique et d'en étudier ses avantages et inconvénients, on suppose ici que la structure fine non-transmise est synthétisée par un générateur de bruit. Le diagramme de fonctionnement est décrit Figure 3.12.

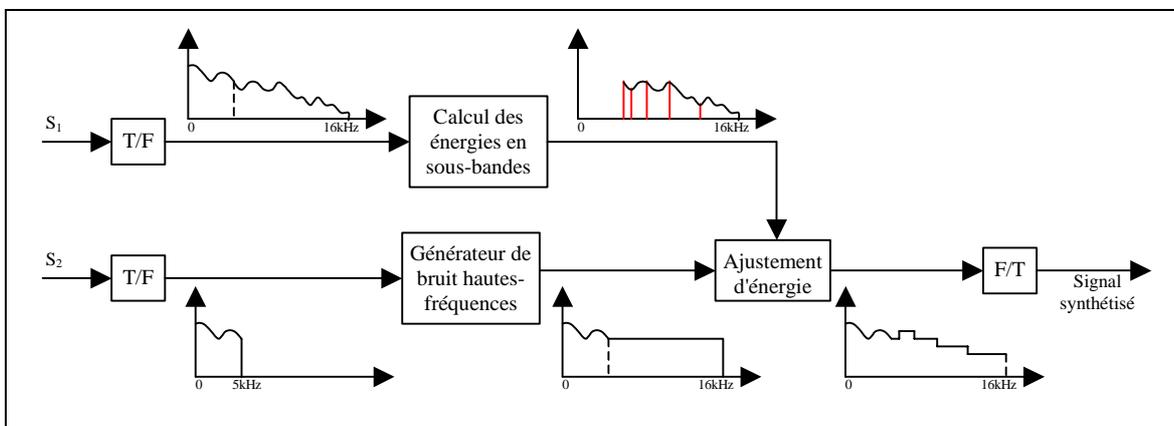


Figure 3.12 : Ajustement d'enveloppe par facteurs d'échelle dans le domaine fréquentiel

Le signal original pleine bande S_1 et le signal à bande limitée S_2 sont segmentés en trames de quelques dizaines de millisecondes. On réalise une transformée temps/fréquence sur chacune des trames puis on calcule les énergies correspondantes sur N sous-bandes haute-fréquence du spectre original. On ajuste ensuite les hautes fréquences synthétisées sur le signal à bande limitée par les facteurs d'échelles transmis.

3.4.2. Avantages de l'estimation d'enveloppe dans le domaine fréquentiel

3.4.2.1. Choix de la mise en œuvre

Divers types de transformées temps/fréquences sont envisageables pour ce type d'application. Nous nous sommes concentrés durant ces trois années de thèse sur deux transformées particulières, la MDCT (Modified Discrete Cosine Transform) et la DFT (Discrete Fourier Transform). On trouvera une description succincte de ces transformées en annexe A. Pour de plus amples informations sur la MDCT, on se référera à [DER 00].

3.4.2.2. Précision de l'enveloppe

La précision de l'enveloppe spectrale est régie par le nombre et la largeur des sous-bandes. Selon la précision de l'enveloppe requise, et selon le débit disponible, cette méthode est facilement adaptable et offre de nombreuses possibilités dans la manière de modéliser l'enveloppe.

3.4.2.3. Prise en compte des critères perceptifs

La modélisation d'enveloppe par facteurs d'échelle dans le domaine fréquentiel est particulièrement intéressante d'un point de vue psychoacoustique puisqu'elle permet de prendre en compte la sensibilité de l'oreille selon l'axe fréquentiel. On décomposera ainsi avantageusement cet axe en sous-bandes selon une échelle Bark (paragraphe 2.2.2). Cela aura pour effet de modéliser plus finement les sous-bandes basse-fréquence. On se contentera d'une modélisation plus grossière, et donc de sous-bandes plus larges vers les plus hautes fréquences, là où l'oreille est moins sensible au contenu et aux variations d'énergie du spectre.

3.4.2.4. Choix de la longueur de bande à synthétiser

Cette méthode de modélisation d'enveloppe s'adapte également facilement quelle que soit la longueur de bande à synthétiser, et quelle que soit la fréquence de coupure du signal à bande limitée. Cet aspect présente un intérêt certain dans le cadre de l'enrichissement de spectre de signaux codés par un codeur par transformée (de type AAC par exemple); la fréquence de coupure de tels signaux est en effet amenée à varier au cours du temps.

3.4.3. Ajustements en fonction de la nature des signaux

Le signal représenté Figure 3.13 correspond à la succession d'une période de silence et du signal S_{Noise} décrit Figure 2.6. Le premier signal est le signal original pleine-bande de référence échantillonné à 32 kHz; le second est généré à partir de la version à bande limitée (5 kHz) de l'original selon la méthode d'enrichissement de spectre décrite Figure 3.12. On utilise ici une transformée MDCT avec des fenêtres d'une longueur de 512 échantillons.

L'ajustement d'enveloppe sur la partie transitoire génère du bruit sur deux fenêtres d'analyse/synthèse, soit sur 1024 échantillons. Après l'attaque, ce phénomène est en général imperceptible car le bruit est masqué par l'attaque (post-masquage temporel de plusieurs centaines de millisecondes). Le pré-masquage est en revanche beaucoup plus bref, de l'ordre de 5ms et il apparaît des phénomènes de pré-écho bien connus dans les codeurs par transformée. L'ajustement d'enveloppe dans le domaine MDCT requiert dès lors une gestion des fenêtres d'analyse/synthèse au cours du temps en fonction de la stationnarité du signal afin de contrôler le phénomène d'étalement du bruit.

La stratégie adoptée est celle utilisée par le codeur AAC et consiste à utiliser des fenêtres longues de l'ordre de 32ms pour les périodes stationnaires du signal et des fenêtres 8 fois plus courtes pour les transitoires. Pour plus de détails dans la gestion des fenêtres, on se référera à [FER 96].

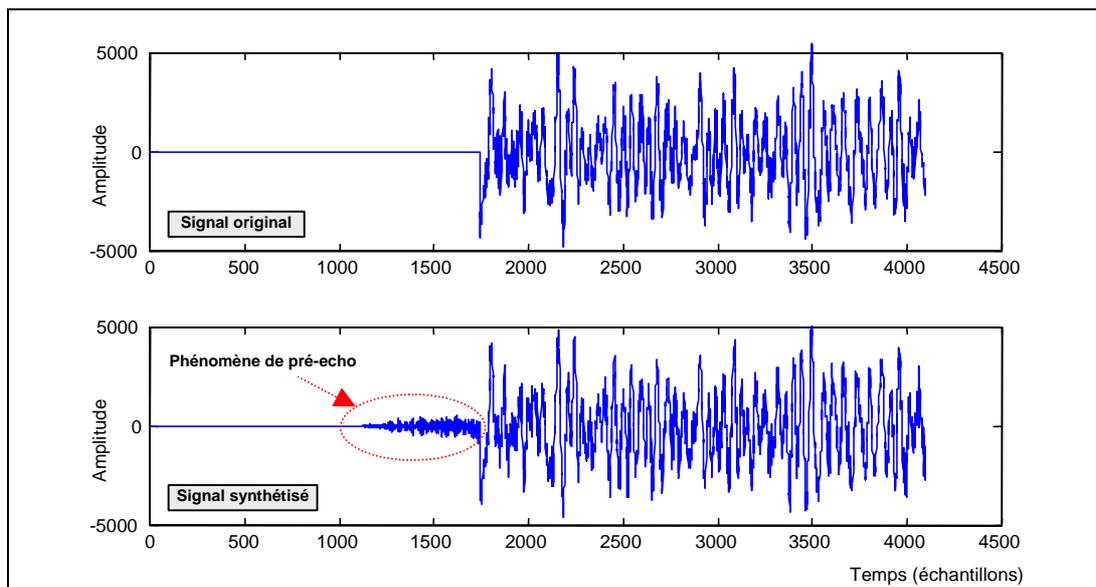


Figure 3.13 : Ajustement d'enveloppe par MDCT d'un signal transitoire

Cette approche atténue les phénomènes de pré-écho mais accroît en contrepartie le nombre de descripteurs d'enveloppe sur les signaux transitoires. Les N facteurs d'échelles décrivant l'enveloppe spectrale dans le cas stationnaire passent à $8*N$ facteurs d'échelles sur les signaux transitoires (pour une même résolution fréquentielle).

3.4.4. Quantification et coût de transmission des facteurs d'échelle

Une fois les facteurs d'énergies en sous-bandes calculés, il reste à les transmettre, à bas débit. On utilise pour ce faire un schéma de compression composé des trois étapes suivantes :

- Une première étape de quantification des énergies. On réalise ici une quantification scalaire logarithmique de chacun des facteurs sur 7 bits, couvrant ainsi une dynamique de ± 128 dB par pas de 1 dB. Notons que ce pas de 1 dB, déduit du JND temporel défini au paragraphe 2.2.4.3, assure une continuité d'énergie du signal imperceptible.
- Une seconde étape de codage différentiel des valeurs d'énergie quantifiées. Le codage différentiel est basé sur l'observation que des données successives sont fortement corrélées. Il est alors plus judicieux d'encoder non pas les données elles-mêmes mais la différence entre deux données successives. Dans notre cas, ce codage différentiel se justifie par le fait que pour la majorité des signaux musicaux, les facteurs d'énergie représentatifs de l'enveloppe spectrale sont effectivement corrélés [JAY 84]. L'enveloppe spectrale correspond en effet à une version lissée du spectre et les grandes variations d'énergies sont rares.
- Une troisième étape de compression des énergies différentielles quantifiées. On utilise pour ce faire un algorithme de codage entropique⁶ (codage sans perte) [HUF 52].

⁶ Le codage entropique exploite les propriétés statistiques des coefficients quantifiés pour diminuer le débit de transmission en utilisant des mots courts pour représenter les événements les plus probables et des mots plus longs pour les occurrences rares.

3.4.5. Conclusion

La modélisation d'enveloppe par facteurs d'échelles dans le domaine fréquentiel présente plusieurs degrés de liberté :

- Sur le choix de la fréquence de coupure du signal à bande limitée et de la largeur de bande à synthétiser. La technique offre une grande souplesse puisqu'elle est facile à adapter quelle que soit la bande à transmettre.
- Sur la précision de l'enveloppe requise. Le fait de travailler directement dans le domaine fréquentiel permet de choisir la précision de l'enveloppe en jouant sur le nombre et sur la largeur des sous-bandes.

Afin de limiter l'étalement du bruit lors de l'ajustement d'énergie du spectre des signaux transitoires, on diminuera la taille des fenêtres de pondération de la transformée temps/fréquences.

En terme de coût de transmission des descripteurs, on compressera efficacement les données à transmettre en adoptant une stratégie de compression basée sur une quantification scalaire, un codage différentiel suivi d'un codage entropique des facteurs d'échelle.

En terme de complexité enfin, le coût correspond à celui de la transformée temps/fréquence et reste de ce fait faible, l'optimisation des transformées étant souvent basée sur des FFT.

3.5. Conclusion du chapitre

Vu la diversité et la pluralité des modes de production des signaux musicaux, il n'existe que peu de corrélations entre les enveloppes hautes et basse-fréquence. Les techniques d'extrapolation d'enveloppe sans transmission d'information s'avèrent dès lors inefficaces.

Deux méthodes d'estimation d'enveloppe et de remise en forme spectrale ont été développées dans ce chapitre. Ces méthodes permettent la représentation et la transmission à bas débit des enveloppes spectrales :

La première technique, basée sur la prédiction linéaire, consiste à modéliser l'enveloppe spectrale des signaux par un filtre auto-régressif. Dans le contexte de l'enrichissement de bande des signaux musicaux, la modélisation de l'enveloppe haute-fréquence implique :

- Une translation des hautes fréquences à modéliser en bande de base afin d'optimiser la prédiction linéaire
- Un ajustement de l'ordre de la prédiction linéaire en fonction de la nature des signaux, un ordre élevé (de 20 environ pour une modélisation d'une bande d'environ 10 kHz) modélisant correctement les signaux bruités, et un ordre plus faible pour les signaux à fortes composantes tonales.
- Un ajustement temporel sur les signaux transitoires

Concernant l'ajustement d'énergie du spectre par l'enveloppe transmise, la technique requiert un blanchiment préalable du signal cible à remettre en forme.

En terme de débit, le coût de transmission des descripteurs d'enveloppes est faible grâce à la conversion des coefficients de filtre en coefficients LSP. Ces derniers offrent de meilleures propriétés de codage et une quantification vectorielle de ces coefficients conduit à un taux de compression élevé.

La seconde technique développée repose sur une modélisation d'enveloppe et une remise en forme spectrale des signaux par facteurs d'échelle directement dans le domaine fréquentiel.

La modélisation d'enveloppe par facteurs d'échelle en sous-bandes comporte un intérêt certain pour ce type d'application. Elle offre en effet de nombreux degrés de liberté :

- Au niveau de la flexibilité et de la mise en œuvre. En jouant sur le nombre et la largeur des sous-bandes, on peut ainsi modéliser une partie quelconque du spectre avec la précision désirée.
- Elle permet de prendre en compte directement les propriétés psychoacoustiques du signal haute-fréquence.

Contrairement à l'ajustement d'enveloppe par filtre de synthèse AR, la remise en forme spectrale des signaux ne demande pas, a priori, de blanchiment spectral.

Sur les signaux transitoires, on limitera la taille des fenêtres d'analyse/synthèse des transformées temps/fréquence afin de limiter les phénomènes de pré-écho.

En terme de débit, le coût de transmission des facteurs d'échelle peut-être avantageusement réduit grâce à un codage différentiel suivi d'une compression sans perte des données.

Les deux méthodes ont été intégrées dans le codeur/décodeur d'enrichissement de spectre développé dans cette thèse. Nous comparerons au chapitre 5 les performances de ces deux techniques sur le système complet d'extension de bande.

3.6. Bibliographie du chapitre 3

- [ABE 95] M. ABE and Y. YOSHIDA
More natural sounding voice quality over the telephone : An algorithm that expands the bandwidth of telephone speech
NTT Review, Vol. 7, N°3, pp 104/109, Mai 1995
- [CHE 94] Y. M. CHENG, D. O'SHAUGHNESSY, P. MERMELSTEIN
Statistical recovery of wideband speech from narrowband speech
IEEE, Transactions on Speech and Audio Processing, vol. 2, n°4, pp 544-548, Octobre 1994
- [CHO 98] E. L. T. CHOY
Waveform Interpolation Speech Coder
Thèse de l'université Mac Gill, Montréal (Canada), août 1998
- [DER 00] O. DERRIEN et al.
Le codeur MPEG-2 AAC expliqué aux traiteurs de signaux
Annales des télécommunications, Septembre/octobre 2000
- [EPP 00] J. EPPS
Wideband extension of narrowband speech for enhancement and coding
Thèse de l'université de New South Wales, Septembre 2000
- [FER 96] J. S. Ferreira
Convolutional effects in Transform Coding with TDAC : An optimal window
IEEE, Transaction on speech and audio Processing, Vol. 4, N°2, March 1996
- [HAY 96] S. Haykin
Adaptive Filter Theory
Prentice-Hall, 1996
- [HUF 52] D. A. HUFFMAN
A Method for the Construction of Minimum-Redundancy Codes
Proceedings of the IRE, pp. 1098-1101, Septembre 1952.
- [JAY 84] N.S JAYANT & P. NOLL
Digital Coding of Waveforms - Principles and Applications to Speech and Video
Prentice-Hall, Englewood Cliffs, New Jersey, 1984.
- [KLE 95] W.B. KLEIJN & K.K. PALIWAL
Speech coding and synthesis
Elsevier, 1995
- [KOV 97] B. KOVESI
Quantification vectorielle des paires de raies spectrales pour la compression de la parole à débit réduit.
Thèse de l'université de Rennes I, Janvier 1997
- [LIN 80] Y.LINDE, A. BUZZOT et R.M. GRAY
An algorithm for vector quantizer design
IEEE Transactions On communications, Com-28(1):84-95, Janvier 1980
- [MAK 75] J. MAKHOUL
Linear prediction: A tutorial review
Proceedings of the IEEE, vol. 63, n°4, pp 561-580, Avril 1975
- [MAR 76] J.D. MARKEL et A.H. GRAY
Linear prediction of speech
Springer-Verlag, Berlin Heidelberg, New York, 1976
- [MIE 00] G. MIET, A. GERRITS, J.C. VALIERE
Low band extension of telephone-band speech
ICASSP 2000, Istanbul, Vol 3, pp. 1851-1854, Juin 2000

-
- [MOR 95] N. MOREAU
Technique de compression des signaux
Masson, collection technique et scientifique des communications, 1995
- [NIS 99] M. NISHIGUCI, A. INOUE, Y. MAEDA, J. MATSUMOTO
Parametric Speech Coding – HVXC at 2.4-4.0 kbps
Proceedings of the IEEE, Workshop on speech coding, 1999
- [SCH 97] D. SCHWARZ
Spectral envelopes in sound analysis and synthesis
Rapport technique de l'université de Stuttgart, Décembre 1997
- [SMI 98] J. SMITH
Linear Predictive Coding
Rapport Technique, MUS 420/EE 265, 1998
- [SPA 94] A.S. SPANIAS
Speech Coding, a tutorial review
Proceedings of the IEEE, Octobre 1994
- [TAD 99] H. TADDEI
Codage hiérarchique faible retard pour les nouveaux réseaux et services
Thèse de l'université de Rennes I, Octobre 1999
- [VAL 00] J. VALIN & R. LEFEBVRE
Bandwidth Extension of Narrowband Speech for Low Bit-rate Wideband Coding,
IEEE Workshop on speech coding, Delavan (USA), pp. 130-132, Septembre 2000
- [YOS 94] Y. YOSHIDA and M. ABE
An algorithm to reconstruct wideband speech from narrowband speech based on codebook
mapping
Proceedings of the ICSLP, Yokohama (Japon), pp1591-1594, 1994

CHAPITRE 4

TECHNIQUES D'EXTENSION DE LA STRUCTURE FINE SPECTRALE

Plan du chapitre

1.1.	INTRODUCTION.....	74
1.2.	TECHNIQUE D'EXTRAPOLATION DE SPECTRE SANS TRANSMISSION D'INFORMATION	75
1.3.	TECHNIQUES PARAMETRIQUES	76
1.4.	TECHNIQUES DE TRANSLATIONS SPECTRALES	79
1.5.	ENRICHISSEMENT DE SPECTRE PAR DISTORSION NON-LINEAIRE	86
1.6.	CONCLUSION DU CHAPITRE.....	98
1.7.	BIBLIOGRAPHIE DU CHAPITRE 4	99

4.1. Introduction

Les techniques développées dans ce chapitre ont pour but de recréer la structure fine du spectre haute-fréquence à partir du signal à bande-limitée, c'est-à-dire de recréer les composantes harmoniques et de bruit haute-fréquence non transmises, et ceci quelle que soit la nature du signal (parole et musique).

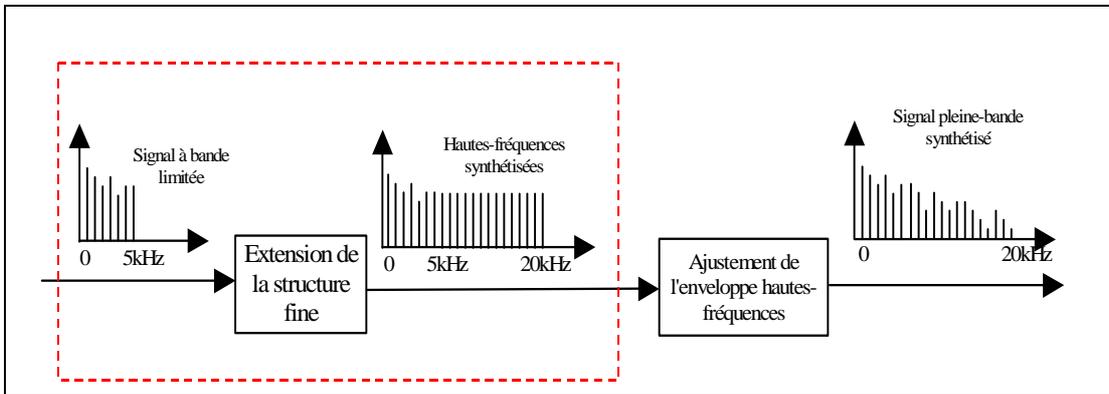


Figure 4.1 : Extension de la structure fine dans la méthode complète d'enrichissement de spectre

La structure fine du spectre de la plupart des signaux musicaux étant sensiblement la même sur tout le spectre (paragraphe 2.3.3.7), on rappelle qu'une technique d'extension efficace devra :

- Générer un bruit haute-fréquence à partir d'un bruit basse-fréquence.
- Étendre une série harmonique (pure ou bruitée) tronquée.
- Étendre la structure fine des signaux plus complexes résultant du mélange de plusieurs séries harmoniques.
- Étendre la structure fine des signaux transitoires sans entacher leur comportement temporel, et en particulier sans générer de pré-échos.

Quatre types de régénération des hautes fréquences sont abordés dans ce chapitre.

Dans le premier paragraphe, nous nous intéressons à une technique qui vise à extrapoler, directement dans le domaine fréquentiel, les échantillons haute-fréquence à partir de ceux contenus en basse-fréquence.

Le second paragraphe est consacré aux techniques d'extension de la structure fine du spectre basées sur une approche paramétrique.

Au paragraphe 4.4, nous présentons les méthodes de translations spectrales qui visent à dupliquer le contenu spectral basse-fréquence en haute-fréquence.

Nous présentons enfin au paragraphe 4.5 les techniques d'extension reposant sur les distorsions non linéaires.

4.2. Technique d'extrapolation de spectre sans transmission d'information

4.2.1. Principe

L'extrapolation linéaire des signaux génériques dans le domaine fréquentiel a fait l'objet d'une étude dans [YOO 01]. L'idée est de prédire les échantillons fréquentiels haute-fréquence à partir des échantillons basse-fréquence dans le domaine MDCT (Modified Discrete Cosine Transform, voir Annexe A). La Figure 4.2 illustre le mode de fonctionnement de la technique.

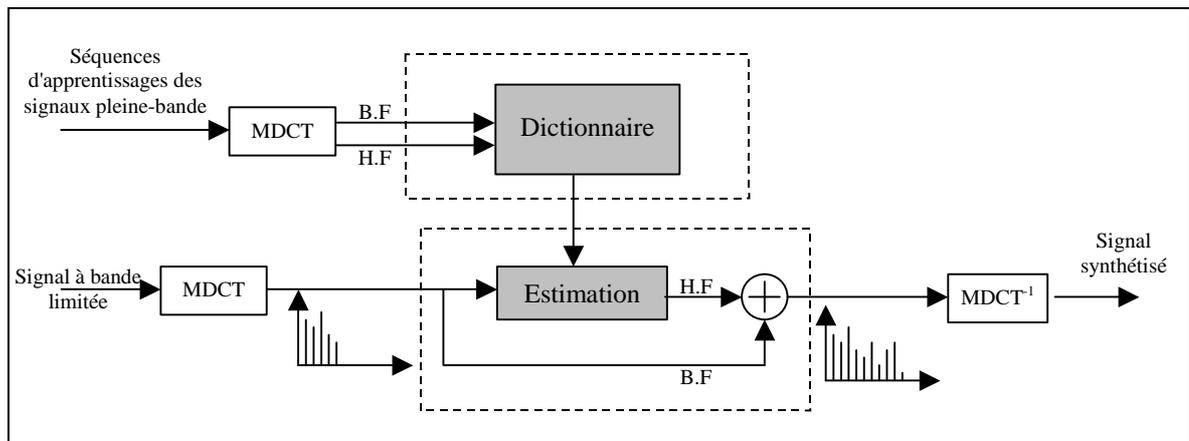


Figure 4.2 : Extrapolation linéaire dans le domaine fréquentiel

On modélise, dans ce modèle, chacune des raies haute-fréquence Y_j comme une combinaison linéaire des raies basse-fréquence X_i :

$$Y_j = \sum_i w_{ji} X_i \quad (4.1)$$

La matrice de pondération w_{ji} est déterminée par minimisation de l'erreur de reconstruction sur une séquence d'apprentissage constituée de plusieurs heures de musique.

4.2.2. Résultats

Sur les signaux purement harmoniques pour lesquels les échantillons haute-fréquence sont corrélés aux basses fréquences, et avec un dictionnaire d'apprentissage adapté (apprentissage sur des signaux harmoniques), la technique synthétise des composantes haute-fréquence perceptivement proches de celles présentes sur le signal original (test sur un extrait de violoncelle dans [YOO 01]), avec toutefois de nombreuses discontinuités temporelles. Les informations d'amplitude et de phase étant en effet mélangées dans la MDCT (transformée à coefficients réels, voir Annexe A), il est délicat de traiter séparément ces deux informations et donc d'assurer, de trame en trame, une continuité dans l'énergie et la phase des signaux.

Concernant les signaux complexes (parole et musique en général), la technique génère du bruit en haute-fréquence et dégrade de ce fait le rendu sonore.

La complexité et la diversité des signaux numériques font qu'il est très difficile (voire impossible) de trouver un modèle générique susceptible d'extrapoler, sans transmission d'information, les raies haute-fréquence d'un spectre à partir de ses basses fréquences.

4.3. Techniques paramétriques

4.3.1. Introduction

Une décomposition paramétrique du signal selon le modèle décrit au chapitre 2.3.3.1 (décomposition en harmoniques + tonales isolées + bruit + transitoires) offre des perspectives intéressantes pour les techniques d'extrapolation de spectre des signaux audio numériques à bas-débit.

Nous passons ici en revue les différentes techniques existantes dans la littérature avant de conclure sur l'efficacité de cette approche dans le cadre de notre étude. Sont développées en particulier les techniques paramétriques utilisées en codage de parole, et une technique paramétrique d'enrichissement de spectre des signaux musicaux, la méthode PlusV.

4.3.2. Synthèse des hautes fréquences sur les signaux de parole

Le modèle source-filtre associé aux signaux de parole, illustré Figure 4.3, est particulièrement adapté à la représentation paramétrique de ces sons.

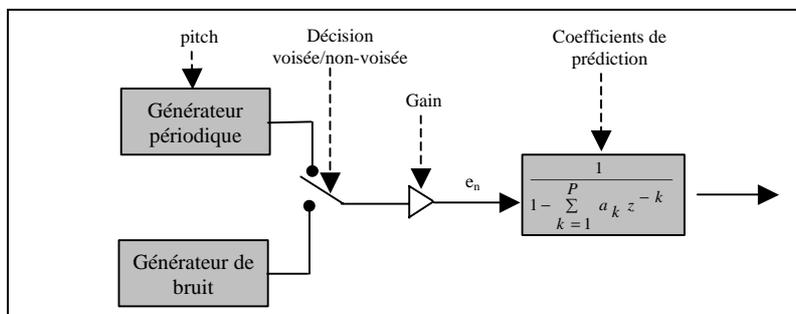


Figure 4.3 : Modèle paramétrique de production de la parole

- Les signaux de parole voisée peuvent être modélisés efficacement par un peigne harmonique mis en forme par une enveloppe spectrale adéquate.
- Les signaux de parole non-voisée peuvent être modélisés efficacement par du bruit mis en forme par une enveloppe spectrale de bruit associée.

Basée sur ce modèle de production de la parole, il devient aisé d'étendre la structure fine spectrale des signaux de parole à bande limitée puisque :

- Pour les sons non-voisés, un simple générateur de bruit synthétisera efficacement la structure fine du spectre haute-fréquence.
- Pour les signaux de parole voisée, la connaissance de la fréquence fondamentale permettra d'étendre aisément la structure fine du spectre à bande limitée ([SPA 94]).

Cette paramétrisation des signaux de parole a donné naissance à plusieurs techniques d'enrichissement de spectre. Citons la méthode d'extension de bande par extraction de pitch, dérivée de la technique CELP et présentée dans [YOS 94], la technique d'extension de bande dérivée des techniques de synthèse MBE (Multi-band excitation) dans laquelle le signal à bande limitée est divisé en sous-bandes centrées sur les harmoniques du signal ([CHA 96]), et la technique d'extension de bande basée sur le modèle STC (Sinusoidal Transform Coding) et présentée au paragraphe suivant. Toutes ces méthodes ont pour but de synthétiser les hautes fréquences non-transmises des signaux de parole en bande téléphonique ("narrowband to wideband conversion")

4.3.2.1. Exemple de mise en oeuvre

L'article [EPP 98] présente une technique d'extrapolation de spectre des signaux de parole dérivée des codeurs de type STC. A partir de signaux de parole voisée de 4 kHz de bande passante, le but du codeur/décodeur d'enrichissement de spectre est de fournir des signaux de 8 kHz de bande.

Le signal haute-fréquence non-transmis est modélisé comme étant la somme de deux composantes :

- Une composante harmonique $\hat{s}_h(n)$, modélisant la partie voisée du signal
- Une composante bruitée $\hat{s}_n(n)$, modélisant la partie non-voisée du signal

Le principe consiste donc à déterminer, pour chaque trame analysée, les composantes haute-fréquence harmoniques et bruitées du signal, dans le but de générer le signal de parole haute-fréquence :

$$\hat{s}_{H.F.}(n) = \hat{s}_h(n) + \hat{s}_n(n)$$

La partie bruitée est synthétisée par un générateur de bruit.

La partie harmonique du signal est synthétisée à partir de la fréquence fondamentale f_0 du signal à bande limitée. Après extraction de cette fréquence⁷, le décodeur synthétise les harmoniques non transmises et génère ainsi une excitation de la largeur de bande souhaitée (extension de la série harmonique par synthèse d'un peigne de sinus séparés par la fréquence fondamentale estimée).

Finalement, le signal synthétisé est remis en forme par une enveloppe adéquate (enveloppe estimée à partir de l'enveloppe du spectre à bande réduite, selon la méthode décrite au paragraphe 3.2.3), puis sommé au signal de base, formant ainsi le signal à bande élargie.

4.3.3. Synthèse des hautes fréquences sur les signaux musicaux

La technique paramétrique d'enrichissement de spectre PlusV ([VLS 01]) consiste à transmettre les composantes élémentaires (bruit + tonales) du signal en haute-fréquence. Un diagramme de fonctionnement est fourni Figure 4.4.

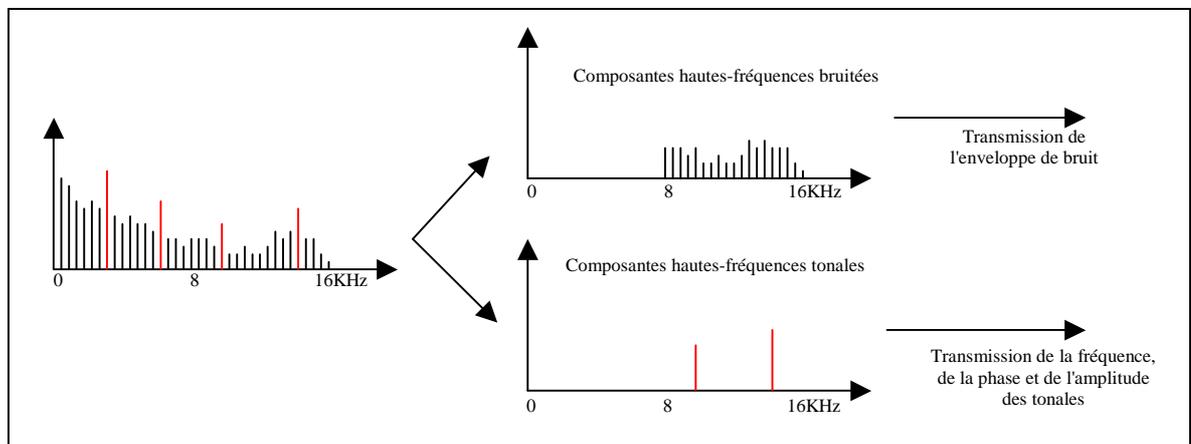


Figure 4.4 : Extension de bande basée sur une approche paramétrique

Le signal d'entrée échantillonné à 32 kHz est passé dans le domaine fréquentiel à l'aide d'une transformée. Le spectre basse-fréquence [0-8kHz] est codé par un codeur cœur classique (MP3). Le codeur d'extension de bande extrait et transmet les composantes tonales du signal haute-fréquence de la bande [8-16kHz]. Le signal résiduel (composantes haute-fréquence bruitées) est modélisé au décodeur par un bruit blanc mis en forme par l'enveloppe transmise (modélisation d'enveloppe en sous-bandes).

⁷ On ne développe pas ici les techniques de détection de fondamentale qui font l'objet d'une vaste étude dans la littérature; un état de l'art et une implémentation efficace sur des signaux mono harmoniques sont proposés dans [BAG 94]

En terme de débit, le coût correspond à la transmission de l'enveloppe de bruit, de la position, de l'amplitude et de la phase des tonales synthétisées (quatre tonales au plus par trame dans le codeur plusV).

Associée à un codeur de type MP3, la technique PlusV offre un codage pleine-bande des signaux stéréophoniques à des débits avoisinant les 64 kbit/s. La technique fonctionne sur des signaux en bande passante relativement élevée (de l'ordre de 8 kHz) et requiert un débit moyen de l'ordre de 8 kbit/s. Qui plus est, elle ne permet de transmettre qu'un nombre limité de tonales et est de ce fait mal adaptée à la modélisation des signaux très harmoniques qui, dans la majorité des cas, présentent de nombreuses tonales au-delà des 8 kHz (instruments à cordes en particulier).

4.3.4. Problèmes liés à l'approche paramétrique sur les signaux musicaux

La détection de pitch et l'extraction des tonales sont des opérations relativement aisées sur des signaux de parole mono-locuteur, en basse-fréquence et sur les signaux fortement harmoniques.

Elle devient toutefois complexe à mettre en oeuvre sur les signaux musicaux. Ces derniers résultant dans la plupart des cas d'un mélange de plusieurs sons (parole, instruments de musiques...), l'extraction des composantes tonales et du bruit devient délicate à réaliser, et plus particulièrement en haute-fréquence. Le niveau de bruit augmente en effet dans la partie haute du spectre et perturbe d'autant plus l'analyse du signal, c'est-à-dire l'extraction et le suivi des composantes tonales.

4.3.5. Conclusion sur l'approche paramétrique

L'approche paramétrique comporte un intérêt certain pour les techniques d'extrapolation de spectre à bas-débit des signaux de parole et des signaux simples composés en particulier d'une seule fréquence fondamentale.

Elle est en revanche actuellement très délicate à mettre en oeuvre sur les signaux musicaux quelconques, et plus particulièrement en haute-fréquence, là où l'analyse paramétrique est fortement perturbée par le bruit.

L'extension de la structure fine du spectre basée sur une telle approche a donc été rejetée dans cette thèse.

4.4. Techniques de translations spectrales

Nous traitons tout d'abord dans ce paragraphe des translations de spectre par repliement spectral et par modulation d'amplitude, réalisables dans le domaine temporel, c'est-à-dire sans transformée temps/fréquence, avant d'aborder les translations de spectre reposant sur de telles transformées.

4.4.1. Translations spectrales réalisées dans le domaine temporel

4.4.1.1. Translation par repliement spectral (spectral folding)

Cette technique, développée pour la première fois dans [MAK 79a] et [MAK 79b] puis reprise dans [AVE 95] et [ENB 99], est utilisée dans l'extrapolation de bande des signaux de parole. Le principe est de régénérer un signal d'excitation en bande élargie (signal spectralement blanchi⁸) à partir d'un signal en bande téléphonique. Le signal synthétisé est ensuite remis en forme par le filtre d'enveloppe extrapolé à partir des basses fréquences (technique d'extrapolation d'enveloppe identique à celle décrite au paragraphe 3.2.3).

L'extension de la structure fine du spectre est réalisée par un simple sur-échantillonnage des signaux à bande-limitée.

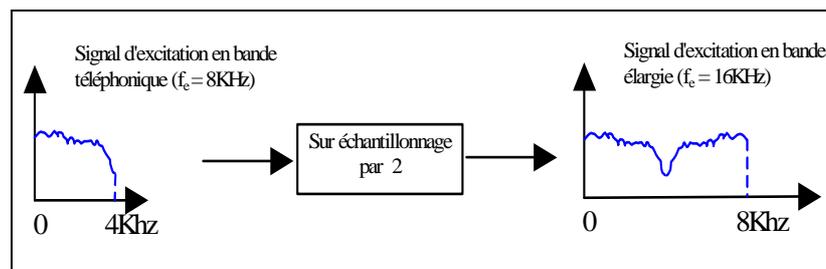


Figure 4.5 : Translation par repliement spectral

Cette opération de sur-échantillonnage a l'intérêt de préserver le signal basse-fréquence et d'assurer une continuité de l'énergie du spectre au niveau du repliement. Notons toutefois que si le signal de base n'est pas pleine bande, il apparaît une discontinuité énergétique dans le spectre au niveau du repliement ("trou" autour de 4 kHz sur la Figure 4.5).

Un autre inconvénient dans le cadre de l'élargissement de bande est que la bande [0-2kHz] qui est en général très harmonique sur les signaux de parole voisée (chapitre 2.3.1), se retrouve traduite en [6-8kHz]. Cette partie du spectre, normalement plus bruitée que la partie basse fréquence, se retrouve alors "sur-harmonisée".

La Figure 4.6 illustre ce phénomène. Elle superpose le spectre blanchi d'un signal résiduel de parole voisée échantillonné à 16 kHz et le spectre du signal synthétisé par repliement spectral.

⁸ Nous avons étudié au paragraphe 3.3.5 l'intérêt de ce blanchiment dans le cadre de l'estimation d'enveloppe par prédiction linéaire

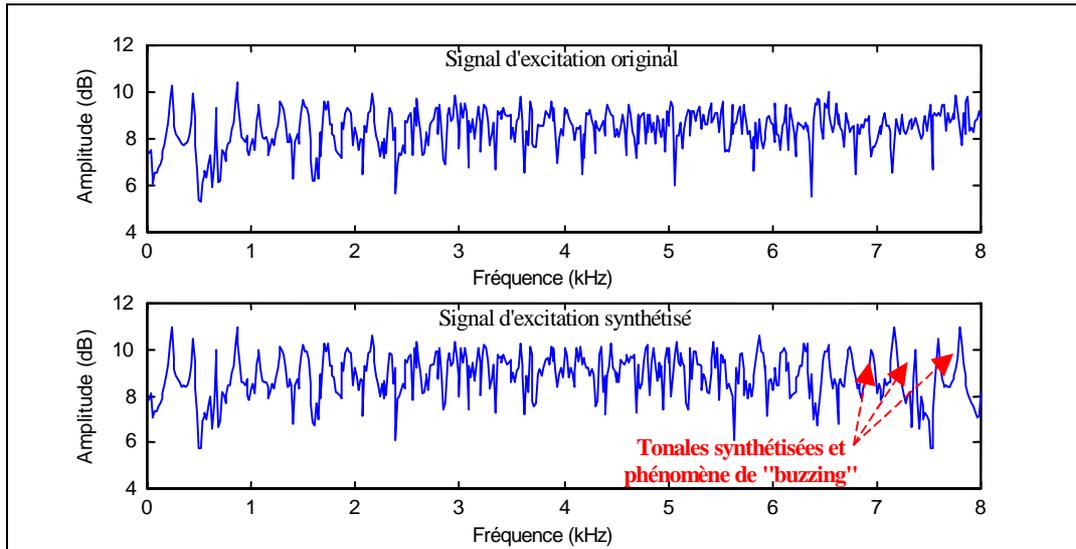


Figure 4.6 : Extension de bande par repliement spectral

Ce phénomène de "sur-harmonisation" ("buzzing" en anglais), très perceptible par l'oreille, donne à la voix un côté nasillard. Il peut être limité en utilisant la technique sur des signaux de bande passante plus élevée (fréquence d'échantillonnage de l'ordre de 16 kHz) afin de translater les basses fréquences en plus haute-fréquence, là où l'oreille est moins sensible.

4.4.1.2. Translation par modulation d'amplitude

On module le signal à bande limitée $x(t)$ par un signal sinusoïdal $\cos 2\pi f_0 t$

$$x_m(t) = x(t) \cos 2\pi f_0 t \quad (4.2)$$

Cette multiplication dans le domaine temporel se traduit par un dédoublement du signal et une translation du signal dans le domaine des fréquences autour de la fréquence $\pm f_0$.

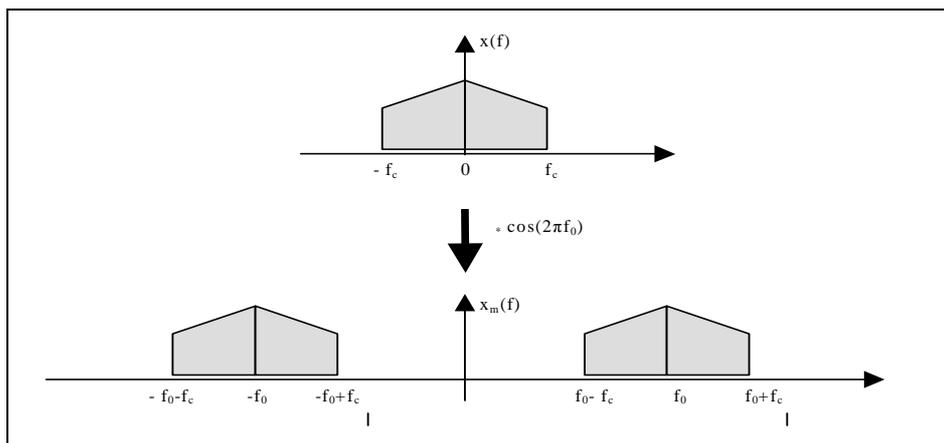


Figure 4.7 : Modulation numérique

Cette technique, facile à mettre en œuvre et peu complexe, est très proche de la technique de repliement de spectre présentée précédemment. Cette dernière s'apparente en effet à une modulation d'amplitude autour de la fréquence $2f_c$. Les modulations d'amplitude, bien qu'offrant plus de degrés de liberté dans le choix des translations, restent néanmoins limitées, en ce sens qu'elles ne permettent de manipuler que des signaux à spectre complet. La translation d'une ou de plusieurs parties du spectre et/ou la suppression du spectre image par modulation numérique requièrent des opérations de filtrage; la fréquence de coupure du signal à bande de base étant variable au cours du temps, un filtrage dans le

domaine temporel devient délicat à mettre en œuvre et l'emploi d'une transformée temps/fréquence s'avère dès lors incontournable.

4.4.2. Translations spectrales mettant en œuvre des transformées temps/fréquences

Après une définition et une description des techniques de mise en œuvre, nous étudions dans ce paragraphe le comportement des translations de spectre sur les signaux musicaux.

4.4.2.1. Définition

La translation spectrale, dont le principe est illustré Figure 4.8 consiste à transférer une partie du spectre d'une position à une autre.

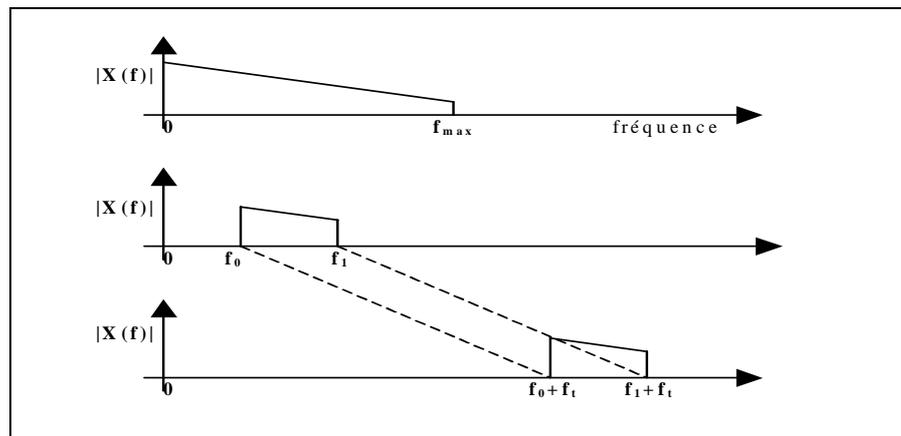


Figure 4.8 : Translation de spectre

4.4.2.2. Mise en œuvre

Une méthode de mise en œuvre des translations de spectre est illustrée Figure 4.9. Les opérations de translations sont ici réalisées à l'aide d'un banc de filtres. L'analyse en banc de filtres consiste à découper un signal en une pluralité de signaux de largeur de bande plus fine. On se référera à [MAL 92], [VAI 93] et [RAU 87] pour plus de détails sur les différentes familles de bancs de filtres.

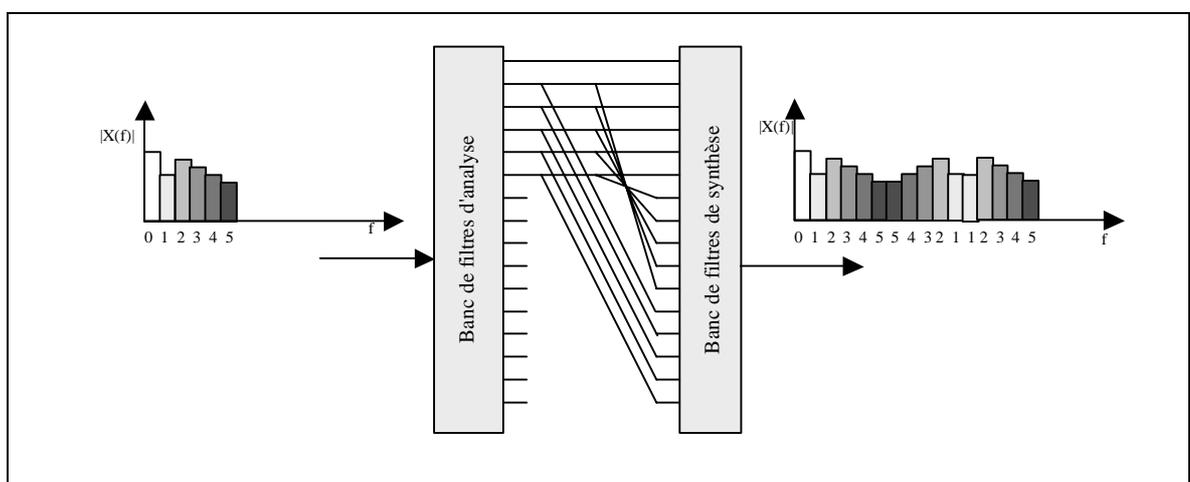


Figure 4.9 : Translations de spectre par banc de filtres

Ce type de réalisation comporte différents avantages puisqu'il offre la possibilité de traduire n'importe quelle partie du spectre basse-fréquence une ou plusieurs fois en haute-fréquence. La technique permet par exemple de s'affranchir de certaines sous-bandes (comme par exemple la sous-bande 0 sur l'exemple Figure 4.9). On notera également que deux types de translations sont également possibles, une translation de spectre directe par simple recopie des sous-bandes basses vers les sous-bandes hautes, et une translation de spectre inversée par recopie inverse des sous-bandes. On réalise sur l'exemple Figure 4.9 une double translation de spectre par recopie inverse et par recopie directe d'une partie des sous-bandes basses.

Plusieurs transformées sont envisageables pour mettre en œuvre les translations de spectre. Nous nous sommes concentrés tout au long de cette thèse sur deux types d'implémentation particuliers : Le premier dans le domaine MDCT, et le second dans le domaine DFT.

4.4.2.3. Comportements sur les signaux musicaux

Sur les signaux complexes (musique en général) résultant d'un mélange de bruit et de plusieurs fréquences fondamentales, la structure fine est sensiblement la même sur tout le spectre. Sur de tels signaux, les translations de spectre dupliquent les sous-bandes basses en haute-fréquence et synthétisent de ce fait un signal pleine-bande de contenu spectral fin proche de l'original.

Contrairement aux techniques développées au paragraphe 4.4.1, les translations spectrales de spectre dans le domaine transformé offrent plus de liberté quant au choix de la partie basse-fréquence à traduire et quant à la longueur de bande du signal à synthétiser. Cet aspect comporte un intérêt certain pour l'enrichissement de spectre des signaux à fréquence de coupure variable (signaux AAC en particulier).

Elles offrent de plus la possibilité de s'affranchir des sous-bandes trop harmoniques en très basse-fréquence (bande [0-2kHz]) qui sont sur les signaux de parole voisée peu liées perceptivement aux plus haute-fréquence.

Elle offre enfin l'avantage de synthétiser en haute-fréquence un signal à spectre complet, c'est-à-dire sans trou d'énergie dans la bande (problème illustré Figure 4.5), et sont de ce fait plus intéressantes que les techniques de translation développées au paragraphe 4.4.1.

Les translations de spectre dans le domaine fréquentiel comportent donc un intérêt certain pour l'extension de la structure fine de la majorité des signaux musicaux. Nous nous intéressons maintenant au comportement des translations de spectre sur les signaux musicaux particuliers, les signaux purement harmoniques et les signaux transitoires.

4.4.2.3.1. Comportement sur un peigne de sinus

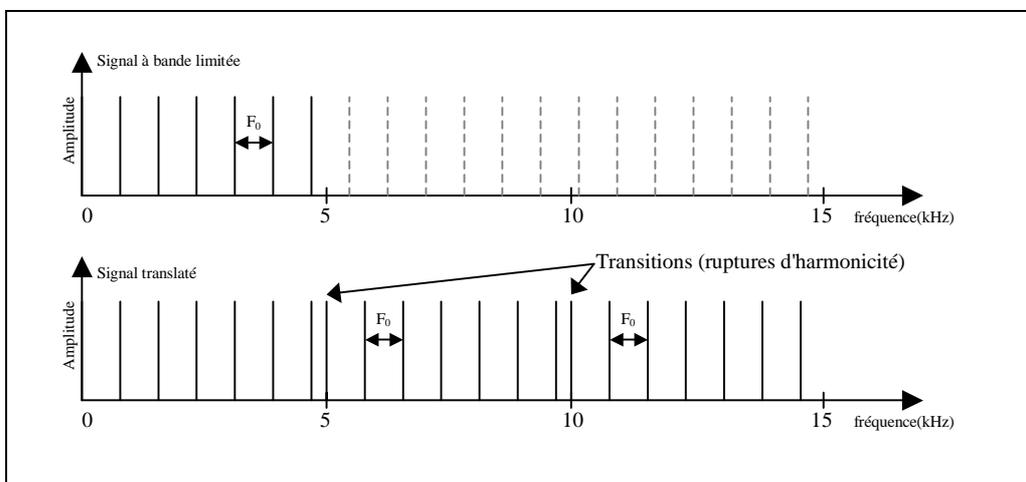


Figure 4.10 : Translations d'un peigne harmonique

Sur les signaux harmoniques, les translations de spectre préservent l'écart entre les fréquences excepté au niveau des transitions entre les blocs traduits.

Le nombre de ruptures d'harmonicité dépend du nombre de translations, et donc de la longueur de la bande du signal à traduire et de la largeur de bande à synthétiser. Sur l'exemple illustré Figure 4.10, la bande [0-5kHz] est traduite deux fois en haute-fréquence, réalisant ainsi une extension de spectre d'une largeur de bande de 10 kHz.

Cet exemple de mise en œuvre génère deux ruptures d'harmonicité. Ces ruptures, non gênantes sur des signaux bruités ou faiblement harmoniques, peuvent en revanche générer des tonales très proches les unes des autres sur les signaux fortement harmoniques. Ces cassures engendrent alors des phénomènes de dissonance très perceptible par l'oreille (paragraphe 2.2.4.1). Ce point sera plus amplement développée au paragraphe 5.2.4.3.

L'utilisation de telles transformées requiert ainsi des translations de sous-bandes de largeur spectrale suffisamment importante afin de limiter le nombre de ruptures d'harmonicité, et donc le risque de dissonance.

4.4.2.3.2. Comportement sur les signaux transitoires

La Figure 4.11 représente respectivement de haut en bas :

- Le spectrogramme d'un signal transitoire échantillonné à 16 kHz
- Le spectrogramme résultant de la somme du signal original filtré à 4 kHz et de ce même signal traduit en [4-8kHz]. La translation est réalisée par une transformée de type MDCT utilisant des fenêtres d'une durée de 32 ms.
- Le spectrogramme résultant de la somme du signal original filtré à 4 kHz et de ce même signal traduit en [4-8kHz]. La translation est réalisée par une transformée de type MDCT utilisant des fenêtres plus courtes de 4 ms.

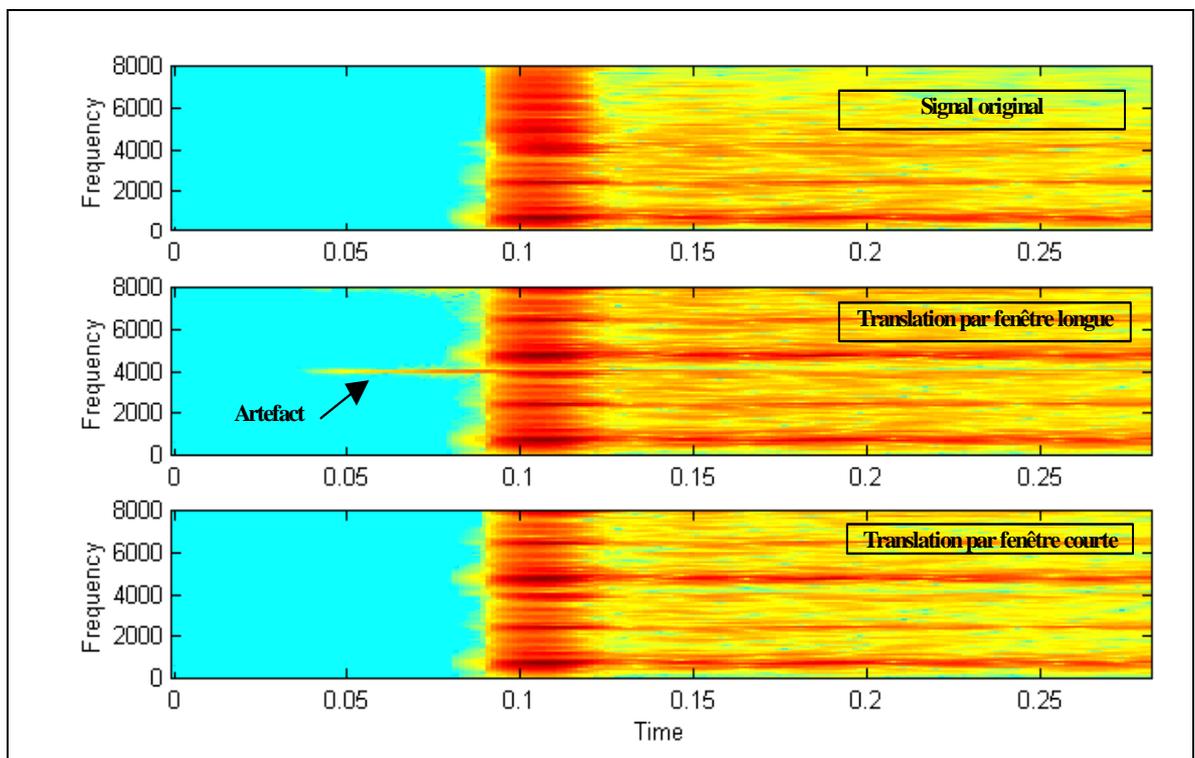


Figure 4.11 : Translation par MDCT d'un signal transitoire

Fréquemment, les translations de spectre préservent la structure fine du signal.

Temporellement en revanche, les transformées temps-fréquences génèrent du bruit au niveau des translations de spectre (autour de 4 kHz sur la Figure 4.11). La durée de ce bruit est proportionnelle à celle des fenêtres de pondérations utilisées par la transformée.

En utilisant des fenêtres de 32 ms, le bruit généré est trop étendu temporellement pour être masqué par l'attaque et il apparaît un phénomène de pré-écho. Pour éviter ce problème, il est nécessaire de réduire la taille des fenêtres afin de limiter les phénomènes d'étalement du bruit et de descendre en dessous du seuil de masquage temporel (de l'ordre de 5 ms, comme vu au paragraphe 2.2.3.2).

Notons que pour un signal transitoire d'énergie décroissante dans le temps (fin d'un signal transitoire par exemple), le passage en mode court n'est pas nécessaire puisque l'oreille est beaucoup moins sensible au post-écho suivant l'attaque (masquage temporel postérieur de plus de 100 millisecondes).

4.4.2.3.3. Remarque sur la translation de spectre d'un signal harmonique

Dans le cas de signaux harmoniques, la translation de spectre en fenêtres courtes (4ms) génère des artefacts (bruit) au niveau de la transition. Ce phénomène est illustré Figure 4.12 qui représente une translation dans le domaine MDCT d'un peigne harmonique filtré à 5 kHz. On translate sur cet exemple la bande [2-5kHz] en [5-8kHz].

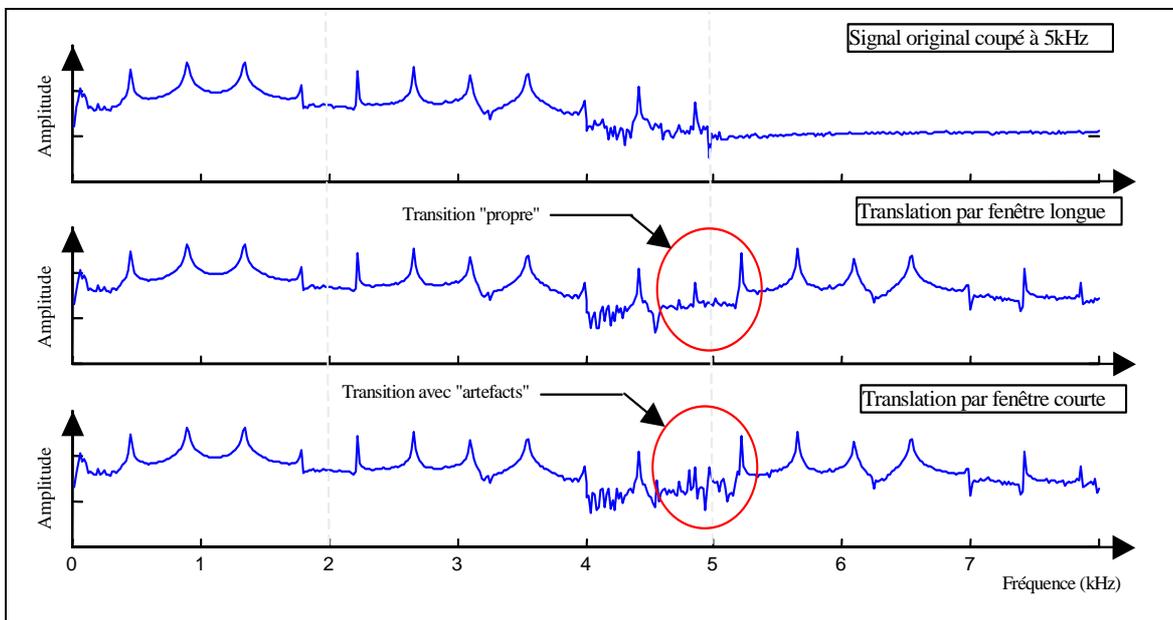


Figure 4.12 : Translation par MDCT d'un signal harmonique

Une translation en fenêtres longues (32 ms) offre une meilleure résolution fréquentielle (concentration de l'énergie des tonales sur quelques raies) et constitue donc un moyen efficace pour traduire les composantes tonales sans générer de bruit parasite, comme illustré Figure 4.12.

4.4.2.4. Conclusion sur les techniques de translations spectrales

Les techniques de translations spectrales par repliement de spectre et par modulation numérique ne requièrent pas de transformée temps fréquences et sont de ce fait simples à mettre en œuvre et peu coûteuses. Elles offrent toutefois des degrés de liberté limités dans la manipulation des différentes parties du spectre.

Les translations de spectre dans le domaine fréquentiel offrent des perspectives plus intéressantes pour l'extension des signaux musicaux.

Elles préservent la structure fine des signaux bruités et faiblement harmoniques (parole et signaux musicaux complexes en général).

Sur les signaux de parole voisée, elles offrent la possibilité de s'affranchir des très basses fréquences dont le contenu fréquentiel, fortement harmonique, est en général perceptivement éloigné du contenu haute-fréquence.

Sur les signaux transitoires, les translations dans le domaine fréquentiel requièrent des fenêtres d'analyse d'une durée de moins de 5 ms, durée du pré-masquage, afin de rendre inaudible les artefacts générés par les transformées temps/fréquence.

Concernant les signaux purement harmoniques composés de une ou de plusieurs fréquences fondamentales, l'harmonicité du signal est préservée, excepté au niveau des transitions entre les différents blocs spectraux. On prendra soin de traduire des blocs de largeur spectrale suffisante afin de limiter les ruptures d'inharmonicité et les phénomènes de dissonance qu'elles engendrent. Nous étudierons au chapitre 5.2.4.3 les effets de telles cassures dans le spectre et tenterons de mettre en place des solutions pour gérer ces discontinuités.

Un compromis sera dès lors à trouver entre la fréquence de coupure du signal à bande limitée, la longueur de bande à traduire, et le fait que sur les signaux de parole voisée, on souhaite s'affranchir des sous-bandes basses ([0-2kHz]) peu liée perceptivement aux plus hautes fréquences au-delà de 4 kHz. Tous ces éléments seront repris au chapitre 5 dans l'élaboration de la technique complète.

4.5. Enrichissement de spectre par distorsion non-linéaire

4.5.1. Introduction

Le filtrage non-linéaire est un outil très répandu en synthèse sonore. Il offre en effet la possibilité de générer, de manière simple et peu coûteuse, des harmoniques en haute-fréquence, créant ainsi des effets diversifiés.

L'approche non-linéaire a suscité quelques études dans le cadre de l'extension de bande des signaux de parole. La littérature est en revanche très pauvre concernant l'utilisation de telles distorsions dans le cadre de l'enrichissement de spectre des signaux audio génériques.

Nous étudions tout d'abord au paragraphe suivant l'effet des non-linéarités sur les signaux transitoires. Ce premier exemple illustre les avantages de cet outil dans le cadre de notre étude.

Après un passage en revue au paragraphe 4.5.3 des différentes techniques non linéaires utilisées en codage de parole, nous analyserons le comportement de ces distorsions sur ces signaux particuliers.

Nous étudions enfin au paragraphe 4.5.5 le comportement de différentes fonctions non-linéaires sur les signaux plus complexes.

Tous les éléments développés nous amèneront à conclure sur cette approche dans le cadre de l'extension de la structure fine spectrale des signaux génériques.

4.5.2. Etude des distorsions non-linéaires sur les signaux transitoires

Nous étudions ici l'effet de deux fonctions non-linéaires, la fonction x^2 et la fonction $|x|$, sur les signaux transitoires. La Figure 4.1 reprend le signal S_{Trans} , défini au paragraphe 2.3.3.6, échantillonné à 32 kHz et formé respectivement d'une attaque de castagnettes, d'une attaque de clavecin et enfin d'une transitoire de cloche. Le second signal, représenté Figure 4.1, correspond au signal S_{Trans} filtré passe-bas à 5 kHz.

Le troisième et le quatrième signal représentent respectivement :

- le signal S_{Trans} filtré à 5 kHz et élevé au carré
- Le signal S_{Trans} filtré à 5 kHz sur lequel on applique une valeur absolue

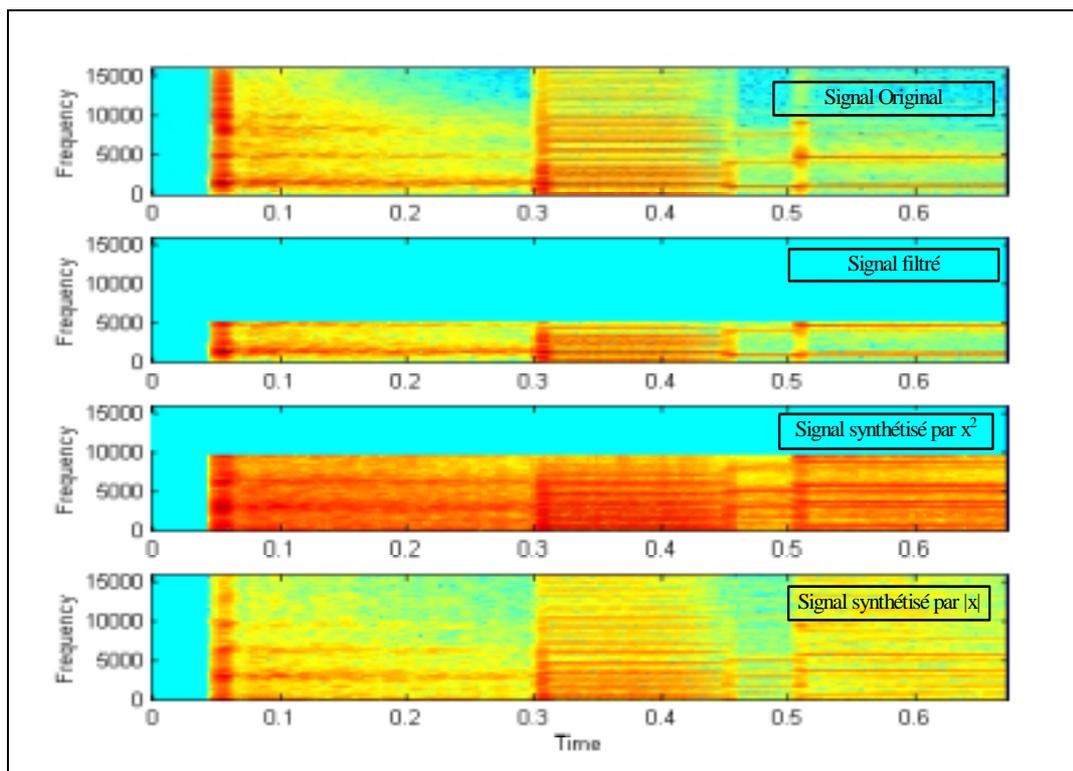


Figure 4.13 : Effets des non-linéarités sur des signaux transitoires

Ce premier exemple nous amène aux deux constatations suivantes :

- Fréquemment, les non-linéarités synthétisent des hautes fréquences a priori de même nature que celles présentes en basse-fréquence. Notons que, à partir du signal de 5 kHz de bande, la fonction x^2 double la bande passante du signal alors que la fonction $|x|$ synthétise des hautes fréquences jusqu'à la fréquence de Nyquist.
- Temporellement, les distorsions non-linéaires génèrent des hautes fréquences tout en préservant la résolution temporelle des signaux : les pré-échos sont absents. Cet outil comporte donc un intérêt certain pour notre étude puisqu'il permet de s'affranchir de toute transformée temps-fréquences qui, comme nous l'avons vu au paragraphe 4.3, génère le plus souvent des problèmes de bords (étalement du bruit du au fenêtrage)

4.5.3. Etat de l'art

Les non-linéarités ont fait l'objet de quelques études dans le domaine de l'extension de bande des signaux de parole. Nous présentons ici les différentes méthodes de mises en œuvre et les différents types de non-linéarité utilisés dans la littérature.

4.5.3.1. Codeur RELP

Le RELP (Residual Excited Linear Predictive Coding) a été développé dans les années 1980 en codage de parole pour des débits variant de 4.8 à 16 kbit/s. Un mode de fonctionnement de ce codeur est illustré Figure 4.14 ([ATA 75]).

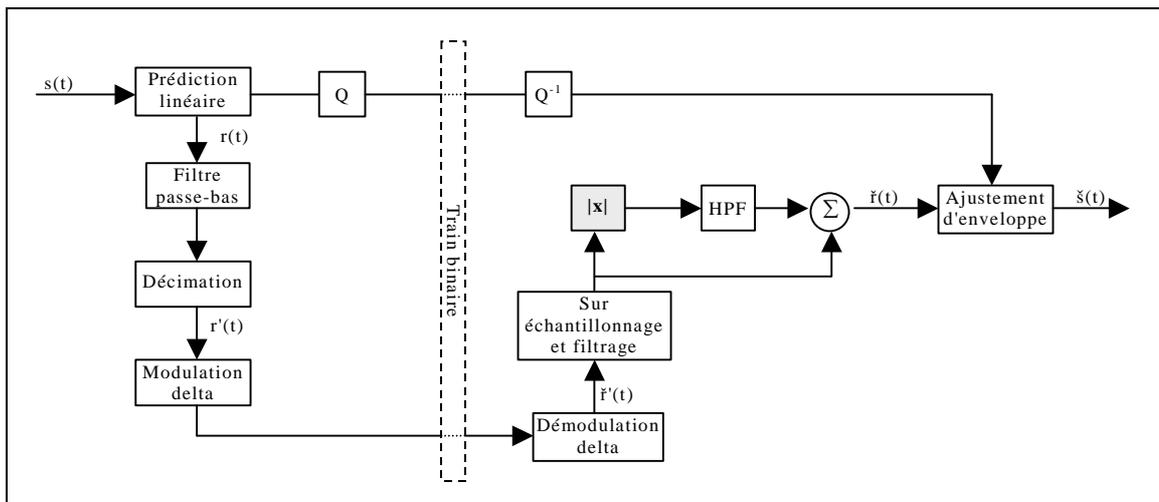


Figure 4.14 : Codage RELP

Au codeur, le RELP effectue une prédiction linéaire sur le signal original $s(t)$. Les coefficients de prédiction sont transmis au décodeur, et utilisés pour blanchir le signal $s(t)$. Le signal résiduel $r(t)$ est passé dans un filtre passe-bas de fréquence de coupure de l'ordre de 1 kHz, puis sous-échantillonné. Le signal résultant $r'(t)$ est codé par un algorithme de type ADM (Adaptative Delta Modulation [JAY 84]) avant d'être transmis.

Au décodeur, le démodulateur Delta permet de retrouver $r'(t)$ qui est ensuite interpolé par non-linéarité (Sur-échantillonnage de $r'(t)$ à la fréquence de $s(t)$ et synthèse des hautes fréquences non-transmises par distorsion non-linéaire de la forme $|x|$).

Le signal ainsi synthétisé est filtré puis sommé au signal de base $r(t)$. Le signal résultant est alors remis en forme par le filtre d'enveloppe déterminé par les coefficients transmis.

L'avantage du RELP est qu'il permet de s'affranchir de la détection et du suivi du pitch, tâches parfois délicates à réaliser en codage de parole. Sur les signaux de parole, la structure fine de la bande [0-1kHz] contient l'information nécessaire (degré de voisement notamment) pour étendre efficacement le spectre en plus haute-fréquence.

4.5.3.2. Autres types de non-linéarité

La technique HFR (High Frequency Regeneration) décrite dans [MAK 79b] vise à étendre la structure fine du spectre des signaux de parole en bande téléphonique par l'utilisation de la fonction non-linéaire:

$$y(t) = \frac{(1 + \alpha)|x(t)| + (1 - \alpha)x(t)}{2}, \quad 0 \leq \alpha \leq 1 \quad (4.3)$$

Les articles [ATA 75], [VIS 82], [VALL 00] proposent différents schémas de codage basés sur la non-linéarité $|x|$. Le double avantage de la valeur absolue est que, contrairement à la fonction de type x^2 par exemple, elle ne requiert pas de normalisation d'énergie et qu'elle accroît le niveau de bruit en haute-

fréquence (point développé au paragraphe 4.5.4.4), limitant ainsi le phénomène de "buzzing" défini au paragraphe 4.4.1.1;

Dans [PAT 81], deux autres types de non-linéarités sont étudiés, la fonction x^3 et la fonction W (Figure 4.15). Seules quelques sonorités particulières (/s/) sont améliorées par ces deux fonctions qui conduisent toutes deux à des résultats similaires.

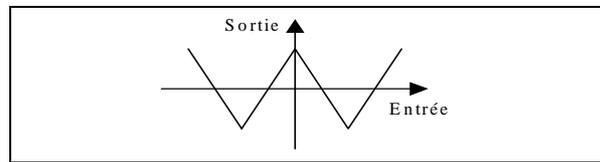


Figure 4.15 : Fonction non linéaire de type W

Notons enfin l'article [WEI 75] qui utilise la fonction non-linéaire :

$$y(t) = \begin{cases} x & , x \geq 0 \\ \frac{|x|}{2} & , x < 0 \end{cases} \quad (4.4)$$

dans le but de synthétiser les harmoniques des sons voisés en bande téléphonique.

4.5.4. Etude des non-linéarités sur les signaux de parole

Afin de bien comprendre l'intérêt porté sur les non-linéarités dans le cadre de l'enrichissement de spectre des signaux de parole, nous introduisons dans ce paragraphe quelques développements mathématiques liés aux fonctions non-linéaires. Nous étudions ensuite le comportement de deux fonctions non-linéaires particulières sur les signaux de parole, c'est-à-dire sur les signaux bruités et/ou harmoniques composés d'une seule fréquence fondamentale.

4.5.4.1. Développement en série de Taylor et phénomènes de repliement spectral

Toute fonction définie et dérivable N fois en 0 admet un développement en série de Taylor de la forme :

$$f(x) = \sum_{n=0}^N a_n x^n \quad (4.5)$$

Prenons l'exemple de la fonction non-linéaire $\ln(1+x)$ qui se décompose en:

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots = \sum_{n=1}^{\infty} (-1)^{n-1} \frac{x^n}{n} \quad , \quad 1 < x \leq 1 \quad (4.6)$$

En remplaçant x par $A \cos(2\pi f_0 t)$, avec $|A| < 1$ c'est-à-dire en injectant un signal de fréquence f_0 , il vient :

$$\ln(1 + A \cos 2\pi f_0 t) = A \cos 2\pi f_0 t - \frac{A^2}{2} \cos^2 2\pi f_0 t + \frac{A^3}{3} \cos^3 2\pi f_0 t - \frac{A^4}{4} \cos^4 2\pi f_0 t + \dots \quad (4.7)$$

et avec les considérations trigonométriques suivantes :

$$\cos^2 x = \frac{1}{2}(1 + \cos 2x) \quad , \quad \cos^3 x = \frac{3}{4} \cos x + \frac{1}{4} \cos 3x \quad , \quad \cos^4 x = \frac{3}{8} + \frac{1}{2} \cos 2x + \frac{1}{8} \cos 4x \quad (4.8)$$

il vient :

$$\ln(1 + A \cos 2\pi f_0 t) = -\left(\frac{A^2}{4} + \frac{2A^4}{32}\right) + \left(A + \frac{3A^3}{12}\right) \cos 2\pi f_0 t - \left(\frac{A^2}{4} + \frac{A^4}{8}\right) \cos 4\pi f_0 t + \frac{A^3}{12} \cos 6\pi f_0 t - \frac{A^4}{32} \cos 8\pi f_0 t + \dots \quad (4.9)$$

Ce qui fait apparaître des partiels aux multiples de la fréquence f_0 .

La Figure 4.16 superpose la réponse en fréquences d'un signal $\cos(2\pi f_0 t)$ composé d'un cosinus pur à la fréquence $f_0 = 3$ kHz, du développement en série de Taylor à l'ordre 4 de la fonction $\ln(1 + \cos(2\pi f_0 t))$ et de $\ln(1 + x)$.

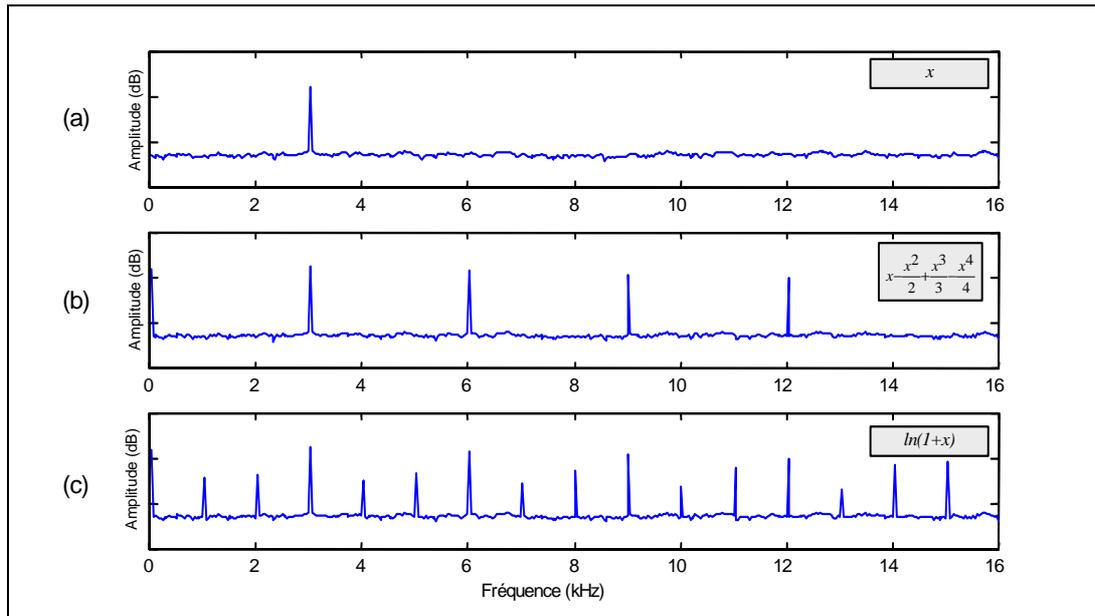


Figure 4.16 : Comportement de la fonction $\ln(1+x)$ sur un sinus

La fonction $\ln(1+x)$ génère (Spectre (c) sur la Figure 4.16) de nombreuses tonales, et par le phénomène de repliement spectral⁹, des tonales non multiples de la fréquence fondamentale. Afin de limiter ces phénomènes, on prendra soin dans la suite du chapitre d'utiliser des signaux suffisamment sur-échantillonnés, ou d'utiliser des fonctions non-linéaires se limitant à un développement en séries de Taylor d'ordre faible (Spectre (b) sur la Figure 4.16).

Nous étudions aux paragraphes suivants l'effet des non-linéarités de type x^2 et de $|x|$ sur les signaux de parole. Notons que la fonction $|x|$ peut être approchée par le développement en série de Taylor à l'ordre 10 :

$$|x| \approx 4.77x^2 - 18.8x^4 + 40.43x^6 - 40x^8 + 14.6x^{10}, \quad -1 \leq x \leq 1 \quad (4.10)$$

Sont étudiés en particulier, les effets de ces deux fonctions sur:

- Un bruit blanc en bande étroite, représentatif des signaux de parole non-voisée
- Un signal harmonique faiblement bruité, représentatif des signaux de parole fortement voisée
- Un signal harmonique bruité, représentatif des signaux de parole faiblement voisée.

⁹ Le repliement spectral (*aliasing* en anglais) se manifeste lorsqu'un signal devant être échantillonné à la fréquence F_e , présente des composantes de fréquence supérieure à la fréquence de Shannon, soit $F_e/2$. Pour les signaux réels, toutes les fréquences situées au-dessus de $F_e/2$ se replient, c'est-à-dire qu'elles ne sont pas perdues mais réapparaissent dans la zone de fréquence $[0, F_e/2]$.

4.5.4.2. Effet des non-linéarités sur un bruit blanc en bande étroite

L'effet des non-linéarités sur le bruit est plus difficile à décrire mathématiquement que sur les signaux sinusoïdaux. Une trame de bruit à un instant t peut toutefois se modéliser par une somme finie de tonales (décomposition en série de Fourier). C'est ainsi qu'une non linéarité de type x^2 appliquée sur un bruit en bande étroite de $[0-4\text{kHz}]$ de bande passante générera un bruit à spectre étendu de bande $[0-8\text{kHz}]$ (similitude avec les signaux harmoniques).

Mais le problème réside dans la description et la perception du bruit synthétisé. Prenons l'exemple d'un bruit blanc pleine bande de densité de probabilité uniforme. A partir de l'expérience décrite Figure 4.17, on génère trois bruits haute-fréquence en bande étroite $[4-8\text{kHz}]$.

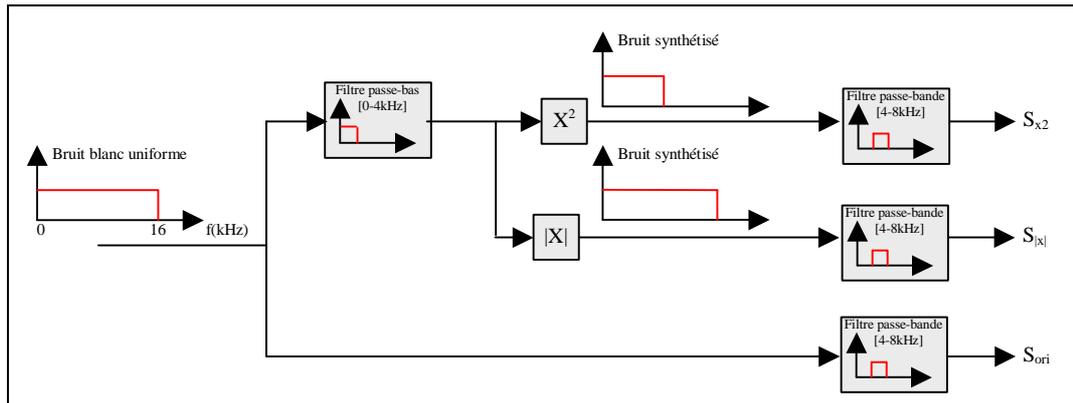


Figure 4.17 : Synthèse de bruit haute-fréquence par non-linéarité

Les trois bruits synthétisés sont de nature et de perception auditive très différentes. Le signal synthétisé S_{x^2} par la non-linéarité x^2 est plus impulsif et est très éloigné perceptivement du signal original S_{ori} . Sa densité de probabilité (Figure 4.18) est très différente de celle du signal cible. Le signal synthétisé $S_{|x|}$ est en revanche plus proche perceptivement du signal S_{ori} . Sa densité de probabilité se rapproche mieux de celle du signal cible. Notons que les densités de probabilité représentées Figure 4.18 ont été normalisées.

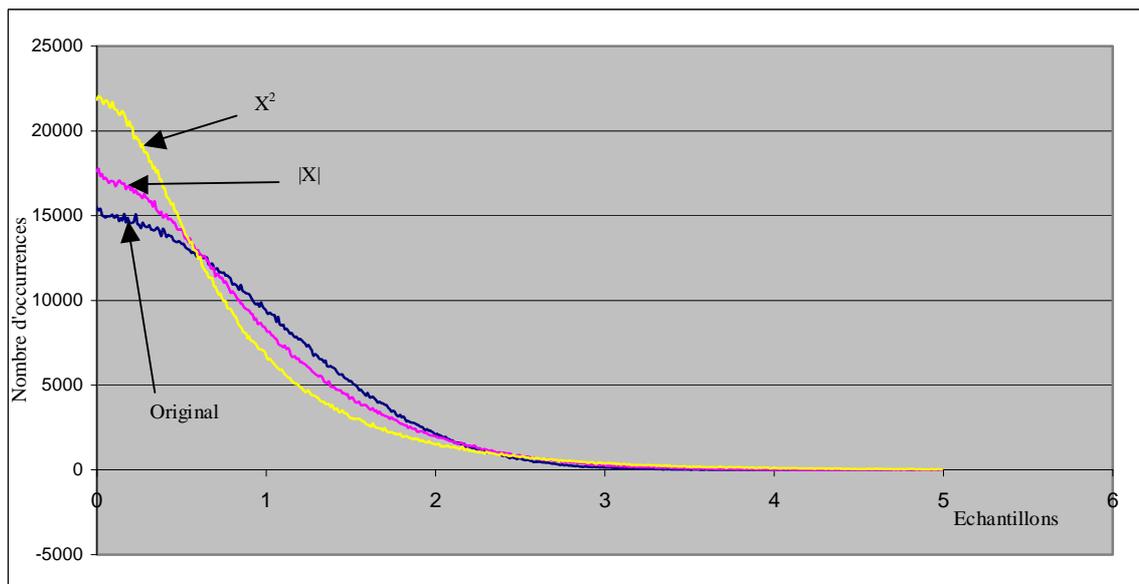


Figure 4.18 : Densités de probabilité des bruits synthétisés par non-linéarité

4.5.4.3. Effets des non-linéarités sur un signal harmonique peu bruité

La Figure 4.19 représente successivement le signal $S_{\text{Mono_harm}}$ pleine bande décrit au chapitre 2, le même signal filtré entre 0 et 4 kHz, le signal filtré élevé au carré et le signal filtré sur lequel on applique une valeur absolue.

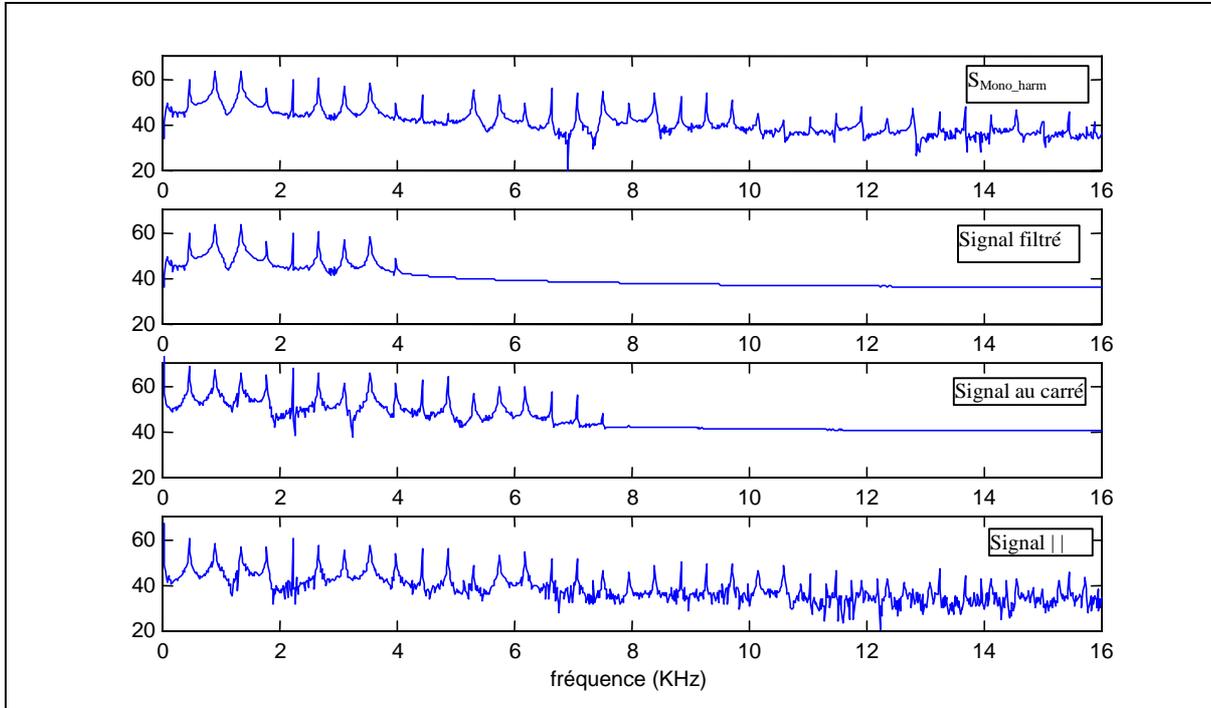


Figure 4.19 : Effets des non-linéarités sur un peigne de sinus

Les deux non-linéarités synthétisent des harmoniques multiples de celles présentes dans le signal à bande limitée. Dans le cas de la valeur absolue, les problèmes de repliement apparaissent principalement en haute-fréquence, le signal résultant devenant alors plus bruité.

Les non-linéarités constituent donc un outil simple pour extrapoler ce type de signaux puisqu'elles ont pour effet d'étendre les séries harmoniques tronquées, sans cassure dans le spectre au niveau des transitions (en 4 kHz sur la Figure 4.19).

Notons enfin que les tonales synthétisées en haute-fréquence sont à phase "continue", en ce sens que le signal synthétisé ne comporte pas de discontinuités temporelles.

4.5.4.4. Effet des non-linéarités sur un signal harmonique bruité

On reprend ici l'expérience réalisée précédemment mais avec un niveau de bruit plus élevé (signal $S_{\text{Mono_Harm}} + S_{\text{Noise}}$ définis au paragraphe 2.3.3).

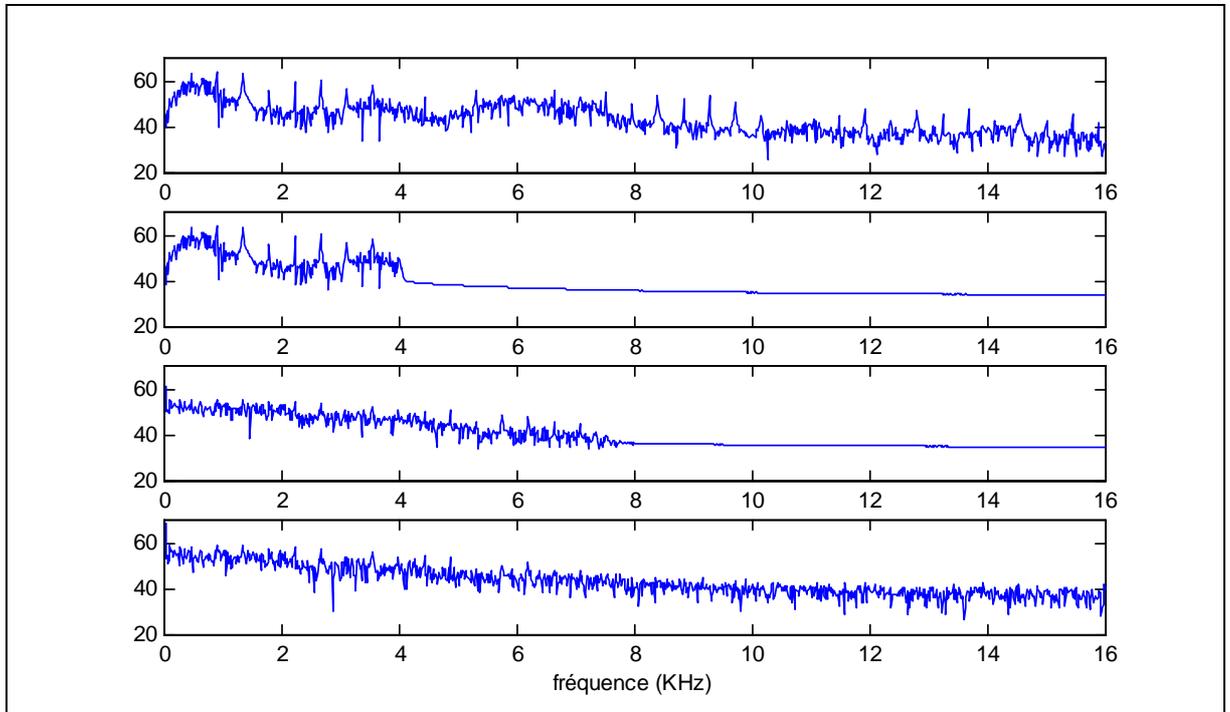


Figure 4.20 : Effets des non-linéarités sur un signal harmonique bruité

Les hautes fréquences synthétisées sont de même nature que les composantes spectrales présentes en basse-fréquence sur le signal d'entrée, avec toutefois un niveau de bruit croissant vers les hautes fréquences. Le rapport harmonique/bruit en haute-fréquence sur les signaux synthétisés est dès lors très proche de celui du signal original. Notons que, au delà des 8 kHz dans le cas de la non-linéarité $|x|$, les tonales sont pratiquement "noyées" dans le bruit.

4.5.4.5. Conclusion

Cette première étude sur les non-linéarités nous a permis de comprendre tout l'intérêt de la fonction $|x|$ dans le cadre de l'enrichissement de spectre des signaux de parole. Pour les signaux non-voisés, la non-linéarité $|x|$ synthétise du bruit en haute-fréquence perceptivement proche de celui présent sur le signal original, et pour les signaux voisés (signaux assimilables à celui décrit Figure 4.20), elle extrapole la structure harmonique avec un niveau de bruit croissant vers les hautes fréquences.

4.5.5. Etude sur les signaux plus complexes

Après avoir étudié les signaux de parole et les signaux transitoires pour lesquels les non-linéarités donnent de bons résultats, nous allons maintenant nous intéresser au comportement de ces fonctions sur les signaux plus complexes.

Nous introduisons tout d'abord les phénomènes d'intermodulation générés par les non-linéarités.

4.5.5.1. Phénomènes d'intermodulation

4.5.5.1.1. Définition

L'intermodulation se produit lorsque deux ou plusieurs composantes fréquentielles interagissent entre elles. Le signal résultant contient alors non seulement les harmoniques des composantes fréquentielles de l'original, mais également des composantes à la somme et la différence (qui ne sont en général pas des harmoniques des fréquences de l'original).

4.5.5.1.2. Exemple

Considérons le signal $s(t)$, dont le spectre est représenté Figure 4.21, constitué de deux composantes sinusoïdales aux fréquences respectives de $f_a=500$ et $f_b=700$ Hz (spectre a) et la fonction non-linéaire d'ordre 3, $f(x)=x^2 + x^3$ (spectre b).

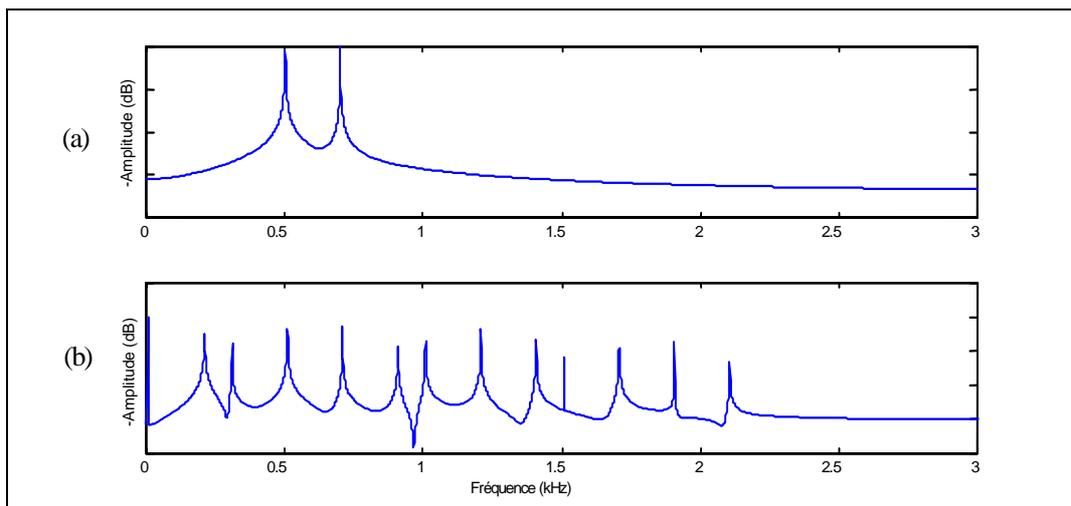


Figure 4.21 : Phénomènes d'intermodulation

Le deuxième ordre génère des composantes en $2f_a$ et $2f_b$, ainsi que les composantes quadratiques différences en $f_a \pm f_b$. Le troisième ordre génère des composantes en $3f_a$ et $3f_b$, ainsi que les composantes différences en $2f_a \pm f_b$ et $f_a \pm 2f_b$.

A partir des deux fréquences, f_a et f_b , on génère ainsi 4 composantes considérées comme "utiles" puisque multiples de ces deux fréquences ($2f_a$, $2f_b$, $3f_a$ et $3f_b$), et 6 composantes d'intermodulation considérées comme "gênantes" puisque inharmoniques par rapport au signal de base.

4.5.5.1.3. Limitation des phénomènes d'intermodulation

Un moyen simple de limiter les phénomènes d'intermodulation consiste à filtrer (filtrage passe-haut) le signal d'entrée du système. Prenons l'exemple d'un signal à bande limitée, de 4 kHz de bande passante, et composé de deux peignes de sinus de fréquence fondamentale $f_0=700\text{Hz}$ et $f_1=1200\text{Hz}$. Ce signal S_1 est représenté Figure 4.22.

Appliquons la fonction x^2 sur ce signal S_1 et filtrons le signal résultant par un filtre passe-haut afin de ne garder que les fréquences supérieures à 4kHz. On obtient le signal S_2 représenté Figure 4.22.

Appliquons maintenant la fonction x^2 non plus sur le signal S_1 , mais sur une version filtré passes-haut à 2 kHz de ce même signal (Signal S_3 correspondant au signal S_1 privé des 3 trois premières tonales comprise entre 0 et 2 kHz). Filtrons ensuite le signal résultant par un filtre passe-haut de façon à ne garder que les hautes-fréquences au-delà de 4kHz. On obtient le signal S_4 .

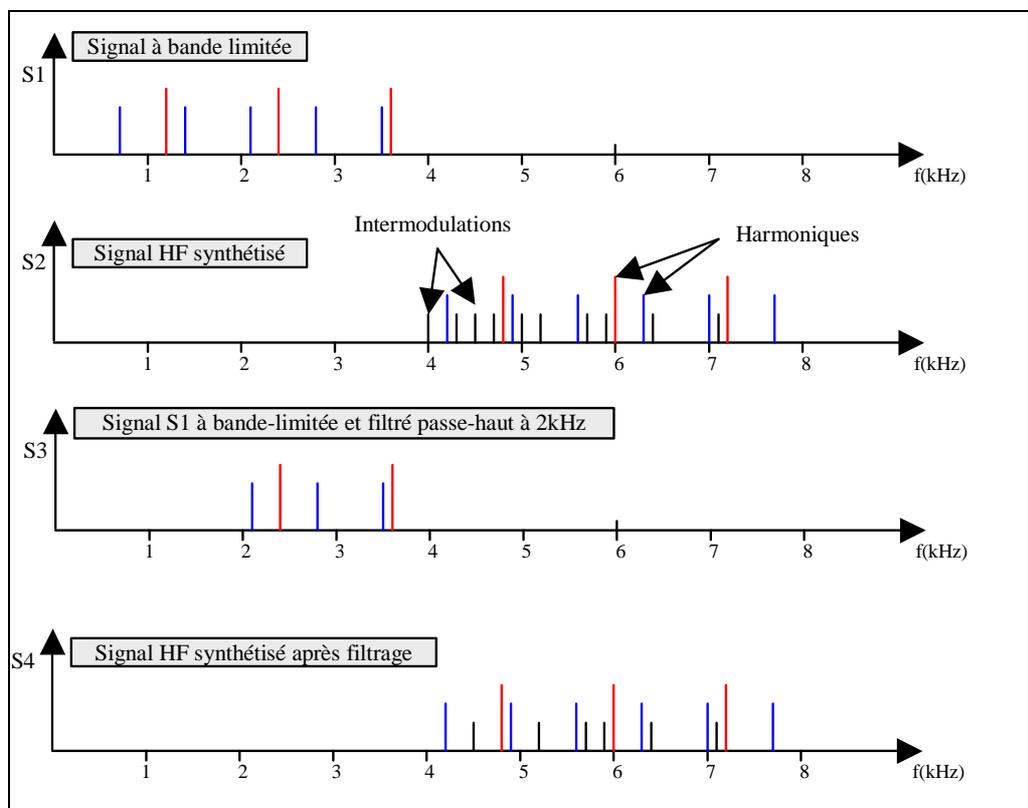


Figure 4.22 : Limitation des phénomènes d'intermodulation

Dans les deux cas (S_2 et S_4), on génère bien les harmoniques multiples du signal à bande limitée S_1 . Le signal S_2 contient en plus tous les termes croisés de fréquence $f_0 \pm f_1$ (10 tonales d'intermodulations). Concernant S_4 , ce nombre passe à 6 puisque le nombre d'interactions diminue. Le pré-filtrage passe-haut des signaux réduit ainsi les phénomènes d'intermodulations sur les signaux multi-harmoniques et inharmoniques.

4.5.5.2. Effet des non-linéarités sur les signaux complexes

Les signaux musicaux résultent en général du mélange de différents instruments et de parole. Ils peuvent ainsi être modélisés dans la majorité des cas par du bruit mélangé à plusieurs peignes de sinus de fréquences fondamentales diverses. Nous étudions ici le comportement des fonctions x^2 et $|x|$ sur le signal $S_{\text{Multi_harm}}$ décrit au paragraphe 2.3.3.4. Ce signal est composé de bruit et de deux peignes de sinus de fréquences fondamentales respectives $f_0 = 440\text{Hz}$ et $f_1 = 750\text{Hz}$.

Afin de limiter les phénomènes d'intermodulations précédemment introduits, on applique les deux non-linéarités sur le signal filtré passe-bande entre 2 et 4 kHz. Les spectres synthétisés sont représentés Figure 4.23.

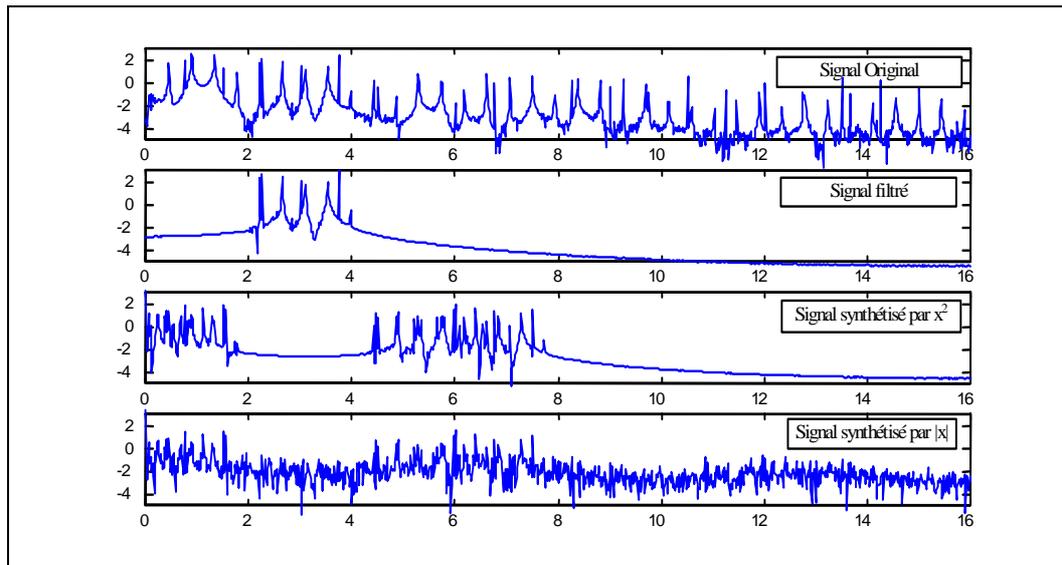


Figure 4.23 : Effets des non-linéarités sur deux peignes de sinus

Dans le cas de la non-linéarité en x^2 , les tonales synthétisées par intermodulation viennent "polluer" les harmoniques haute-fréquence. Le signal reste toutefois perceptivement proche de l'original.

Dans le cas de la non-linéarité $|x|$ en revanche, les phénomènes d'intermodulations (malgré le filtrage passe-bande) et de repliement sont beaucoup plus prononcés et la structure fine s'éloigne fortement de celle du signal original. Le nombre de tonales est si élevé que le signal synthétisé devient assimilable à un bruit.

C'est en fait là le problème essentiel de la technique d'enrichissement de spectre des signaux musicaux par distorsions non-linéaires. Ce type de signaux, composé d'un mélange d'harmoniques, est en effet très fréquent en musique (mélange d'instruments harmoniques, traînée résultant de la transition douce entre deux notes consécutives de piano...).

4.5.6. Conclusions sur les distorsions non linéaires

Les non-linéarités constituent un outil simple et prometteur dans le cadre de l'enrichissement de spectre. Outil réalisable sans transformée temps/fréquence, il a le double avantage d'être très peu complexe en ressources CPU et de ne pas générer d'artefacts gênants sur les signaux transitoires.

La non-linéarité de type $|x|$ est particulièrement appropriée pour enrichir la structure fine des signaux de parole, et plus généralement des signaux bruités et/ou composés d'une seule fréquence fondamentale puisque :

- A partir d'un bruit à bande-limitée, elle synthétise du bruit haute-fréquence perceptivement proche du bruit original.

-
- A partir d'un signal harmonique tronqué, elle étend la série harmonique, sans cassure au niveau des transitions et tout en gardant une cohérence dans la phase des harmoniques synthétisées.
 - A partir d'un signal harmonique bruité, la fonction $|x|$ étend la structure harmonique avec un niveau de bruit croissant en haute-fréquence, respectant ainsi les propriétés spectrales des signaux de parole en haute-fréquence (paragraphe 2.3.1).

Concernant l'extension de la structure fine des signaux musicaux en revanche, les non-linéarités s'avèrent être un outil totalement inefficace. Sur les signaux complexes composés de différentes fréquences fondamentales, les intermodulations sont telles que la structure fine du spectre synthétisée à partir des signaux à bande-limitée s'apparente à du bruit et se retrouve de ce fait perceptivement très éloignée de celle présente sur le signal original.

Ces phénomènes d'intermodulations peuvent être limités par filtrage passe-bande des signaux à étendre mais ce filtrage demeure toutefois insuffisant pour la plupart des signaux musicaux.

Vu la complexité du modèle mathématique associé aux non-linéarités, on maîtrise actuellement mal leurs effets sur les signaux complexes, et notamment dans le domaine fréquentiel.

4.6. Conclusion du chapitre

Il n'existe pas de méthode générique susceptible d'étendre la structure fine de tous les signaux musicaux et les méthodes d'extrapolation des échantillons fréquentiels sont inefficaces vue la diversité des signaux musicaux.

Les techniques paramétriques requièrent une analyse précise du signal. Sur les signaux musicaux, et plus particulièrement en haute-fréquence, le niveau de bruit étant en général assez élevé, l'extraction et le suivi des composantes sinusoïdales deviennent délicats à mettre en œuvre. L'approche paramétrique n'est toutefois pas à exclure car elle constitue un moyen efficace pour transmettre, à bas débit, des tonales isolées (extension de bande des signaux inharmoniques en particulier).

Les techniques de translations de spectre dans le domaine temporel ont l'avantage de ne pas requérir de transformée temps/fréquence mais restent toutefois limitées dans le choix des translations.

Deux techniques d'extension de la structure fine ont fait l'objet d'une étude plus approfondie tout au long de cette thèse

- Les translations spectrales dans le domaine fréquentiel.
- Les distorsions non-linéaires

Ces deux techniques sont en effet plus appropriées pour les applications d'enrichissement de spectre à bas-débit. Sur les signaux de parole et de musique composés d'un seul peigne harmonique, les deux techniques donnent des résultats comparables puisqu'elles étendent toutes deux le spectre en préservant la structure fine. Sur les signaux purement harmoniques, les non-linéarités étendent la structure harmonique des signaux sans cassure dans le spectre et sont de ce fait plus efficaces que les translations spectrales.

Sur les signaux composés de plusieurs signaux harmoniques en revanche, les intermodulations générées par non-linéarité sont telles que le signal synthétisé en haute-fréquence se retrouve très éloigné perceptivement du signal original.

On a donc retenu la technique reposant sur les translations spectrales dans le domaine fréquentiel. Rappelons qu'elle offre une grande flexibilité dans le choix :

- De la mise en œuvre (MDCT, DFT, Banc de filtres)
- De la largeur de bande à synthétiser
- De la largeur des bandes à traduire
- Du type de translation (avec ou sans retournement de spectre)

Les translations de spectre préservent la structure fine de la majorité des signaux audio-numériques, qu'ils soient bruités, harmoniques ou composés de plusieurs séries harmoniques. Elles requièrent toutefois quelques ajustements sur les signaux transitoires et sur les signaux harmoniques :

- Concernant les signaux transitoires, on prendra soin de conserver une résolution temporelle suffisante dans l'utilisation des transformées temps/fréquences afin de pas générer de pré-écho.
- Concernant les signaux purement harmoniques, quelques précautions sont à prendre afin de limiter les phénomènes de dissonance.

Tous ces éléments seront repris au chapitre suivant dans lequel nous présentons la technique d'extension de bande complète développée durant ces trois années de thèse.

4.7. Bibliographie du Chapitre 4

- [ATA 75] B.S. ATAL, M.R. SCHOEDER, V. STOVER
Voiced-excited predictive coding system for low bit-rate transmission of speech
Proceedings from International conference on communications and Process, San Francisco, 1975
- [AVE 95] C. AVENDANO, H. HERMANSKY and E.A. WAN
Beyond Nyquist : Towards the recovery of broad-bandwidth speech from narrow-bandwidth speech
Proceedings of EuroSpeech , 4th, Madrid, pp 165-168, Septembre 1995
- [BAG 94] P. BAGSHAW
Automatic prosodic analysis for computer aided pronunciation teaching
Thèse de l'université d'Edimbourg, 1994
- [CHA 96] C.F CHAN and W.K HUI
Wideband re-synthesis of narrow-band CELP coded speech using multi-band excitation model
Proceedings ICSLP, Philadelphia, Vol. 1, pp 667-670, 1996
- [ENB 99] N. ENBOM, W. B. KLEIJN
Bandwidth Expansion of Speech based on Vector Quantization of the Mel Frequency Cepstral Coefficients
IEEE, Workshop on speech coding, Porvoo, Finland, pp 171-173, Juin 1999
- [EPP 98] J. EPPS & W.H. HOLMES
Speech enhancement using STC-based bandwidth extension
ICSLP, Sydney, Décembre 1998
- [JAY 84] N.S. JAYANT & P. NOLL
Digital coding of waveforms
Prentice Hall, 1984
- [MAK 79a] J. MAKHOUL and M. BEROUTI
High frequency regeneration in speech coding systems
IEEE, ICASSP, pp 428-431, Avril 1979
- [MAK 79b] J. MAKHOUL and M. BEROUTI
Predictive and residual encoding of speech
Acoustical Society of America, Décembre 1979
- [MAL 92] H. MALVAR
Signal processing with lapped transforms
Norwood, Artech House, 1992
- [MIE 00] G. MIET, A. GERRITS, J.C. VALIERE
Low band extension of telephone-band speech
ICASSP 2000, Istanbul, Vol 3, pp. 1851-1854, Juin 2000
- [PAT 81] P.J. PATRICK and C.S. XYDEAS
Speech quality enhancement by high frequency band generation
Digital Processing of Signals in Communication
Proceedings of the IERE, n°49, Loughborough, pp 365-373, Avril 1981
- [RAU 87] J.B. RAULT
Algorithme de réduction de débit pour le codage des voix son haute qualité
Thèse de l'université de Rennes 1, Mai 1987
- [SPA 94] A.S. SPANIAS
Speech Coding, a tutorial review
Proceedings of the IEEE, Octobre 1994
- [VAI 93] P. P. VAIDYANATHAN
Multirate systems and filter banks
Prentice Hall, Englewood Cliffs, 1993

-
- [VAL 00] J. VALIN & R. LEFEBVRE
Bandwidth Extension of Narrowband Speech for Low Bit-rate Wideband Coding,
IEEE Workshop on speech coding, Delavan (USA), pp. 130-132, Septembre 2000
- [VIS 82] V. R. VISWANATHAN, A.L. HIGGINS, W.H. RUSSEL
Design of a robust baseband LPC coder for speech transmission over 9.6 kbit/s noisy channels
IEEE Transaction Communication, Vol. Com. 30, n° 4, Avril 1982
- [VLS 01] VLSI Solution OY
The PlusV specification
Rev. 1.0, October 2001, www.plusV.org
- [WEI 75] C.J. WEINSTEIN
A linear prediction vocoder with voice excitation
EASCON 1975
- [YOO 01] S.W YOON et D. CHOI
Bandwidth Extrapolation for Audio Signal
Final Report, <http://ise0.stanford.edu/class/ee368c>, Février 2001
- [YOS 94] Y. YOSHIDA and M. ABE
An algorithm to reconstruct wideband speech from narrowband speech based on codebook
mapping
Proceedings of the ICSLP, Yokohama (Japon), pp1591-1594, 1994

CHAPITRE 5

TECHNIQUES COMPLETES D'ELARGISSEMENT DE BANDE

Plan du chapitre

1.1.	INTRODUCTION.....	102
1.2.	TECHNIQUE PAT (PERCEPTUAL AUDIO TRANSPOSITION).....	103
1.3.	TECHNIQUE SBR (SPECTRAL BAND REPLICATION)	128
1.4.	RESULTATS DES TESTS MPEG-4.....	131
1.5.	CONCLUSION DU CHAPITRE.....	134
1.6.	BIBLIOGRAPHIE DU CHAPITRE 5.....	135

5.1. Introduction

Après avoir présenté, aux deux chapitres précédents, les techniques d'estimation d'enveloppes spectrales et les techniques d'extension de la structure fine, nous développons dans ce chapitre deux solutions complètes d'élargissement de bande.

La première d'entre elles concerne la technique d'enrichissement de spectre PAT (Perceptual Audio Transposition) développée pendant cette thèse, et proposée en normalisation dans les projets DRM et MPEG-4. Deux implémentations particulières sont détaillées : la première basée sur une modélisation d'enveloppe par prédiction linéaire et la seconde basée sur une modélisation d'enveloppe par facteurs d'échelle dans le domaine fréquentiel.

Nous donnons au paragraphe 3 une brève description de la technique concurrente testée dans MPEG-4, la technique d'enrichissement de spectre SBR (Spectral Band Replication).

Nous présentons pour conclure ce chapitre les tests comparatifs entre ces deux techniques.

5.2. Technique PAT (Perceptual Audio Transposition)

5.2.1. Introduction

La technique d'enrichissement de spectre développée durant ces trois années de thèse repose sur l'architecture fournie Figure 5.1

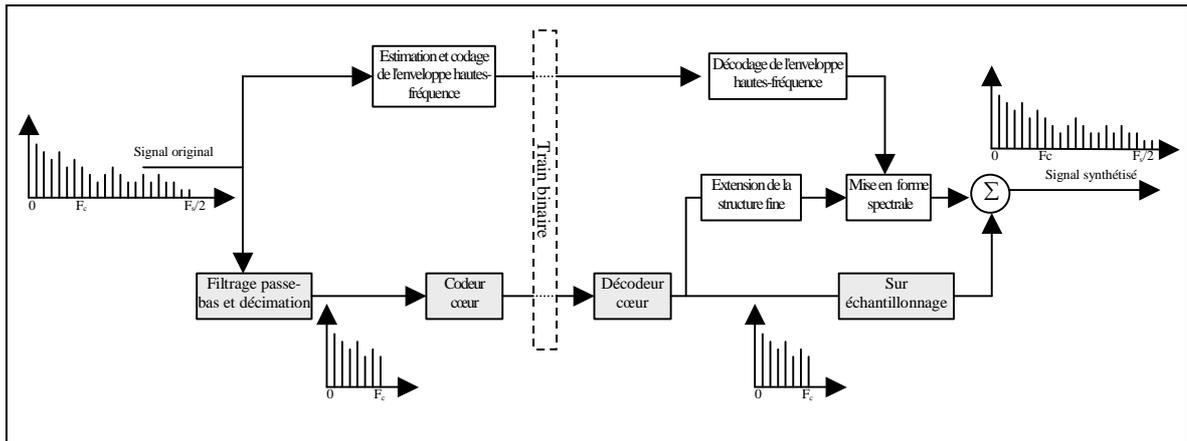


Figure 5.1 : Diagramme de fonctionnement du codeur/décodeur PAT

Le signal original pleine-bande est filtré et sous-échantillonné avant d'être codé par le codeur cœur bas-débit. Afin de simplifier la description de la technique, on travaille dans ce chapitre avec des signaux originaux échantillonnés à 48 kHz, et des signaux codés échantillonnés à 24kHz. Notons que la technique PAT implémentée offre une grande souplesse d'utilisation concernant la fréquence d'échantillonnage des signaux traités. Elle fonctionne notamment avec des signaux d'entrée de fréquences d'échantillonnage diverses (8, 22.05 et 24 kHz) et peut délivrer en sortie des signaux de fréquences variées (16, 32, 44.1 et 48 kHz). La technique associée au codeur de parole G-729, et dont nous présentons les tests d'écoute au paragraphe 5.4.1, fonctionne par exemple avec des signaux codés échantillonnés à 8 kHz.

Les hautes fréquences, non-transmises par le codeur cœur, sont injectées dans un module d'estimation d'enveloppe. Les descripteurs d'enveloppe sont ensuite quantifiés avant d'être transmis, à un débit avoisinant les 2 kbit/s dans un flux distinct de celui du codeur cœur. Ce débit nous est imposé par des contraintes pratiques liées aux applications visées. Il résulte en effet d'un compromis entre le partage du débit requis par le codeur AAC et celui requis par le codeur d'extension de bande pour des applications à 24 kbit/s dans les projets DRM et MPEG-4. Le choix s'est porté sur un débit de 22 kbit/s pour l'AAC (codage des signaux d'une bande passante comprise entre 4 et 7 kHz, cf Tableau 1.1) et de 2 kbit/s pour le module d'extension de bande.

Au décodeur, le décodeur cœur synthétise le signal à bande-limitée. Celui-ci est injecté dans le module d'extension de la structure fine qui synthétise les hautes fréquences. Celles-ci sont ensuite ajustées par les descripteurs d'enveloppe transmis et décodés. Le signal résultant est sommé au signal cœur, générant ainsi le signal de sortie à bande élargie.

Concernant le module d'extension de la structure fine, la technique basée sur les translations de spectre dans le domaine fréquentiel (conformément à celle étudiée au paragraphe 4.4.2) a été retenue. Nous revenons au paragraphe 5.2.4 sur les ajustements requis par cette méthode afin d'étendre efficacement la structure fine des signaux de parole et de musique, et plus particulièrement celle des signaux harmoniques.

Concernant le module d'estimation et de transmission d'enveloppe, deux techniques ont été implémentées et testées sur la chaîne complète. L'une basée sur la prédiction linéaire (paragraphe 3.3) et l'autre utilisant les facteurs d'échelles dans le domaine transformé (paragraphe 3.4). Nous étudions au paragraphe 5.2.5 les avantages et inconvénients de chacune de ces techniques dans la technique complète d'enrichissement de spectre. L'étude du comportement de ces deux méthodes sur les signaux musicaux, ainsi qu'un test d'écoute final, nous permettra de retenir la meilleure des deux techniques.

Les solutions décrites dans ce chapitre requièrent des modules auxiliaires d'analyse des signaux. Nous développons aux paragraphes 5.2.2 et 5.2.3 les techniques de détection d'attaque et d'analyse harmonique des signaux utilisées par la suite. Rappelons que le module de détection d'attaque est requis afin de contrôler les phénomènes d'étalement de bruit générés par les transformée temps-fréquences (paragraphe 3.4.3) et que l'analyse harmonique du signal permet de contrôler le rapport tonales à bruit en hautes fréquences par blanchiment spectral du signal à bande-limitée. (paragraphe 3.3.1.2)

5.2.2. Détection de signaux transitoires

5.2.2.1. Introduction

Un signal transitoire, communément appelé "attaque" en codage audio, correspond à une brusque variation d'énergie de tout, ou d'une partie du spectre du signal. Elle correspond, la plupart du temps, à l'apparition d'un instrument ou à un changement de note au cours du temps. La détection d'attaque est requise dans le codage par transformée afin d'ajuster la taille des fenêtres d'analyse et de synthèse, l'amélioration de la résolution temporelle entraînant une diminution des phénomènes d'étalement de bruit et des phénomènes de pré-écho (problème abordé au paragraphe 3.4.3).

Reprenons le signal S_{Trans} défini au chapitre 2.3.3.6

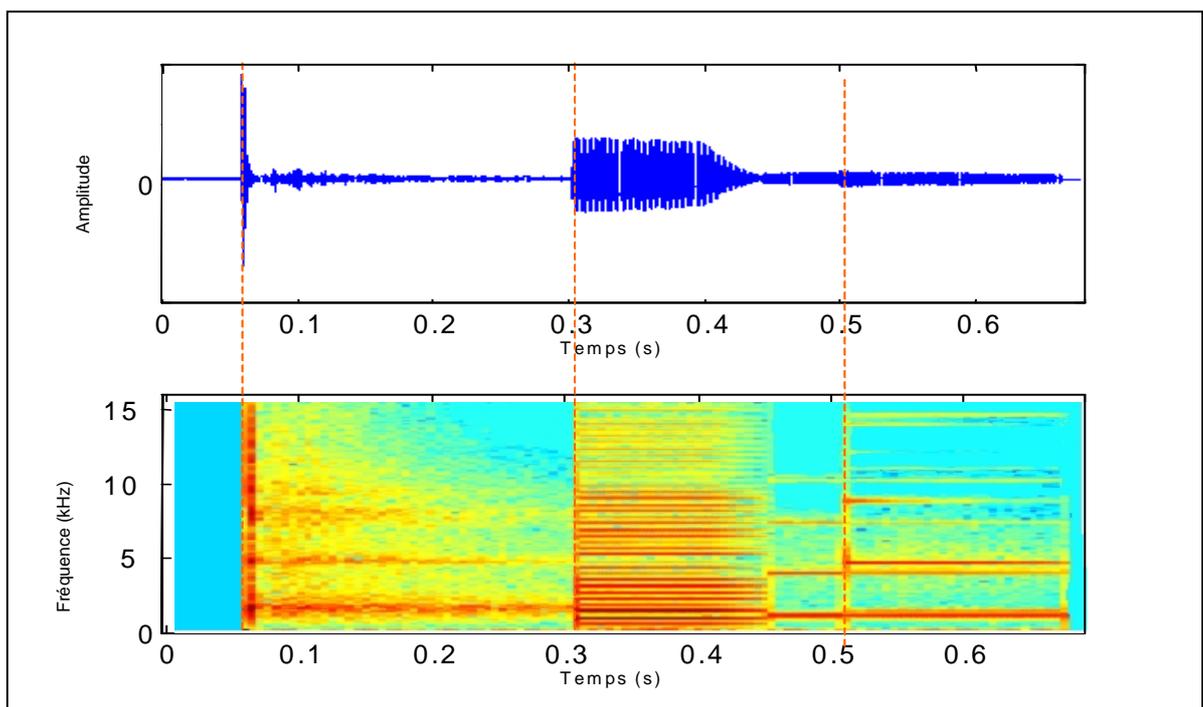


Figure 5.2 : S_{Trans}

Les deux premières attaques (situées en 0.05s et 0.3s) correspondent respectivement à une attaque de castagnette et à une attaque de clavecin. La dernière attaque (en 0.5s), correspondant à une transitoire de clochette.

5.2.2.2. Méthodes de détection d'attaque

5.2.2.2.1. Méthodes temporelles

Simple et facile à mettre en oeuvre, les méthodes temporelles consistent à suivre l'évolution de l'énergie du signal. Un découpage de l'axe temporel en trames de quelques millisecondes permet de détecter les brusques variations d'énergie du signal.

Les deux premières attaques sur le spectrogramme Figure 5.2 sont facilement détectables par des techniques temporelles puisque l'énergie du signal varie fortement en des temps brefs.

Ce type de critère est cependant insuffisant car des modifications importantes du spectre peuvent se produire dans des zones fréquentielles particulières, sans variation notable de l'énergie totale du signal. Il est par exemple impossible, avec un tel critère, de détecter les attaques en haute-fréquence (attaque

de clochette sur la Figure 5.2 située en 0.5s) puisque les variations d'énergies en haute-fréquence sont totalement masquées par la stationnarité du signal en basse-fréquence.

5.2.2.2.2. Méthodes fréquentielles

La détection de signaux transitoires développée repose sur un quadrillage temps/fréquence du signal. On donne ici l'implémentation réalisée pour des signaux échantillonnés à 48 kHz.

Le signal original est segmenté en trames d'analyse de 1024 échantillons (21,33ms) et une décision stationnaire/transitoire est prise pour chacune de ces trames.

La trame d'analyse est découpée en 8 sous-trames de $N=128$ échantillons (2,66ms). Chacune des sous-trames est ensuite pondérée par une fenêtre de Hanning définie par :

$$Window_{Hann}[i] = 0.5 * (1 - \cos(2 * \pi * i / N)) \quad , \quad 0 \leq i < N \quad (5.1)$$

avant d'être découpée en 8 sous-bandes fréquentielles de 3 kHz de large. Le passage dans le domaine fréquentiel est ici réalisé par une transformée de Fourier discrète sur les signaux fenêtrés.

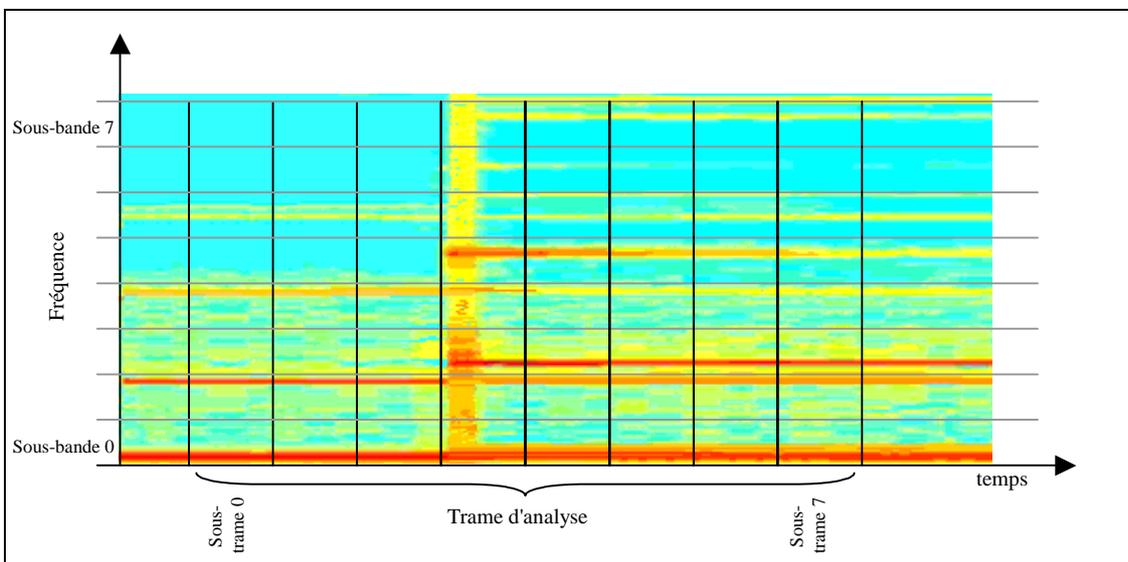


Figure 5.3 : Détection d'attaque sur un signal de clochette

Ce quadrillage permet de suivre l'évolution temporelle de chacune des sous-bandes et offre ainsi la possibilité de détecter les transitoires dans n'importe quelle partie du spectre.

Une sous-bande est considérée comme transitoire lorsqu'on détecte une différence d'énergie supérieure à un seuil (typiquement 20 dB) entre une sous-trame et au moins l'une des quatre précédentes.

Finalement, la trame d'analyse est jugée transitoire lorsque au moins une sous-bande est marquée comme transitoire.

On représente Figure 5.3 le quadrillage temps/fréquence réalisé sur le spectrogramme de la dernière attaque de S_{Trans} (signal de clochette). Les sous-bandes 0, 1, 3, 5, 6, 7 sont considérées comme stationnaires par le détecteur d'attaque (variation d'énergie de sous-trames en sous-trames inférieure à 20dB). Les sous-bandes 2 et 4 voient en revanche leur énergie varier fortement lors du passage de la sous-trame 2 à la sous-trame 3. Cette trame est alors considérée comme transitoire.

5.2.2.3. Gestion des fenêtres d'analyse

Le module de détection d'attaque contrôle le type de fenêtre d'analyse utilisé par les transformées temps/fréquences requises dans le codeur/décodeur d'extension de bande. Rappelons que cette gestion des fenêtres est essentielle afin de réduire le pré-écho sur les signaux transitoires.

Ainsi, si le signal est stationnaire, la transformée utilise des fenêtres d'analyse longues (2048 échantillons, soit 42,66ms pour des signaux échantillonnés à 48 kHz), c'est-à-dire du double de la longueur de la trame (1024 échantillons). Le chevauchement des fenêtres est ici de 50%.

Si une transition est détectée à l'intérieure de la trame, la transformée temps/fréquence utilise des fenêtres huit fois plus courtes, soit d'une longueur de 5,33ms (pré-écho inaudible).

La forme et la gestion des fenêtres d'analyse et de synthèse sont celles retenues dans la norme MPEG AAC [ISO 01a] et sont de la forme :

$$h_a(n) = h_{s(n)} = \sin \left[\left(n + \frac{1}{2} \right) \frac{\pi}{2M} \right], \quad 0 \leq n \leq 2M - 1, \quad (5.2)$$

Où M désigne la taille de la transformée.

La Figure 5.4 illustre les différents types de fenêtre appliqués au cours du temps en fonction de la nature du signal (partie stationnaire et transitoire)

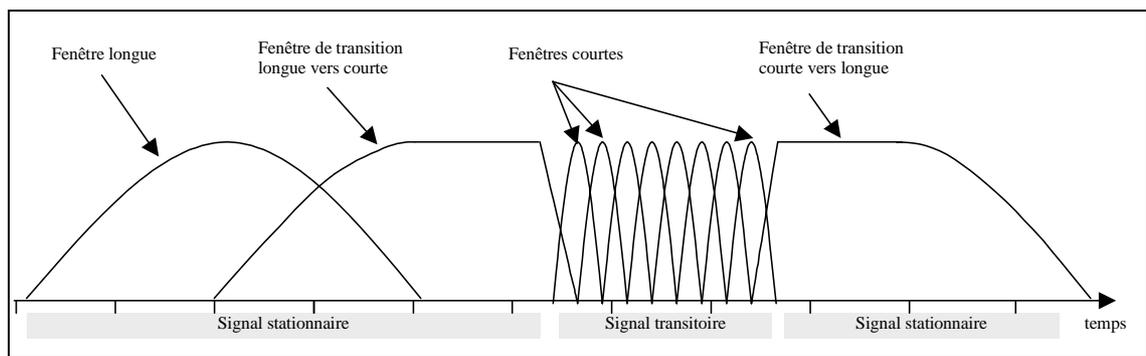


Figure 5.4 : Agencement des fenêtres lors d'un changement de taille de fenêtre

On dispose ainsi de quatre types de fenêtre (longue, longue vers courte, courte et courte vers longue). Cette information est utilisée au décodeur dans les modules de translation et d'ajustement d'enveloppe et doit donc être transmise dans le train binaire.

5.2.3. Détection d'harmonicité

5.2.3.1. Introduction

L'harmonicité d'un signal, telle que nous la définissons ici, correspond au rapport entre le niveau d'énergie des tonales par rapport au niveau d'énergie du bruit. Cette information est essentielle dans la technique d'enrichissement de spectre développée puisqu'elle permet de contrôler le contenu fréquentiel des sous-bandes synthétisées. Ce critère offre ainsi la possibilité:

- De savoir si les sous-bandes synthétisées en haute-fréquence ont la même structure fine que celles présentes sur le signal original, contrôlant ainsi le rapport tonal sur bruit des sous-bandes synthétisées en haute-fréquence par injection éventuelle de bruit ou par modification de l'ordre du filtre de blanchiment (point développé au paragraphe 5.2.4.4)
- D'adapter l'ordre de prédiction linéaire pour la modélisation d'enveloppe par prédiction linéaire afin d'éviter le phénomène d'accrochage des harmoniques sur les signaux à fortes composantes tonales.

Pour ce faire, nous avons développé un outil susceptible d'analyser finement l'harmonicité des signaux.

5.2.3.2. Méthodes de détection d'harmonicité

De nombreuses méthodes de détection d'harmonicité ont été développées dans la littérature; Nous rejetons ici les techniques temporelles (technique basée sur l'autocorrélation du signal [RAB 77], sur le nombre de passage par zéro (zero-crossing), la technique AMDF (Average magnitude difference function) décrite dans [LAR 95]). Ces méthodes, plutôt dédiées à l'analyse des signaux composés d'une seule fréquence fondamentale (parole notamment), donnent des résultats imprécis dans notre cadre d'étude.

La mesure de platitude spectrale (SFM, Spectral Flatness Measure [JOH 88]), définie par le rapport entre la moyenne arithmétique et la moyenne géométrique de la densité spectrale de puissance du signal, est en revanche moins sensible au bruit et donne une bonne indication sur le degré d'harmonicité des signaux. La mesure de platitude spectrale, comprise entre 0 et 1, prend des valeurs proche de 1 pour des signaux décorrélés (bruit) et proche de 0 pour des signaux harmoniques. Cette mesure n'offre toutefois pas la possibilité d'analyser des bandes spectrales avec précision et ne permet notamment pas de savoir le nombre exact de tonales présentes dans le signal. Une analyse précise est en effet incontournable pour réaliser un blanchiment spectral efficace (paragraphe 3.3.4.2).

5.2.3.3. Méthode retenue

La méthode retenue repose sur une analyse paramétrique du signal dans le domaine spectral. L'analyse consiste à repérer les pics qui ressortent du spectre dans le domaine DFT. La technique utilisée est similaire à celle développée dans le modèle psychoacoustique numéro 1 de la norme MPEG-1 [ISO 92].

On ne considère comme composantes sinusoïdales que les maximums du spectre dont l'amplitude est supérieure à 7 dB par rapport aux raies adjacentes ou voisines.

Une fois un maximum détecté, on considère que l'énergie de cette tonale est répartie sur la raie déterminée et sur les deux raies adjacentes (principe illustré Figure 5.5).

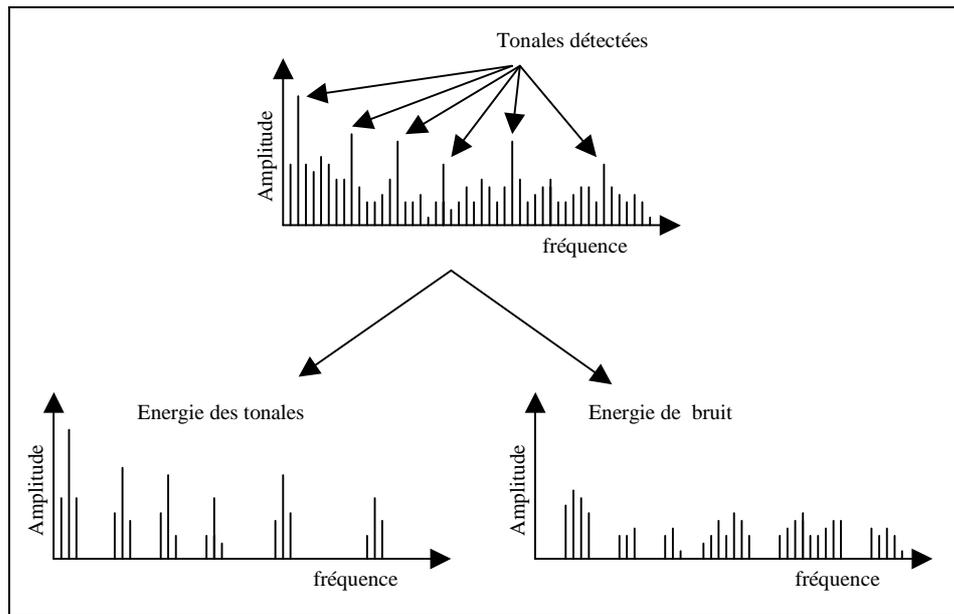


Figure 5.5 : Estimation de l'harmonicité d'un signal

On estime de cette façon l'énergie concentrée par toutes les tonales présentes dans le signal; l'énergie de bruit correspond alors à l'énergie totale du signal moins l'énergie des tonales.

En comparant l'harmonicité des parties basses et haute-fréquence du signal original, on détermine l'ordre de blanchiment requis pour la synthèse des hautes fréquences. Les ordres de blanchiment choisis sont explicités au paragraphe 5.2.5.2.3.

5.2.4. Module d'extension de la structure fine

La technique d'extension de la structure fine par des opérations de translation dans le domaine transformé a été retenue pour sa souplesse d'utilisation et ses performances. On développe ici les détails de la technique concernant le type de translation utilisé, la longueur de la bande à traduire, le comportement de la méthode sur les signaux harmoniques et les ajustements requis afin de réaliser une extension de bande efficace.

Nous développons enfin les détails de l'implémentation de ce module dans la technique PAT.

5.2.4.1. Types de translation

La Figure 5.6 illustre la technique d'enrichissement de spectre par translations spectrales.

- Le premier spectre correspond au signal original pleine bande (signal de parole non-voisée)
- Le second spectre correspond au signal synthétisé par translations spectrales directes. La bande comprise entre 0 et 5 kHz est celle de l'original (signal cœur), et les hautes fréquences au-delà de 5 kHz sont obtenues par 3 translations successives de la bande [1,5-5kHz]. Le choix de cette largeur de bande est justifié au paragraphe suivant.
- Le troisième spectre correspond au signal synthétisé par translations spectrales inversées. La bande comprise entre 0 et 5 kHz est celle de l'original (signal cœur). La bande [5-8,5kHz] est obtenue par translation inversée de la bande [1,5-5kHz], la bande [8,5-12kHz] est obtenue par translation directe de la bande [1,5-5kHz], et enfin la bande [12-15.5kHz] est obtenue par translations inversée de la bande [1,5-5kHz].

On observe ainsi 3 transitions sur la Figure 5.6 (droite en pointillée) située respectivement en 5, 8,5 et 12 kHz. Il apparaît clairement, sur le spectre du milieu, des ruptures d'énergies (sauts d'énergie de plus de 10 dB) au niveau des transitions, contrairement au spectre du bas qui est naturellement "lissé".

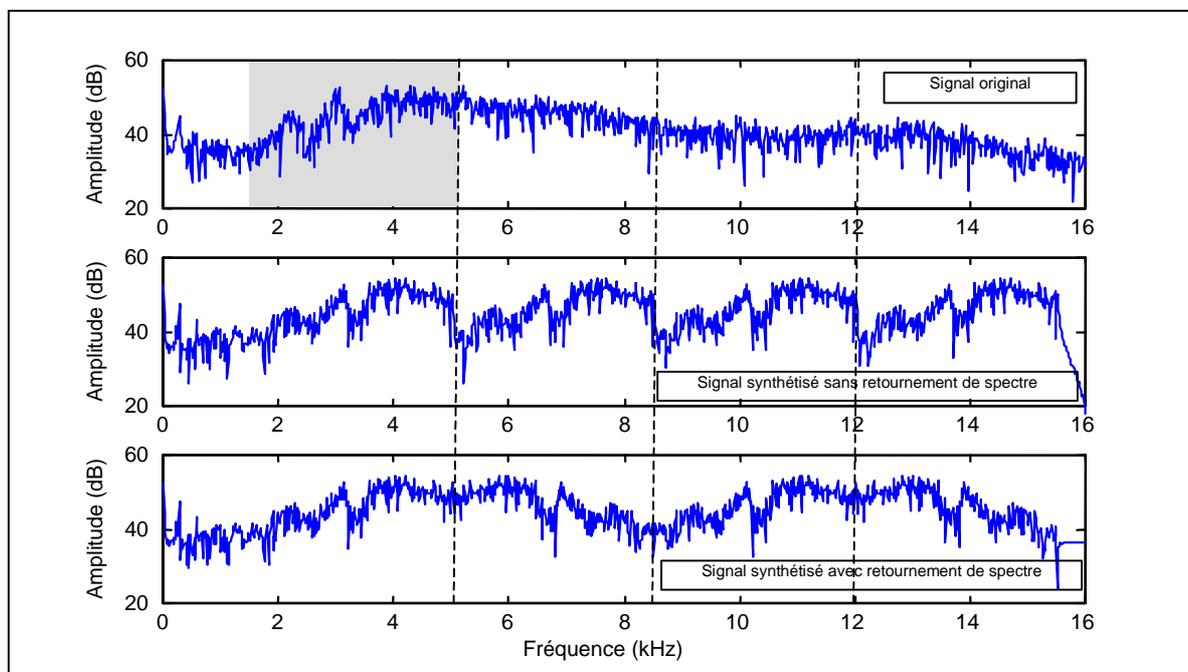


Figure 5.6 : Translations spectrales avec et sans retournement de spectre

La translation par inversion de spectre a donc l'avantage de garder une continuité dans l'énergie des hautes fréquences synthétisées.

5.2.4.2. Choix de la bande basse-fréquence à traduire

Les signaux numériques, et en particulier les signaux de parole voisée, ont un niveau de bruit faible en basse-fréquence (par rapport au niveau des tonales). Cette caractéristique est illustrée sur la Figure 5.7 qui représente le spectre d'un signal de parole chantée et voisée. La bande comprise entre 0 et 1 kHz, très harmonique (fortement voisée), est très éloignée du reste du spectre.

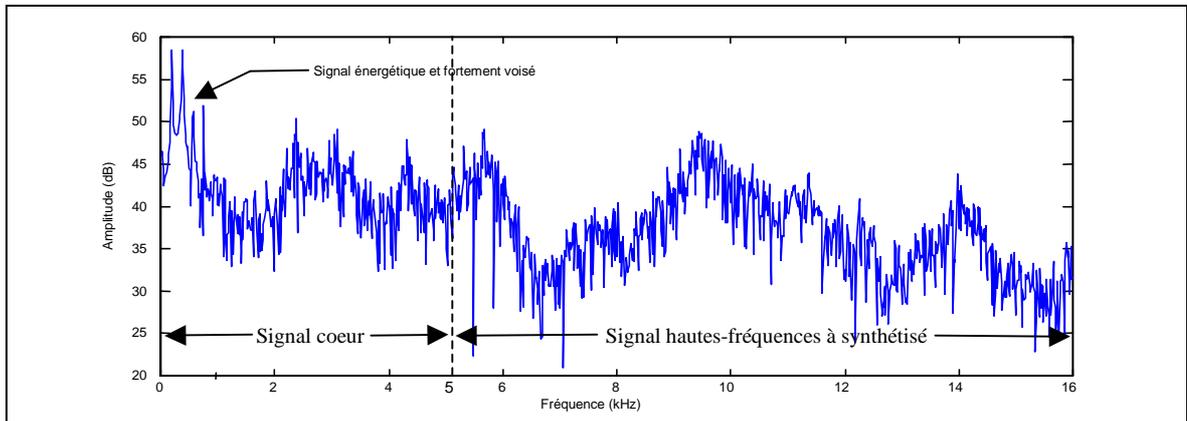


Figure 5.7 : Spectre d'une trame de parole chantée voisée (Suzanne Vega)

Des tests subjectifs réalisés auprès de 5 experts audio sur de nombreuses séquences de parole et de musique, ont permis de déterminer quelle partie du spectre basse-fréquence traduire afin d'obtenir le meilleur rendu sonore. Pour des signaux coupés à 5 kHz, et pour un enrichissement de spectre jusqu'à 16 kHz, le choix s'est porté sur la translation de la bande [1,5-5kHz].

Une largeur de bande plus longue ([0-5kHz]) accentue le phénomène de "buzzing" décrit au paragraphe 4.4.1.1 puisque le contenu [0-1,5kHz] se trouve replié en [8,5-10kHz].

Une largeur de bande plus étroite ([3-5kHz] par exemple) augmente le nombre de translations nécessaire à la synthèse de la bande [5-16kHz]. Le rendu sonore se dégrade alors sur les signaux harmoniques puisque cette augmentation du nombre de translations accroît le nombre de ruptures d'harmonicité, et augmente les risques de dissonances (paragraphe 5.2.4.3).

5.2.4.3. Comportement sur les signaux harmoniques

Sur les signaux bruités et/ou faiblement harmoniques, la technique développée précédemment étend de manière efficace la structure fine du spectre, sans dégradation gênante.

L'extension de bande par des opérations de translations fréquentielles requiert en revanche un ajustement sur les signaux fortement harmoniques afin d'éviter les phénomènes de dissonances. Les translations de spectre étant réalisées indépendamment du pitch (la détection et gestion de la continuité étant très délicates à mettre en oeuvre sur les signaux musicaux composés de plusieurs fréquences fondamentales, paragraphe 4.3), il se crée dans la majorité des cas des ruptures d'harmonicité au niveau des transitions; le cas d'une translation de spectre avec rupture d'harmonicité en 5 kHz est illustré Figure 5.8.

Afin de mesurer l'effet de ce phénomène sur les signaux harmoniques, on a réalisé l'expérience psychoacoustique décrite ci-dessous.

On cherche dans cette expérience à connaître l'influence des ruptures d'harmonicité en 5 kHz sur deux signaux synthétiques.

- Le premier, S_{333} est composé d'un peigne de sinus de mêmes amplitudes, de fréquence fondamentale 333 Hz et de 10 kHz de bande passante (30 tonales).
- Le second, S_{700} est composé d'un peigne de sinus de mêmes amplitudes, de fréquence fondamentale 700 Hz et de 10 kHz de bande passante (14 tonales).

Pour chacun de ces signaux, on génère une rupture d'harmonicité (ou déviation harmonique) en 5 kHz; en d'autres termes, on décale la série harmonique de quelques dizaines de Hertz comme illustré Figure 5.8 sur le signal S_{700} .

Notons que, dans la technique PAT présentée au paragraphe 5.2, cet exemple correspond à une translation de spectre inversée autour de 5 kHz.

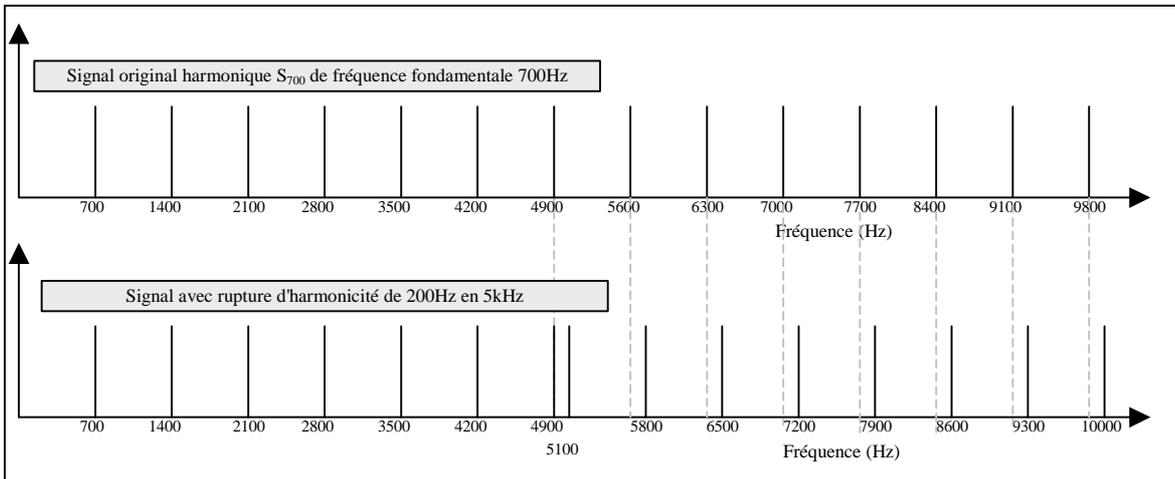


Figure 5.8 : Rupture d'harmonicité

Sur les deux signaux S_{333} et S_{700} , on génère quatre déviations d'harmonicité, de fréquences respectives 20, 50, 100 et 200 Hz.

La Figure 5.9 présente les résultats du test d'écoute réalisé auprès de six sujets experts en tests d'écoute audio. Pour chacun des signaux, le test consistait à noter, sur une échelle de qualité de 1 (mauvaise qualité) à 5 (qualité parfaite), le signal inharmonique dévié par rapport au signal original. Les données correspondent aux valeurs moyennes assorties des intervalles de confiance à 95%.

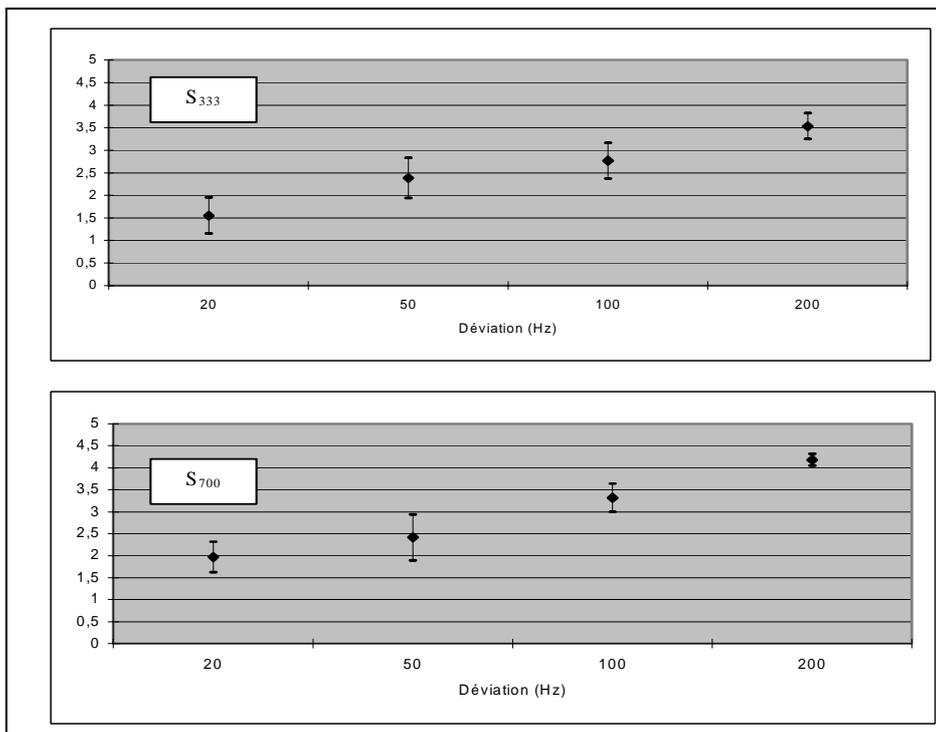


Figure 5.9 : Perception des ruptures d'inharmonicité

Pour des déviations faibles, inférieures à 100 Hz, la qualité du signal est fortement dégradée. Les deux tonales de fréquences très proches interagissent entre elles et génèrent un phénomène de battement gênant.

Il ne s'agit pas, à proprement parlé, d'un phénomène de dissonance, comme défini au paragraphe 2.2.4. Pour une déviation harmonique de 20Hz par exemple, les deux tonales sont en effet comprises dans moins de 5% de la largeur de bande critique dans laquelle elles sont situées, et sont donc perçues comme une seule raie (paragraphe 2.2.4.1).

Cette interaction correspond plutôt à une modulation du signal en amplitude et donc à un phénomène de battement temporel.

Pour une déviation de 200 Hz, les ruptures d'harmonicité sont en revanche moins perceptibles et la qualité de restitution devient tout à fait acceptable. Au-delà de 200 Hz, on supprime une harmonique sur les signaux de faible fréquence fondamentale et on s'éloigne alors perceptivement des signaux originaux.

Le choix s'est donc porté sur une largeur de déviation harmonique de 200 Hz dans le module de translations spectrales. Dans le codeur d'enrichissement de bande développé, on atténue ainsi l'énergie au niveau des transitions inter-blocs, sur une largeur de 200 Hz.

5.2.4.4. Nécessité de blanchir le signal synthétisé

Les signaux audionumériques ont en général en haute-fréquence un niveau de bruit plus élevé qu'en basse-fréquence et par conséquent une structure harmonique moins prononcée (signal de parole Figure 5.10). Lors de l'extension de bande, les opérations de translations conservent la structure fine du spectre et donc l'harmonicité du signal. Le niveau de bruit et le rapport tonales sur bruit étant les mêmes en hautes et basse-fréquence, les signaux se retrouvent "sur-harmonisés" et cela entraîne le phénomène de "buzzing" décrit au paragraphe 4.4.1.1

En revanche, sur les signaux harmoniques purs et sur les signaux à fortes composantes tonales, la structure fine et l'harmonicité du signal est comparable sur tout le spectre (signal de clavecin Figure 5.10); un ordre de blanchiment faible (voire nul) devient nécessaire pour une extension de bande efficace sur ce type de signaux.

Le blanchiment spectral du signal haute-fréquence synthétisé permet de jouer sur le rapport d'harmonicité. En modifiant l'ordre du filtre de blanchiment, on est ainsi capable de conserver ou diminuer le rapport tonal à bruit.

Sur les signaux de parole et de musique complexe (signaux bruités et/ou faiblement tonals), un ordre de blanchiment élevé diminuera l'énergie des tonales et augmentera le niveau de bruit. En translatant ce résidu vers les hautes fréquences, on approchera la structure fine du spectre haute-fréquence originale.

Sur les signaux harmoniques purs en revanche, on utilisera un ordre de blanchiment faible afin de ne pas entacher la structure harmonique du spectre. Un ordre trop élevé aura en effet tendance à atténuer, voire supprimer certains partiels (problème abordé paragraphe 3.3.4.2).

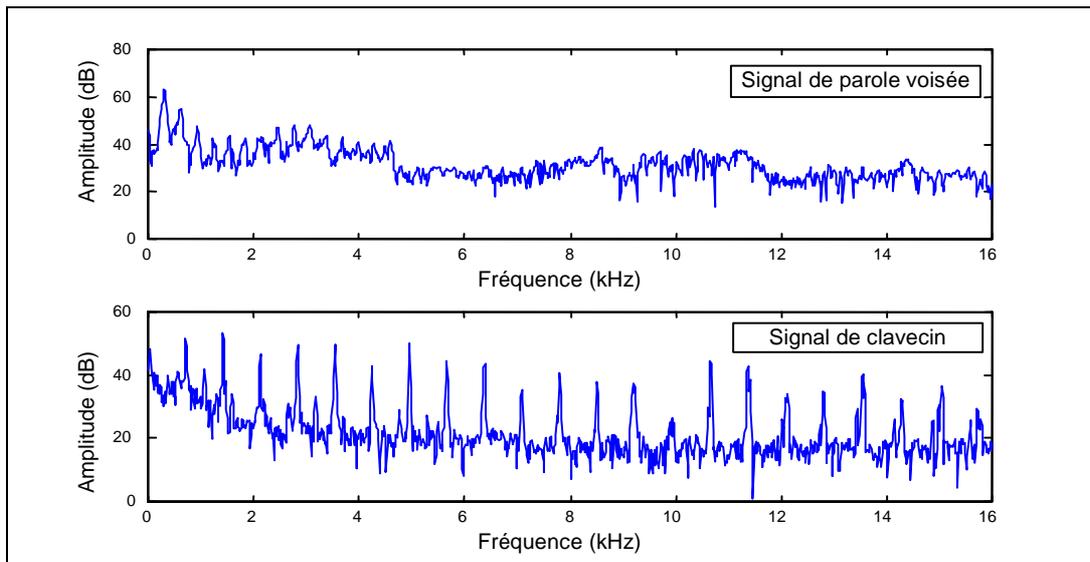


Figure 5.10 : Corrélations hautes-basses fréquences sur un signal de parole et de musique

Dans le PAT, cet ordre de blanchiment est déterminé au codeur sur le signal original, grâce au module d'analyse harmonique, avant d'être transmis au décodeur

5.2.4.5. Implémentation du module d'extension de la structure fine

Conformément au diagramme proposé Figure 5.11, on synthétise, à partir d'un signal cœur de 5 kHz de bande et échantillonné à 24 kHz, un signal pleine-bande échantillonné à 24 kHz. Ce signal de 12 kHz de bande complètera le signal cœur afin de produire le signal complet à 48 kHz de 17 kHz de bande passante. Une translation de spectre sera donc requise en sortie du module d'extension afin de générer le signal haute-fréquence désiré.

Le module d'extension est utilisé au décodeur et prend en entrée :

- Le signal cœur décodé
- Le type de fenêtre utilisé par les transformées temps/fréquence (DFT sur cet exemple)
- L'ordre de blanchiment à réaliser sur le signal synthétisé afin de contrôler le rapport tonales sur bruit.

Notons que ces deux derniers paramètres sont déterminés au codeur et transmis dans le train binaire PAT.

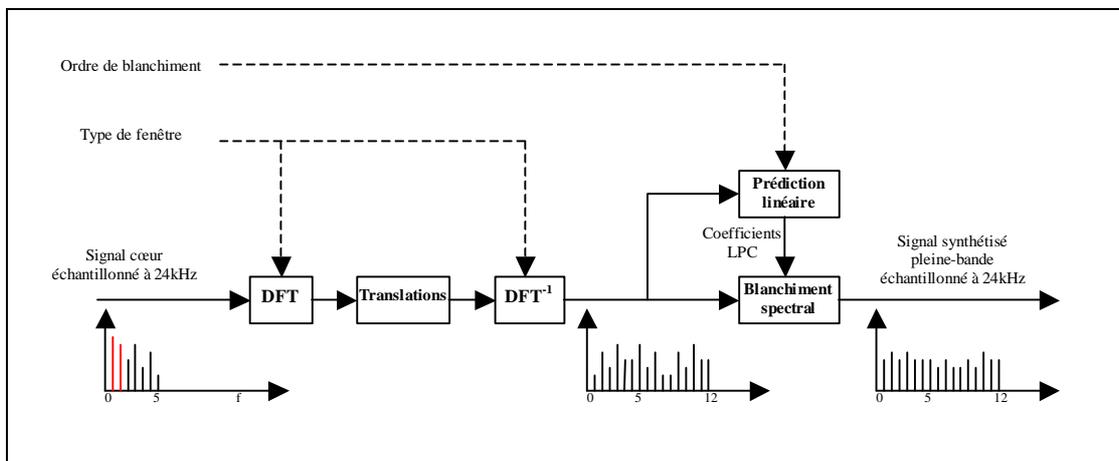


Figure 5.11 : Module d'extension de la structure fine

On utilise ici des fenêtres d'analyse de 512 échantillons pour les signaux stationnaires et des fenêtres 8 fois plus courtes (64 échantillons) pour les signaux transitoires, afin de limiter les artefacts générés par la transformée temps/fréquence.

Une fois le signal passé dans le domaine transformée, on réalise 4 translations de spectre de la bande [1,5-5kHz], selon la méthode décrite au paragraphe 5.2.4.1 (translations directe/inversée afin de conserver une continuité d'énergie dans le spectre). On réalise ensuite une DFT inverse sur ce spectre synthétisé, générant ainsi le signal pleine-bande échantillonné à 24 kHz.

On effectue enfin une prédiction linéaire sur le signal synthétisé. L'ordre de prédiction est déterminé par le codeur et transmis dans le train binaire. Une fois les coefficients LPC calculés, ceux-ci sont utilisés pour blanchir le signal (principe du blanchiment spectral décrit au paragraphe 3.3.1.2).

On obtient ainsi en sortie du module d'extension de la structure fine, le signal de 12 kHz de bande correspondant aux hautes fréquences comprises entre 5 et 17 kHz.

5.2.4.6. Conclusion

Les considérations développées dans ce paragraphe nous ont permis de choisir une technique d'extension de spectre efficace sur la majorité des signaux. La stratégie adoptée consiste donc :

- A traduire non pas la totalité de la bande basse du codeur cœur, mais seulement une partie du spectre afin de limiter les phénomènes de "buzzing".
- A copier le spectre plusieurs fois par inversion successive des différents blocs afin de garder une continuité d'énergie dans le spectre synthétisé
- A atténuer l'énergie aux transitions sur une largeur de bande de 200 Hz afin d'éviter les phénomènes de dissonance et de battement.
- A blanchir le signal afin de contrôler le rapport tonal sur bruit du signal synthétisé par rapport au signal original.

Une fois la structure fine haute-fréquence synthétisée, il reste à remettre en forme spectralement ce signal afin d'approcher au mieux l'enveloppe du signal original. Ce point fait l'objet du paragraphe suivant.

5.2.5. Module d'estimation et d'ajustement d'enveloppe

5.2.5.1. Introduction

Nous reprenons dans ce paragraphe les deux techniques retenues au chapitre 3, à savoir :

- L'estimation et l'ajustement d'enveloppe par prédiction linéaire.
- L'estimation et l'ajustement d'enveloppe par facteurs d'échelle dans le domaine fréquentiel.

Rappelons que ces deux méthodes offrent un moyen efficace de transmettre l'enveloppe haute-fréquence à bas débit et s'avèrent donc intéressantes pour les applications d'extension de bande. Nous tentons d'étudier dans ce paragraphe l'influence de ces deux techniques sur la méthode complète d'enrichissement de spectre.

Pour chacune des deux méthodes, on donne une description précise du codeur, de la quantification des descripteurs d'enveloppe, du formatage du train binaire et enfin du fonctionnement du décodeur. Des tests d'écoute nous permettront de choisir la solution la plus adaptée.

Rappelons que le rôle du codeur consiste à :

- Analyser le signal à coder (signal stationnaire, transitoire, harmonique...)
- Estimer l'enveloppe spectrale haute-fréquence du signal original
- Quantifier les descripteurs d'enveloppe (coefficients de filtre dans le cas de la prédiction linéaire et facteurs d'échelle dans le cas de la modélisation d'enveloppe en sous-bandes)
- Formater le train binaire afin de transmettre toutes les informations requises par le décodeur

Le rôle du décodeur consiste à :

- Lire les informations transmises dans le train binaire par le codeur
- Ajuster l'énergie du spectre haute-fréquence synthétisé par le module d'extension de la structure fine afin d'approcher au mieux l'enveloppe du signal original. L'énergie du spectre est ajustée par les descripteurs d'enveloppes transmis dans le train binaire.
- Sommer les hautes fréquences synthétisées au signal cœur à bande limitée dans le but de générer le signal de sortie large-bande.

5.2.5.2. Estimation et ajustement d'enveloppe par prédiction linéaire

La première version du codeur d'enrichissement de spectre reposait sur une technique d'estimation d'enveloppe par prédiction linéaire (cette version fut présentée dans le cadre de la normalisation DRM).

5.2.5.2.1. Principe

Comme nous l'avons vu au paragraphe 3.3, le principe consiste à ramener la partie haute-fréquence à modéliser en bande de base, à modéliser l'enveloppe du signal ainsi translaté, à transmettre les coefficients LSP correspondant et à les appliquer au décodeur pour remettre en forme le signal extrapolé par les opérations de translations.

5.2.5.2.2. Diagrammes de fonctionnement

5.2.5.2.3. Codeur

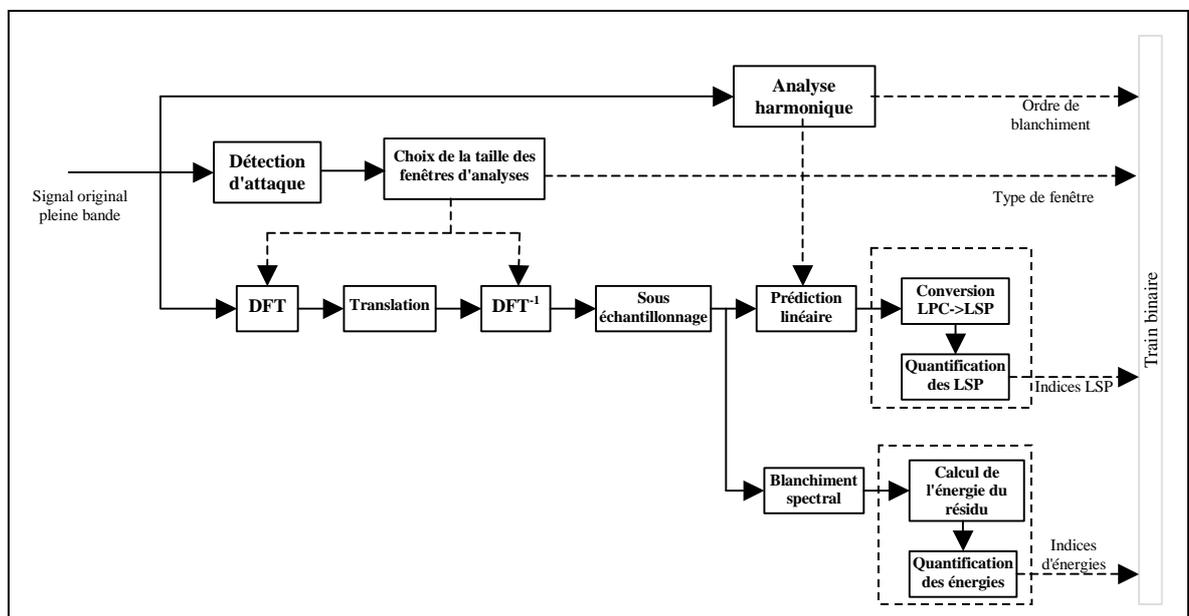


Figure 5.12 : Diagramme de fonctionnement du codeur par prédiction linéaire

Le signal original pleine-bande, échantillonné à 48 kHz, est segmenté en trame d'analyse de 1024 échantillons.

Une détection d'attaque réalisée sur la trame d'analyse permet de choisir la taille des fenêtres utilisées pour traduire le spectre haute-fréquence en bande de base. On réalise également une analyse harmonique (paragraphe 5.2.3) afin de déterminer l'ordre de prédiction linéaire à utiliser pour blanchir le signal au décodeur et pour déterminer l'ordre de prédiction linéaire à utiliser pour estimer l'enveloppe spectrale du signal original.

Le signal original haute-fréquence à modéliser est ensuite translaté en bande de base, puis sous-échantillonné. On utilise pour ce faire une transformée de Fourier discrète. Dans le cadre d'une modélisation d'enveloppe de la bande au-delà de 5 kHz par exemple, on translate la bande haute-fréquence [5-17kHz] en [0-12kHz] puis on la sous-échantillonne à 24 kHz.

On effectue ensuite une prédiction linéaire sur ce signal pleine bande. L'ordre de prédiction est déterminé par le module d'analyse harmonique et par le module de détection d'attaque :

- Pour les signaux transitoires, vu les contraintes de débit imposées, le choix s'est porté sur un ordre de 8 (paragraphe 3.3.4.1)

- Pour les signaux à fortes composantes tonales, le choix s'est porté sur un ordre de prédiction de 8 afin de limiter le phénomène d'accrochage des tonales (paragraphe 3.3.4.2)
- Pour les autres signaux, le choix s'est porté sur un ordre de prédiction de 20 (paragraphe 3.3.4.2)

Les coefficients LPC ainsi déterminés sont ensuite convertis en coefficients LSP, conformément à la technique fournie en annexe B, puis quantifiés (quantification vectorielle décrite au paragraphe 3.3.2.1).

- Pour un ordre de filtre égal à 8, le choix s'est porté sur une quantification en deux dictionnaires de 4 LSP de 1024 vecteurs en entrée (quantification des 4 premiers LSP sur 10 bits et des 4 derniers sur 10 bits également). Notons que l'on utilise un dictionnaire propre aux signaux harmoniques et un second dictionnaire adapté aux signaux transitoires. Les 8 coefficients LSP sont ainsi quantifiés sur 20 bits.
- Pour un ordre de filtre égal à 20, le choix s'est porté sur une quantification en trois dictionnaires respectivement de 5, 5 et 10 LSP de 1024 vecteurs en entrée. Les 20 coefficients LSP sont ainsi quantifiés sur 30 bits.

Une fois la prédiction linéaire réalisée, on blanchit le signal à l'aide des coefficients LPC déterminés. On estime ensuite l'énergie du signal résiduel :

- Dans le cas d'un signal stationnaire, on calcule l'énergie sur toute la trame (un seul facteur d'énergie pour la trame de 21,33ms).
- Dans le cas d'un signal transitoire, la trame est divisée en quatre sous-trames de 5,33ms et l'énergie pour chacune de ces sous-trames est estimée. Ces facteurs d'énergie sont ensuite quantifiés sur 6 bits chacun sur une échelle logarithmique.

5.2.5.2.4. Train binaire

Mode	Harmonique	Transitoire	Normal
Informations			
Ordre de blanchiment	2 bits		
Type de fenêtre	2 bits		
indices LSP	2*10 bits	2*10 bits	3*10 bits
Facteurs d'énergie	1*6 bits	4 * 6 bits	1 * 6 bits
Total	30 bits	48 bits	40 bits
Débit associé	1,4 kbit/s	2,25 kbit/s	1,88 kbit/s

Tableau 5.1 : Débit associé aux différents modes de fonctionnement

5.2.5.2.5. Décodeur

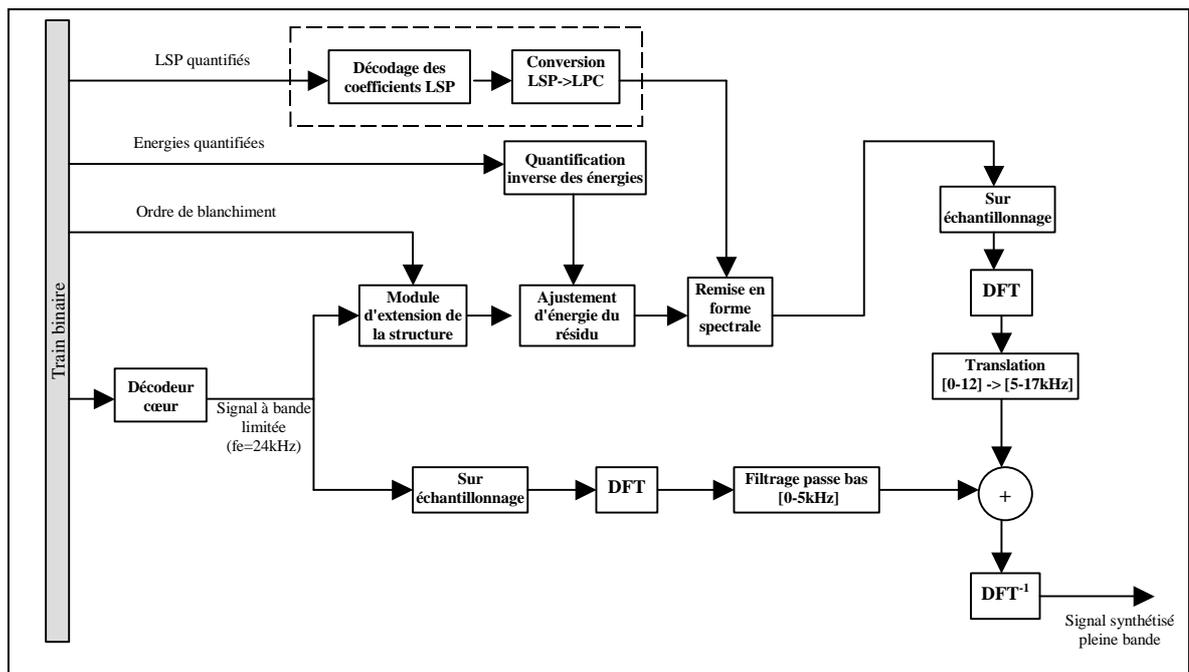


Figure 5.13 : Diagramme de fonctionnement du décodeur

Le train binaire génère les informations requises par le PAT pour synthétiser les hautes fréquences, à savoir:

- L'ordre de blanchiment requis par le module d'extension de la structure fine.
- Le type de fenêtre utilisé (non représenté ici pour la clarté du schéma) par les transformées temps-fréquence pour l'extension de la structure fine et pour les différentes translations de spectre.
- Les énergies quantifiées. Après décodage, l'énergie du signal résiduel (en sortie du module d'extension de la structure fine) est ajustée.
- Les indices correspondant aux LSP quantifiés. Les coefficients LPC déduits de ces coefficients LSP permettent de remettre en forme le signal ajusté en énergie décrit précédemment.

Une fois le signal hautes-fréquence remis en forme, il est sommé au signal cœur (dans le domaine DFT), synthétisant ainsi le signal à bande élargie.

5.2.5.2.6. Problèmes rencontrés par l'approche par prédiction linéaire

5.2.5.2.6.1. *Problème de la stabilité de l'ordre du filtre de prédiction*

Selon l'harmonie du signal original haute-fréquence, l'estimation d'enveloppe par prédiction linéaire requiert un ajustement de l'ordre de prédiction (point abordé au chapitre 3.3.4.2).

Le problème d'une telle approche est la gestion du changement d'ordre du filtre au cours du temps. Elle s'avère délicate à mettre en œuvre et les changements d'ordre engendrent le plus souvent des sauts de phase et donc des discontinuités temporelles (lissage difficile à réaliser avec des filtres d'ordre différents).

5.2.5.2.6.2. *Problème des discontinuités temporelles dues à la quantification des paramètres LPC*

Sur les signaux musicaux fortement harmoniques, les amplitudes des partiels varient très lentement au cours du temps. Cette faible variation de l'enveloppe spectrale engendre une faible variation de la position des pôles du filtre de prédiction linéaire. Les coefficients LPC, une fois quantifiés, sont alors

susceptibles de varier fortement d'une trame à l'autre, créant là encore des discontinuités temporelles à la synthèse. Sur les signaux de parole et sur les signaux faiblement harmoniques en général, le lissage temporel des paramètres LSP (paragraphe 3.3.2.2) permet d'atténuer ces discontinuités qui deviennent imperceptibles. Sur les signaux purement harmoniques en revanche, l'oreille étant plus sensible aux variations d'amplitude, ce lissage n'est pas suffisant et les artefacts restent audibles.

5.2.5.2.7. Conclusion

Sur les signaux de parole, la prédiction linéaire modélise parfaitement la structure formantique des signaux, et ceci pour un ordre de prédiction fixe au cours du temps. La technique offre alors de bons résultats dans le schéma complet d'extension de bande. Sur les signaux musicaux en revanche, et plus particulièrement sur les signaux à fortes composantes tonales (signaux harmoniques purs), l'estimation d'enveloppe par prédiction linéaire génère des problèmes. On maîtrise en effet mal les variations de l'ordre du filtre au cours du temps, ainsi que la quantification des coefficients du filtre. La variation de la répartition fréquentielle des pôles, et donc l'enveloppe spectrale, deviennent dès lors très difficile à contrôler.

5.2.5.3. Estimation et ajustement d'enveloppe par facteurs d'échelle dans le domaine fréquentiel

Afin de résoudre les problèmes développés au paragraphe précédent, une seconde version du PAT a été développée, puis proposée en janvier 2001 dans la normalisation MPEG-4. Cette nouvelle technique repose sur une estimation et un ajustement d'enveloppe par facteurs d'échelle dans le domaine fréquentiel. Nous en étudions ici le principe général avant d'en distinguer deux types de mise en œuvre: l'une dans le domaine MDCT et l'autre dans le domaine DFT.

5.2.5.3.1. Principe

Au codeur, on réalise une transformée temps-fréquence sur le signal original. Le spectre est ensuite divisé en sous-bandes selon une échelle perceptuelle linéaire avec les Barks, variable selon le nombre de facteurs d'énergie souhaités; on calcule les différences d'énergie entre chacune des sous-bandes avant de les quantifier et de les coder par un algorithme de Huffman.

Au décodeur, on réalise une transformée temps/fréquence sur le signal cœur. On synthétise ensuite les hautes fréquences par le module d'extension de bande avant de les ajuster par les facteurs d'énergies transmis et décodés.

5.2.5.3.2. Diagramme de fonctionnement

5.2.5.3.3. Codeur

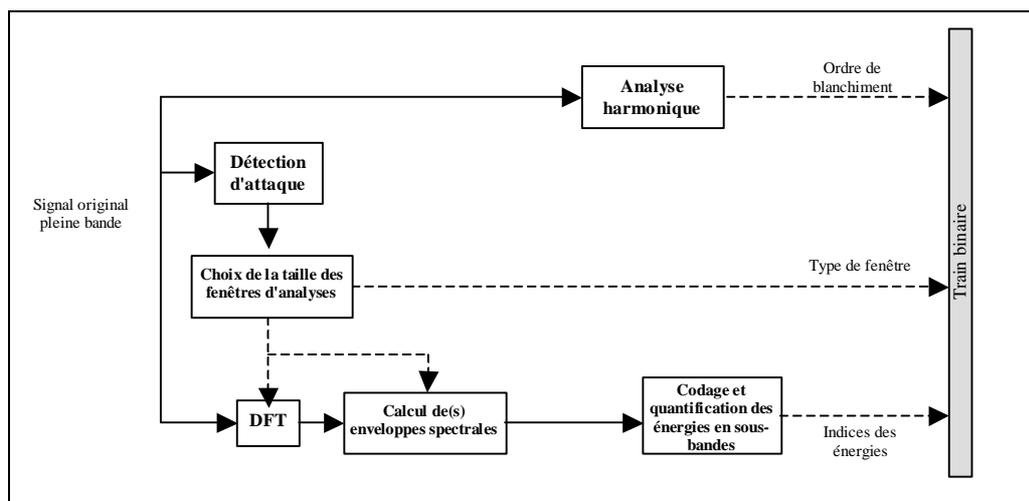


Figure 5.14 : Fonctionnement du codeur par facteurs d'échelle en sous-bandes

On réalise une détection d'attaque et une analyse harmonique sur le signal original d'entrée, conformément aux méthodes développées au paragraphe 5.2.2 et 5.2.3.

Le calcul de l'enveloppe spectrale est réalisé par un quadrillage temps/fréquence du spectre haute-fréquence. Cette grille temps/fréquence, variable dans le temps, est contrôlée par le module de détection d'attaque :

Sur les signaux stationnaires, le spectre calculé sur la trame de 1024 échantillons (21.33ms) est divisé en 8 sous-bandes (échelle linéaire en Barks) et on transmet les facteurs d'énergie associés. Le découpage fréquentiel, fonction de la fréquence de coupure du signal cœur, est donné Tableau 5.2.

Fréquence de coupure du signal ($F_{\text{échantillonnage}} = 48 \text{ kHz}$)	largeur de la sous-bande en raies spectrales (1 raie spectrale = 23,43 Hz)
$3101 \text{ Hz} < F_c < 3790 \text{ Hz}$	16, 24, 32, 48, 64, 96, 104, 128
$4234 \text{ Hz} < F_c < 6891 \text{ Hz}$	32, 32, 48, 48, 64, 64, 112, 112
$7235 \text{ Hz} < F_c$	64, 64, 64, 64, 64, 64, 64, 64

Tableau 5.2 : Découpage fréquentiel en fonction de la fréquence de coupure

La quantification des facteurs d'énergie est faite en prenant en compte le JND temporel (variation maximale de 1dB au cours du temps vue au paragraphe 2.2.4.3), si bien que les variations d'énergie des tonales sur les signaux harmoniques deviennent imperceptibles. Le choix s'est porté sur une quantification sur 6 bits, offrant ainsi une résolution fréquentielle de 1 dB pour une dynamique de 64 dB. Le premier facteur est codé isolément à l'aide d'une première table de Huffman. Les sept suivants sont codés en différentiel à l'aide d'une seconde table.

Sur les signaux transitoires, la trame de 21,33 ms est divisée en 8 sous-trames de 2,66 ms.

A résolution fréquentielle équivalente, le nombre de facteurs d'échelles à transmettre devient 8 fois plus élevé et devient alors bien supérieur au débit cible (2 kbit/s). Un compromis entre le débit et le nombre de paramètres sur les signaux transitoires a donc du être pris. Sur les transitoires, l'oreille étant très sensible aux variations temporelles du signal, le choix s'est porté sur un meilleur codage des énergies temporelles, au détriment des énergies fréquentielles. La stratégie adoptée consiste à transmettre une seule enveloppe (enveloppe moyenne sur la trame de 21.33 ms, pondérée par l'énergie du signal) et les 8 énergies en sous-trames.

On est ainsi amené à transmettre 2 jeux de 8 facteurs d'énergies. La stratégie de codage de ces facteurs est celle utilisée dans le cas stationnaire (quantification sur 6 bits suivi d'un codage différentiel et d'un codage de Huffman avec les tables adaptées aux signaux transitoires).

5.2.5.3.4. Train binaire

La syntaxe complète du train binaire utilisé dans le PAT est fournie en annexe C. Cette description correspond à celle utilisée dans le cadre de la normalisation du codeur AAC+PAT dans MPEG-4. Les paramètres correspondent à un codage des signaux monophoniques échantillonnés à 48 kHz.

Notons que l'utilisation d'un codage entropique de Huffman rend le débit variable. Pour coder les huit facteurs d'énergies fréquentielles, le débit moyen s'élève à environ 1,5 kbit (débit moyen calculé sur les 12 séquences de tests utilisées dans MPEG-4).

Pour les signaux transitoires, la transmission des deux jeux de facteurs d'énergie (temporelle et fréquentielle) requiert un débit plus élevé de l'ordre de 3 kbit/s.

Afin d'illustrer la variation de débit au cours du temps, la Figure 5.15 donne le débit requis au cours du temps par le PAT pour synthétiser la bande [5-15kHz] sur une séquence de parole chantée (séquence test MPEG-4 es01.wav, Suzanne Vega). Le débit moyen sur cette séquence s'élève à 2,1 kbit/s.

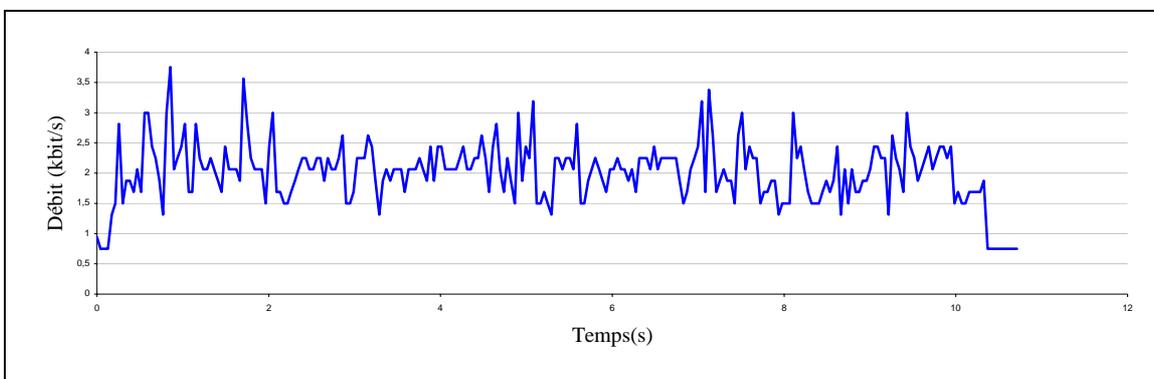


Figure 5.15 : Débit variable du PAT sur la séquence es01.wav

5.2.5.3.5. Décodeur

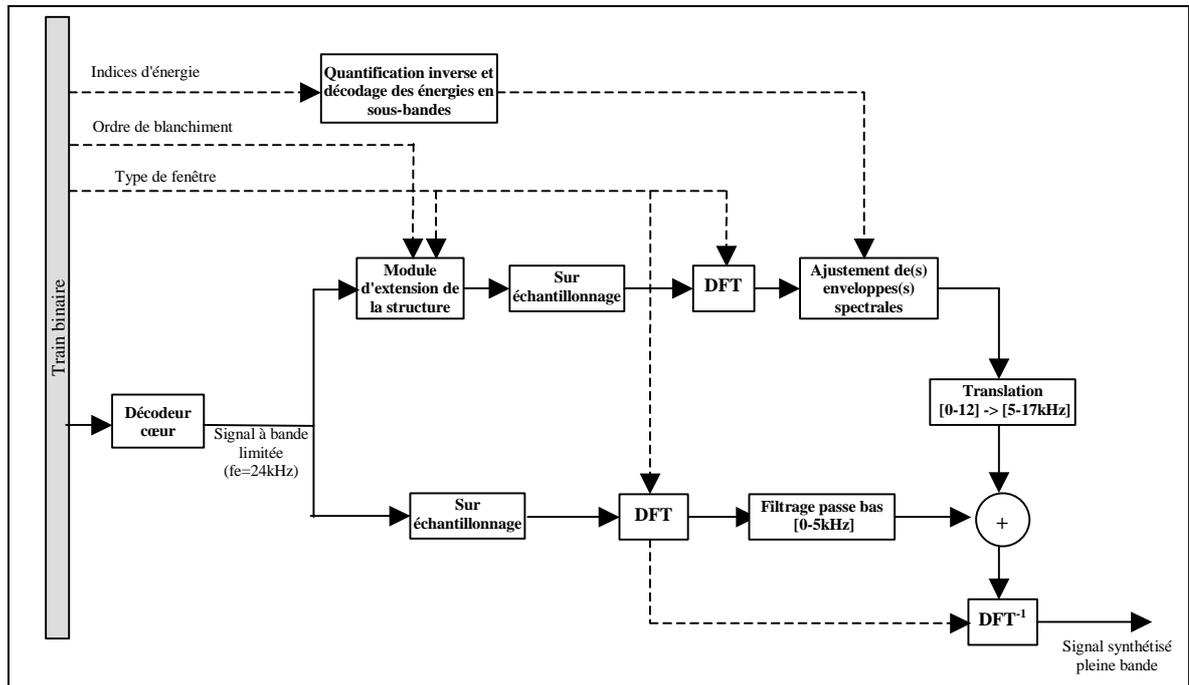


Figure 5.16 : Fonctionnement du décodeur par facteurs d'échelle en sous-bandes

Le signal cœur est décodé avant d'être injecté dans le module d'extension de la structure fine qui génère le signal (blanchi selon l'ordre de blanchiment transmis) de 12 kHz de bande correspondant aux hautes fréquences [5-17kHz]. Ce signal est ensuite sur-échantillonné par deux avant d'être passé dans le domaine fréquentiel. On utilise sur cet exemple une transformée temps/fréquence de type DFT. La justification du choix de cette transformée est précisée au paragraphe suivant.

Le spectre synthétisé est alors translaté en haute-fréquence, puis remis en forme par les facteurs d'énergies en sous-bandes.

Une fois son enveloppe correctement ajustée, le signal est sommé au signal cœur dans le domaine transformé. La DFT inverse synthétise finalement le signal spectralement enrichi.

5.2.5.3.6. Remarque concernant le choix de la transformée

Deux types de transformée temps/fréquence ont été implémentés et testés pour l'ajustement d'enveloppe : une transformée de type MDCT et une transformée de type DFT. Le détail de ces deux transformées est fourni en annexe A.

Le comportement de ces deux transformées est sensiblement le même sur les signaux bruités et/ou faiblement harmonique. En revanche, concernant l'ajustement d'enveloppe des signaux à fortes composantes tonales, la MDCT génère des artefacts que nous tâchons de mettre en évidence sur l'expérience décrite ci-dessous.

A partir d'un signal échantillonné à 32 kHz et constitué d'un sinus de fréquence 3333 Hz, on réalise :

- Une transformée de type MDCT sur 1024 points (2048 points temporels pondérés par une fenêtre en sinus). La MDCT génère 1024 coefficients réels dans le domaine fréquentiel.
- Une transformée de type DFT sur 2048 points (même 2048 points temporels pondérés par la même fenêtre en sinus). La DFT génère 1024 coefficients complexes dans le domaine fréquentiel.

Les Figure 5.17.a et Figure 5.17.b donnent respectivement la représentation du signal dans le domaine MDCT et dans le domaine DFT (première partie du spectre à symétrie hermitienne), les amplitudes des deux spectres étant respectivement égales à :

$$A_{Mdct}[i] = 10 * \log_{10}(MdctSg[i]^2) , 0 \leq i < 1024 \quad (5.3)$$

$$A_{Dct}[i] = 10 * \log_{10}(R[i]^2 + I[i]^2) , 0 \leq i < 1024 \quad (5.4)$$

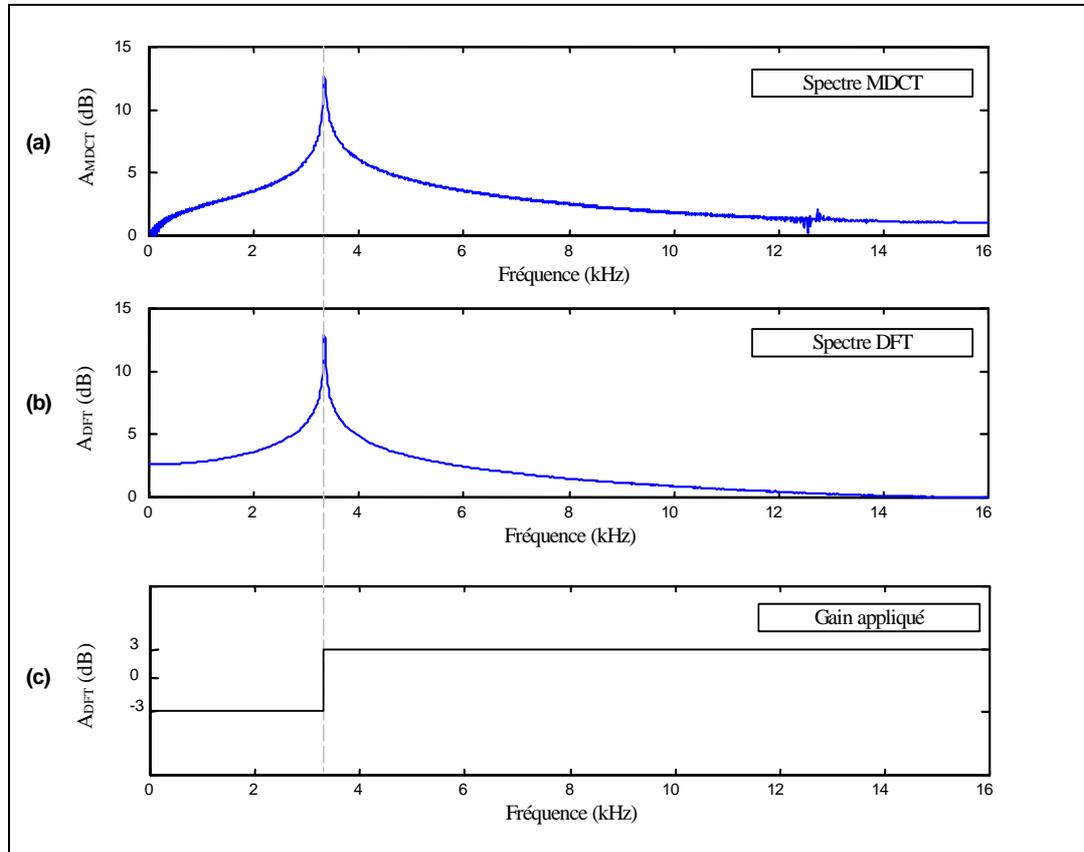


Figure 5.17 : Spectre MDCT et DFT d'un sinus à 3333 Hz

L'énergie des coefficients correspondant au sinus est principalement concentrée autour de sa fréquence fondamentale (3333 Hz).

On simule ensuite un ajustement d'énergie en deux sous-bandes de longueur respective [0-3343.75Hz] et [3343.75-16000Hz], la valeur intermédiaire de 3343.75 Hz correspondant à la raie numéro 214.

L'énergie de la première sous-bande est multipliée par 2, celle de la deuxième sous-bande est divisée par deux (gain illustré Figure 5.17.c). La Figure 5.18 donne le résultat après transformée inverse sur une période de 0.3s.

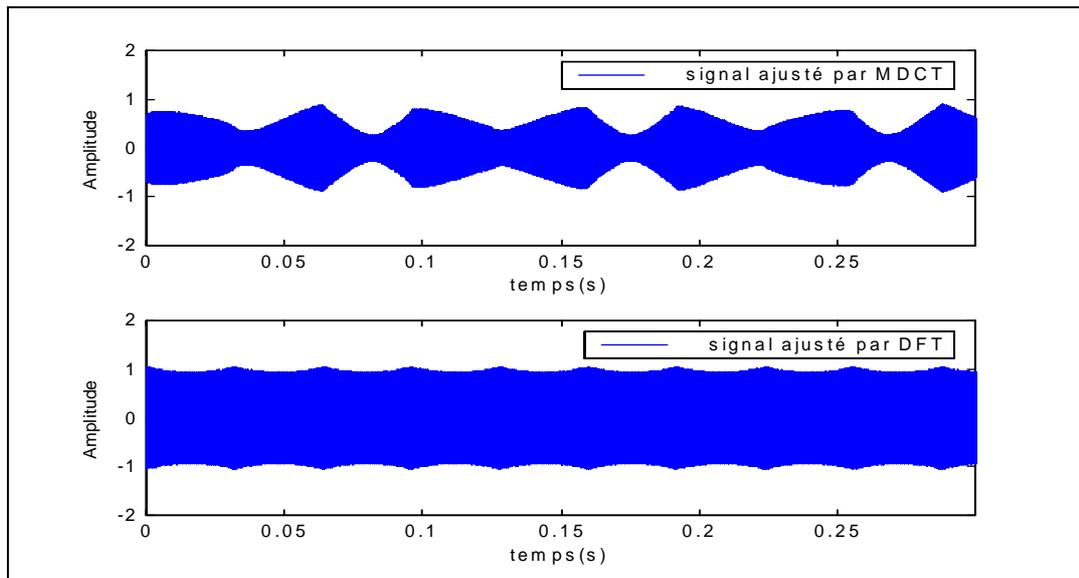


Figure 5.18 : Signaux résultants de l'ajustement d'énergie par MDCT et par DFT

On observe nettement des phénomènes de battement sur le signal ajusté dans le domaine MDCT. La MDCT étant une transformée à coefficients réels, les informations de phase et d'amplitude relatives au sinus sont combinées et difficiles à appréhender. L'ajustement de l'énergie d'une partie du spectre dans le domaine MDCT pose dès lors des problèmes temporels puisque le fait de modifier ces coefficients influe non-seulement sur le module, mais également sur la phase des différentes composantes. En réalisant l'ajustement d'énergie décrit sur l'exemple précédent, on génère ainsi plusieurs raies qui interagissent entre elles (phénomène de modulation d'amplitude). Ce phénomène est particulièrement perceptible sur les signaux harmoniques purs lors de l'ajustement fréquentiel d'une ou d'un ensemble de tonales.

En revanche, l'ajustement d'enveloppe sur ces signaux dans le domaine DFT ne génère pas ces artefacts puisque l'ajustement d'énergie se fait directement sur le module, sans altérer la phase.

5.2.5.4. Tests d'écoute

Afin de déterminer la meilleure des deux techniques de modélisation et d'enveloppe spectrale (par prédiction linéaire et par facteurs d'échelles), des tests d'écoutes ont été réalisés auprès de 5 experts audio sur les deux méthodes complètes d'enrichissement de spectre.

Les deux techniques testées utilisent le même module d'extension de la structure fine, et seule varie la méthode d'estimation et d'ajustement d'enveloppe. Notons que les tests ont été réalisés à partir de signaux originaux filtrés passe-bas à 5 kHz. Dans les deux cas, le débit moyen requis par les deux techniques est contraint à 2 kbit/s.

La méthode DCR [ITU 94] a été retenue pour ce test. Pour chaque séquence, l'auditeur compare deux séquences codées/décodées (A désignant la solution par facteurs d'échelle et B la solution par prédiction linéaire) par rapport à l'original en donnant sa préférence pour l'une ou l'autre des méthodes sur une échelle de -3 à 3 (note de -3 pour A est nettement moins bon que B à 3 pour A est nettement meilleur que B). Les valeurs moyennes assorties des intervalles de confiance à 95% sont présentées Figure 5.19 (test réalisé auprès de 5 auditeurs experts).

Séquence	Nom	Séquence	Nom
es02	German Male	si02	Castanets
es03	English Female	si03	Pitch Pipe
sc02	Orchestral Piece	sm01	Bag Pipes
si01	Harpsichord	sm03	Plucked Strings

Tableau 5.3 : Séquences testées

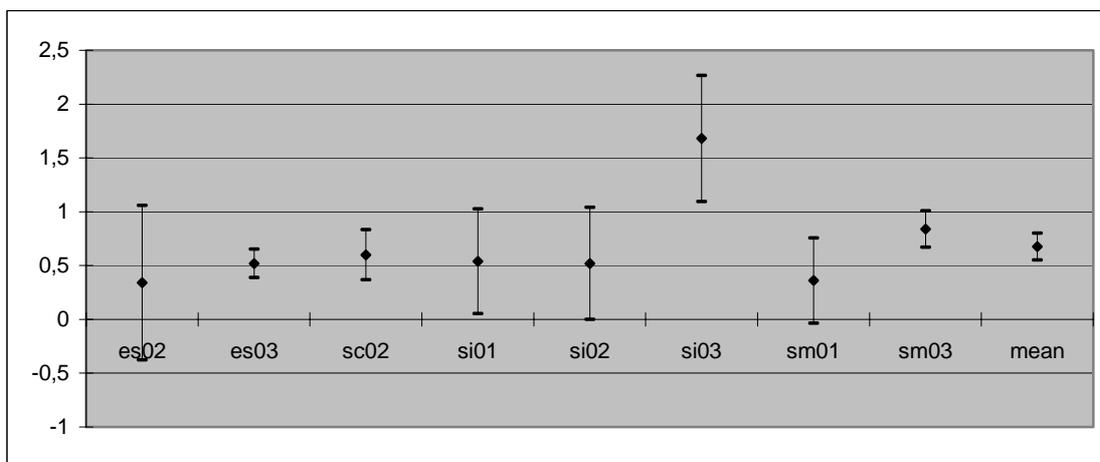


Figure 5.19 : Comparaison entre les deux techniques de modélisation d'enveloppe spectrale

Les tests montrent que :

- Pour deux séquences, les deux méthodes donnent des résultats statiquement identiques (les intervalles de confiance coupant l'axe des abscisses)
- Pour sept séquences, la méthode de modélisation par facteurs d'échelles est meilleure que la méthode par prédiction linéaire.

On peut également remarquer que pour la séquence pitch-pipe (une séquence très harmoniques) la méthode par facteurs d'échelles est bien meilleure que la méthode par prédiction linéaire.

Ces résultats montrent que, dans le contexte de notre étude, la modélisation d'enveloppe par facteurs d'échelle dans le domaine fréquentiel est plus adaptée à celle par prédiction linéaire, et notamment sur les signaux très harmoniques.

5.2.5.5. Conclusion

On a développé dans ce chapitre deux techniques complètes d'extension de bande des signaux musicaux.

La première technique repose sur une extension de la structure fine par des opérations de translations de spectre dans le domaine fréquentiel et sur une modélisation d'enveloppe par prédiction linéaire. La technique implémentée permet, pour un débit moyen de 2 kbit/s, d'étendre la bande passante des signaux musicaux à bande limitée (fréquence de coupure de l'ordre de 5 kHz). La modélisation d'enveloppe par prédiction linéaire, adaptée aux signaux de parole, génère des artefacts perceptivement gênant sur les signaux fortement harmoniques. On maîtrise notamment mal les variations du filtre d'enveloppe au cours du temps.

La seconde technique implémentée repose sur le même module d'extension de la structure fine décrit précédemment mais utilise une méthode de modélisation et d'ajustement d'enveloppe différente. Elle consiste à décrire l'énergie en sous-bandes du spectre haute-fréquence. L'estimation et l'ajustement d'enveloppe par facteurs d'échelles dans le domaine DFT s'avèrent être plus appropriés pour notre application. On contrôle en effet plus aisément la quantification des facteurs d'énergie, limitant ainsi les discontinuités temporelles et les phénomènes de battement au cours du temps.

Concernant la qualité de la technique d'enrichissement implémentée, nous présentons en conclusion de ce chapitre, aux paragraphes 5.4 et 5.5, les tests d'écoute réalisés dans le cadre de la normalisation dans MPEG-4 (technique associée au codeur cœur spécifique MPEG-4 AAC et ITU G-729) face à la technique SBR qui fait l'objet du paragraphe suivant.

5.3. Technique SBR (Spectral Band Replication)

Nous présentons dans ce paragraphe une technique d'extension de bande alternative au PAT, la technique SBR. Le système SBR, développée par CT (Coding Technologies) depuis 1998 consiste à élargir la bande passante d'un signal après décodage. Cette solution propriétaire a été normalisée dans le projet DRM (Digital Radio Mondiale) en 1999 ([DRM 01]) et est en cours de normalisation dans MPEG-4 ([ISO 02]).

La technique SBR est également utilisée dans le codeur MP3 Pro (technique d'enrichissement de spectre associé au codeur MPEG-2 couche 3). Notons que les détails de la techniques SBR, et notamment l'implémentation du décodeur, ne sont connus que depuis Janvier 2002.

5.3.1. Principe

La technique est basée sur un banc de filtres. Le principe est de transmettre l'enveloppe spectrale haute-fréquence sous forme de facteurs d'échelle en sous-bandes et d'étendre la structure fine du spectre par des opérations de transpositions en translatant les sous-bandes du signal cœur vers les hautes fréquences.

Notons qu'actuellement, seule la technique développée au décodeur est parfaitement connue (normalisation du décodeur dans MPEG-4 en janvier 2002). On ne connaît en revanche pas le fonctionnement du codeur. La description faite ci-après est de ce fait totalement indicative.

5.3.1.1. Codeur

Le rôle du codeur est d'estimer et de transmettre l'enveloppe spectrale, ainsi que le niveau de bruit et de blanchiment à appliquer à chacune des sous-bandes signal haute-fréquence.

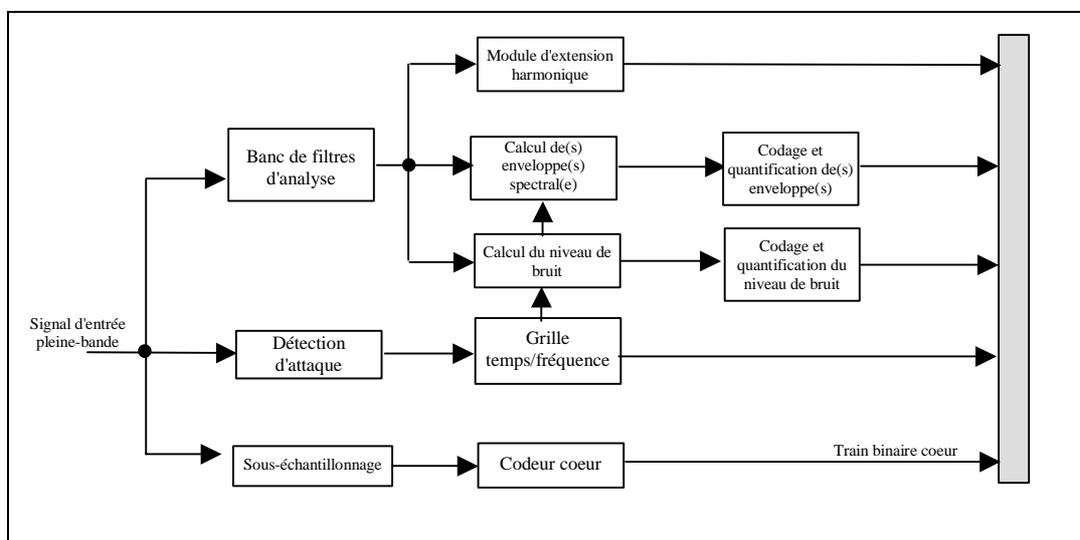


Figure 5.20 : Diagramme de fonctionnement du codeur SBR

Le signal d'entrée est segmenté en trames de 42.6 ms avant d'être injecté dans un banc de filtres à 64 sous-bandes complexes.

Une détection d'attaque détermine un découpage temps fréquences adapté au signal à coder. Ainsi pour un signal stationnaire, le codeur transmet une seule enveloppe spectrale et les niveaux de bruit et de blanchiment à appliquer dans les sous-bandes hautes. Pour un signal transitoire en revanche, la trame est divisée entre 2 et 5 sous-trames (affinement de la résolution temporelle) et le codeur transmet une enveloppe pour chacune de ces sous-trames.

Les enveloppes haute-fréquence et les niveaux de bruits sont calculés à la sortie du banc de filtres, groupés puis codés par un algorithme de Huffman [HUF 52] avant d'être transmis. Le niveau de blanchiment est également estimé et transmis.

Le module d'extension harmonique permet de transmettre (individuellement) des tonales haute-fréquence non synthétisables par la technique d'extension de spectre utilisée au décodeur. Le numéro de la sous-bande dans laquelle la tonale se situe est transmis. Une tonale est synthétisée au milieu d'une sous-bande (afin de modérer son coût de transmission). Cette technique de transmission reste approximative mais offre l'avantage de transmettre des tonales isolées pour un débit très faible.

5.3.1.2. Décodeur

Le signal basse-fréquence est décodé par le décodeur cœur puis injecté dans un banc de filtres d'analyse. Les sous-bandes basses sont éventuellement blanchies (en utilisant l'ordre de blanchiment déterminé et transmis par le codeur) avant d'être translatées en haute-fréquence.

Le signal haute-fréquence ainsi synthétisé est remis en forme par les facteurs de gain transmis décrivant le(s) enveloppe(s) spectrale(s). Le générateur de bruit associé corrige le niveau de bruit des hautes fréquences synthétisées. Le module d'extension des harmoniques injecte les éventuelles tonales dans les sous-bandes visées.

Les sous-bandes haute-fréquence ainsi remises en forme sont sommées au signal cœur décodé lors de la synthèse dans le banc de filtres de synthèse, générant ainsi le signal de sortie large bande.

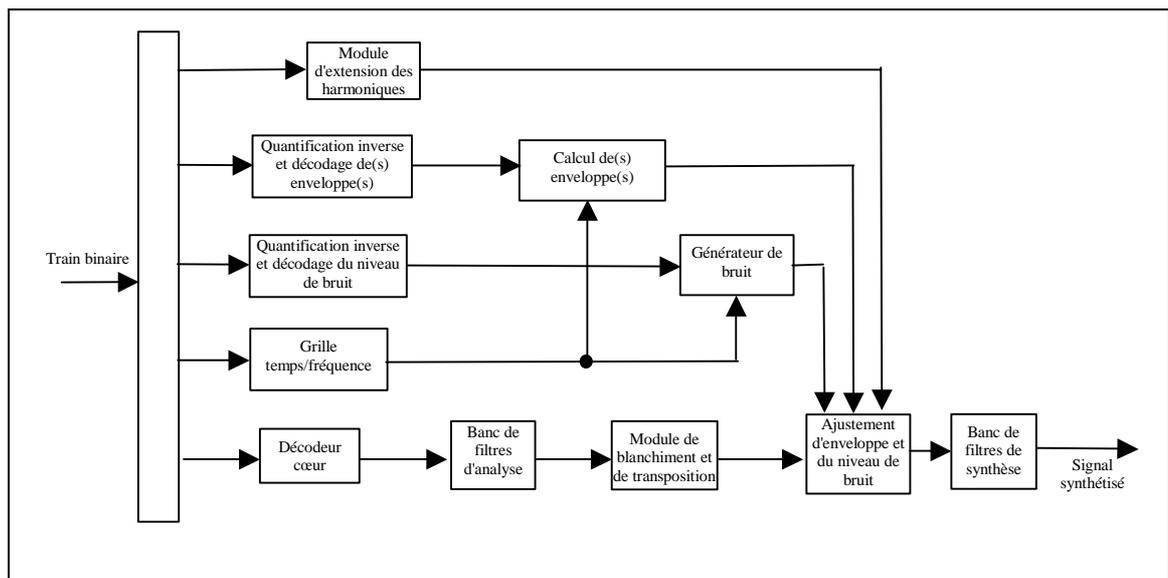


Figure 5.21 : Diagramme de fonctionnement du décodeur SBR

5.3.2. Conclusions

Basé sur un modèle de transposition de la structure fine et de transmission d'enveloppe par facteur d'échelle dans le domaine transformée, la technique SBR permet de synthétiser la bande haute-fréquence pour un débit variable compris entre 1 et 6 kbit/s et un débit moyen de l'ordre de 2.5 kbit/s. Un exemple sur un extrait de parole chantée est illustré Figure 5.22.

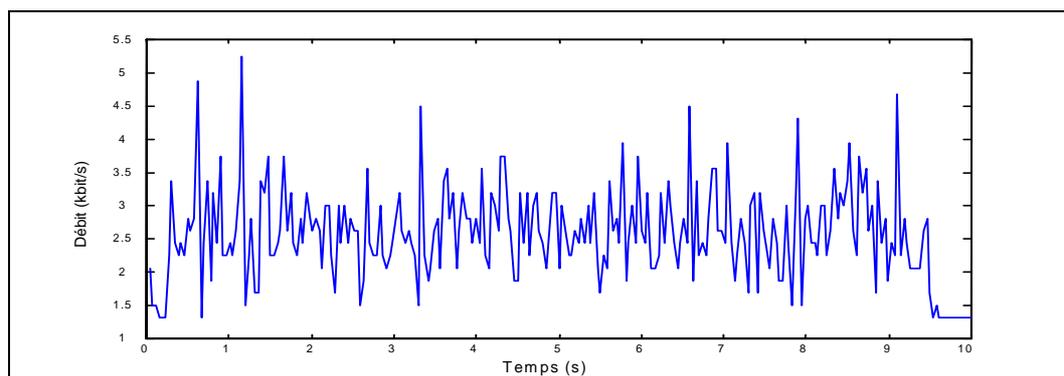


Figure 5.22 : Débit variable du SBR sur la séquence es01.wav

Les originalités du SBR par rapport au PAT concernent essentiellement le module d'extension des harmoniques et le générateur de bruit qui offrent la possibilité de synthétiser des tonales isolées et du bruit dans n'importe quelle partie du spectre haute-fréquence.

Les tests de qualité réalisés en décembre 2001 dans le cadre de la normalisation MPEG-4 sont présentés au paragraphe suivant.

5.4. Résultats des tests MPEG-4

Les deux techniques d'extension de bande PAT et SBR exposées dans ce chapitre ont été proposées en normalisation dans le standard MPEG-4 en juillet 2000. Les différentes phases de normalisation ont fait l'objet de nombreux tests d'écoutes comparatifs afin de montrer l'intérêt de ces nouvelles techniques d'enrichissement de bande dans le domaine du codage audio des signaux génériques.

Nous présentons dans ce paragraphe deux tests réalisés d'une part sur les signaux de parole et associant la technique PAT au codeur de parole ITU G-729, et d'autre part sur les signaux musicaux et associant les techniques PAT et SBR au codeur de musique MPEG-4 AAC.

5.4.1. Technique PAT associée au codeur de parole ITU G-729

Nous présentons ici les tests réalisés en janvier 2001 [ISO 01b] lors de la première étape de normalisation (étape du call for evidence, paragraphe 1.3) dans MPEG-4. La Figure 5.23 donne les résultats des tests comparatifs entre le codeur de parole MPEG-CELP en bande élargie à 24 kbit/s et le codeur de parole ITU G-729 associé à la technique PAT pour un débit total de 13,8 kbit/s (11,8 kbit/s pour le G-729 et 2 kbit/s pour le PAT). Notons que la nouvelle solution proposée correspond à 57.5% du débit de la référence.

Notons également que la technique PAT associée au G-729 fonctionne également à des débits de 8,4 kbit/s (6,4 kbit/s pour le G-729 et 2 kbit/s pour le PAT) et de 10 kbit/s (8 kbit/s + 2 kbit/s).

La méthode DCR [ITU 94] a été retenue pour ce test réalisé sur 17 séquences de parole. Les notes associées sont explicitées Tableau 5.4 (A désignant la solution MPEG CELP et B désignant la solution G-729+PAT).

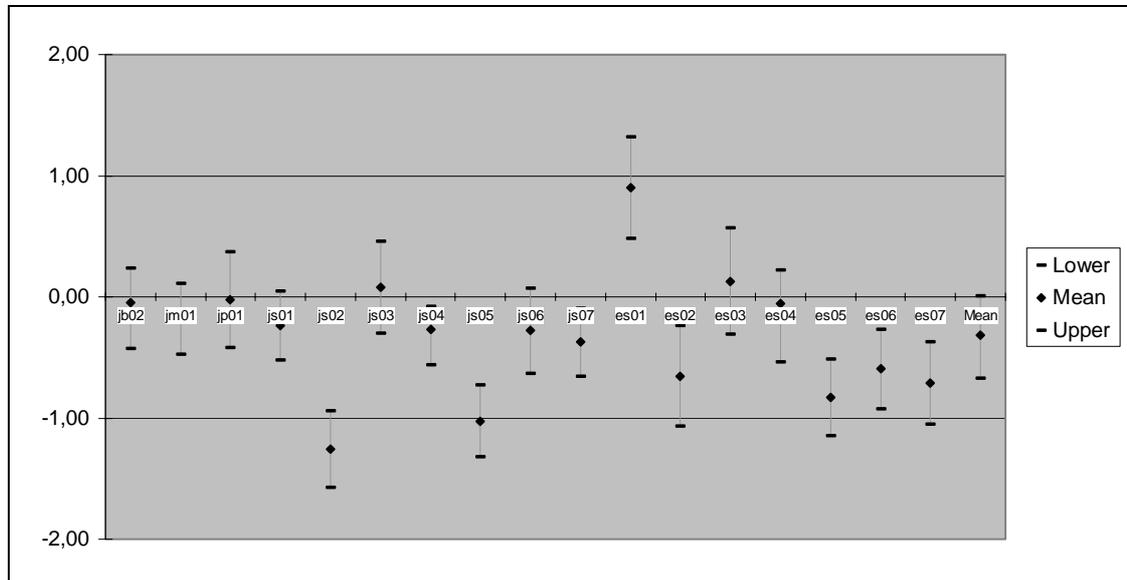


Figure 5.23 : Tests comparatifs entre le CELP (24 kbit/s) et le G-729+PAT (13,8 kbit/s)

Echelle comparative	Score
B est meilleur que A	+2
B est un peu meilleur que A	+1
B est équivalent à A	0
B est un peu moins bon que A	-1
B est largement moins bon que A	-2

Tableau 5.4 : Echelle de tests

On peut ainsi conclure à partir ces tests que :

- La qualité des deux codeurs est statistiquement la même pour 8 séquences
- La qualité est jugée légèrement meilleure à comparable pour la solution G729+PAT sur une séquence
- La qualité est jugée légèrement moins bonne à comparable pour le CELP sur 6 séquences
- Deux séquences sont considérées comme moins bonne avec le G729+PAT

5.4.2. Technique PAT adaptée au codeur de musique MPEG-4 AAC

Nous présentons ici les tests réalisés en décembre 2001 [ISO 01d] lors de la phase comparative entre les techniques PAT et SBR associées au codeur de musique MPEG-4 AAC.

Les systèmes testés sont présentés Tableau 5.5.

Système testé	Mono	Stéréo	Candidat
MPEG-4 AAC	24 kb/s	48 kb/s	FhG
MPEG-4 AAC	30kb/s	60 kb/s	FhG
AAC + PAT (22kb/s+2kb/s)	24 kb/s	48 kb/s	France Télécom
AAC + SBR	24 kb/s	48 kb/s	Coding Technologies

Tableau 5.5 : Systèmes testés

La méthode MUSHRA [ITU 00] a été retenue pour ce test réalisé sur les 12 séquences du Tableau 5.6.

Séquence	Nom	Séquence	Nom
es01	Susan Vega	si01	Harpsichord
es02	German Male	si02	Castanets
es03	English Female	si03	Pitch Pipe
sc01	Trumpet Solo and Orchestra	sm01	Bag Pipes
sc02	Orchestral Piece	sm02	Glockenspiel
sc03	Contemporary Pop music	sm03	Plucked Strings

Tableau 5.6 : Séquences testées

Les Figure 5.24 et Figure 5.25 présentent respectivement les résultats des tests mono et stéréo.

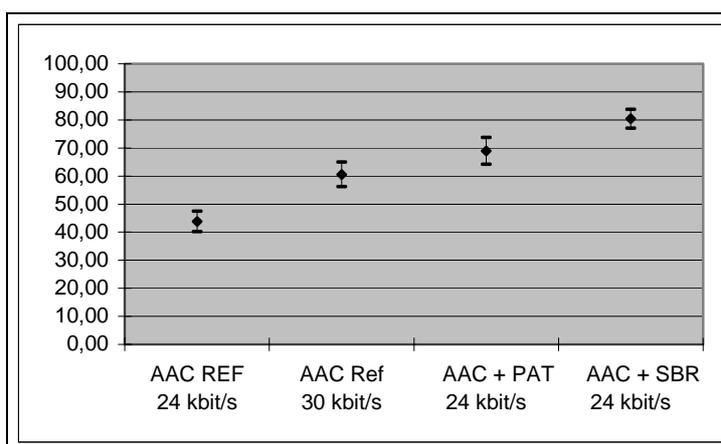


Figure 5.24 : Tests Mono

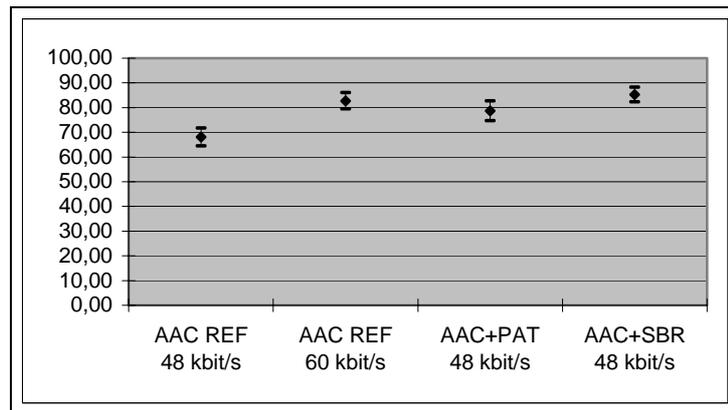


Figure 5.25 : Tests Stéréo

On peut ainsi conclure à partir de ces tests que :

- La configuration AAC+SBR est meilleure que la configuration AAC+PAT
- La qualité du codeur AAC+PAT est meilleur que l'AAC seul à 30 kbit/s.
- La qualité du codeur AAC+PAT à 48 kbit/s est meilleure que l'AAC seul à 48 kbit/s et statistiquement comparable à l'AAC seul à 60 kbit/s

5.4.3. Conclusion

Les tests présentés dans ce paragraphe montrent l'intérêt des techniques d'extension de bande dans le domaine de la compression des signaux audio bas-débit. Ces tests ont entraîné leur normalisation récente dans les deux projets mondiaux DRM et MPEG.

Associée au codeur G-729, la technique PAT permet le codage des signaux de parole en bande élargie à un débit total de 13,8 kbit/s, pour une qualité proche de celle du codeur CELP seul à un débit de 24 kbit/s. La solution proposée offre ainsi une réduction de débit de 57,5%.

Concernant les signaux musicaux, les deux solutions AAC+PAT et AAC+SBR testées offrent une qualité meilleure que l'AAC seul, qui notons-le, était jusqu'à cette date le meilleur codeur audio [ISO 98] et [EBU 00]. Pour une qualité équivalente, la réduction de débit offerte par l'utilisation des modules d'extension de bande est de l'ordre de 25%.

5.5. Conclusion du chapitre

Nous avons présenté dans ce chapitre deux solutions complètes d'enrichissement de spectre, le PAT et le SBR. Ces deux solutions concurrentes, associées au codeur MPEG-4 AAC, ont fait l'objet d'un processus de normalisation dans MPEG-4. Les deux techniques se sont avérées intéressantes pour le codage audio et pour la réduction de débit et constituent de ce fait une avancée certaine dans le domaine de la compression du son.

Elles sont toutes deux basées sur un mode de fonctionnement proche, à savoir :

- Une extension de la structure fine réalisée par translations de spectre successives de tout, ou d'une partie du signal à bande limitée. Cette solution offre ainsi la possibilité de synthétiser des signaux de bande passante de l'ordre de 16 kHz à partir de signaux de bande passante variable de l'ordre de 5 kHz.
- Une estimation, une transmission à bas-débit et un ajustement de l'enveloppe spectrale. La solution retenue dans les deux techniques repose sur une modélisation en sous-bande de l'enveloppe spectrale. Un codage entropique de Huffman permet de réduire les informations à transmettre.
- Un blanchiment spectral variable afin de contrôler le rapport tonal sur bruit des hautes fréquences synthétisées.

Notons que dans la technique SBR, un module additionnel permet en complément d'injecter du bruit dans les sous-bandes haute-fréquence et des tonales isolées.

Toutes les considérations développées dans ce chapitre nous ont permis de constater l'efficacité de ces nouvelles techniques d'enrichissement de spectre des signaux audionumériques. Susceptible de fonctionner avec différents types de codeur cœur, la technique PAT offre par exemple un débit moyen de 2 kbit/s pour une bonne qualité de restitution, et ce, quel que soit le type de signaux traités (parole et musique) et le type de codeur cœur utilisé.

Associée au codeur ITU G-729, la technique PAT offre ainsi un codage des signaux de parole pleine-bande pour un débit total de 13,8 kbit/s. La qualité d'une telle approche est proche de celle des sons codés par le MPEG CELP à 24 kbit/s et représente donc un gain en débit considérable (57,5%).

Concernant les signaux musicaux monophoniques, la solution AAC+PAT développée à 24 kbit/s offre une qualité comparable, voire légèrement meilleure que celle de l'AAC seul à 30 kbit/s, ce qui offre un gain en débit de 25%.

La qualité des signaux stéréophoniques codés en AAC+PAT à 48 kbit/s est équivalente de celle à un son codé à 60 kbit/s en AAC seul (25% du débit également).

La solution SBR a été retenue comme point de départ dans la norme MPEG-4 en décembre 2001. Proche de la technique PAT, le SBR constitue une solution plus aboutie qui offre notamment la possibilité, grâce à l'approche en sous-bandes, de transmettre des tonales isolées, de blanchir des sous-bandes particulières et d'injecter du bruit dans n'importe quelle sous-bande.

5.6. Bibliographie du chapitre 5

- [BOS 97] M. BOSI, et al.
ISO/IEC MPEG-2 Advanced Audio Coding
Journal Audio Engineering Society, pp. 789-813, Octobre 1997
- [DRM 01] Digital Radio Mondiale (DRM) : ETSI TS 101 980 V1.1.1
System Specification
Berlin, September 2001
- [EBU 00] G. STOLL & F. KOZAMERNIK
EBU listening tests on Internet audio codecs
EBU technical Review No. 283, Juin 2000
- [HUF 52] D. A. HUFFMAN
A method for the construction of minimum redundancy codes ”
Proceedings of the IRE, Vol. 40, Numéro 9, Septembre 1952.
- [ISO 01a] ISO/IEC JTC1/SC29/WG11 FDIS 14496
Information technology – Generic Coding of Audio Visual Objects, Part 3 (MPEG-4)
Edition 2001
- [ISO 01b] ISO/IEC JTC1/SC29/WG11 MPEG01/M6706
Results of the Core Experiment on France Telecom's Codec Proposal
P. PHILIPPE, Janvier 2001, Pise, Italie
- [ISO 01c] ISO/IEC JTC1/SC29/WG11 MPEG01/M6708
Results of the Core Experiment on France Telecom's codec Proposal
F. DONKERS, Janvier 2001, Pise, Italie
- [ISO 01d] ISO/IEC JTC1/SC29/WG11 MPEG01/N4378
Report of MPEG-4 Audio Bandwidth Extension RM0 Tests
Pattaya, Décembre 2001
- [ISO 02] ISO/IEC JTC1/SC29/WG11 MPEG02/N4611
WD Text for Backward Compatible Bandwidth Extension for General Audio Coding
March 2002, Jeju, Corée
- [ISO 92] ISO/IEC JTC1/SC29/WG11
Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to About
1.5Mbit/s, standard n°11172, alias "MPEG-1"
ISO-MPEG, Londres, Novembre 1992
- [ISO 98] ISO/IEC JTC1/SC29/WG11
Report on the MPEG-4 audio NADIB verification tests
ISO-MPEG, Juillet 1998
- [ITU 94] International Telecommunication Union (ITU)
Methods for the subjective assessment of small impairments in audio systems including
multichannel sound systems
Recommendation BS.1116, 1994
- [ITU 00] International Telecommunication Union (ITU)
Multi stimulus test with hidden reference and anchor (MUSHRA) - EBU method for subjective
listening tests of intermediate audio quality
Recommendation ITU [10-11Q/62], Janvier 2000
- [JOH 88] J.D. JOHNSTON
Transform coding of audio signals using perceptual noise criteria
IEEE Journal of selected Areas in Communication, Vol. 6, pp. 314-323, 1988

-
- [LAR 95] J. LAROCHE
Traitement des signaux audio-fréquences
Département Signal, Groupe acoustique, Télécom Paris, Février 1995
- [NAJ 00] H. NAJAFZADEH-AZGHANDI
Perceptual Coding of Narrowband Audio Signals
Thèse de l'université de McGill, Montréal, Avril 2000
- [RAB 77] L.R. RABINER
On the use of autocorrelation analysis for pitch detection
IEEE Transactions on acoustics, speech and signal processing
Vol. ASSP-25, pp 24-23, février 1977

CHAPITRE 6

CONCLUSIONS ET PERSPECTIVES

Les techniques d'élargissement de bande offrent de nouvelles perspectives dans le codage audio bas-débit. En témoigne leur normalisation récente dans deux projets mondiaux (DRM et MPEG) et leur utilisation croissante dans différents domaines d'application (codage audio bas-débit sur Internet, radiodiffusion...).

En codage de parole, les techniques d'enrichissement de spectre permettent à bas débit de passer de la bande téléphonique à la bande étendue, offrant ainsi une qualité d'écoute appréciable. Cette approche est intéressante pour les applications téléphoniques fixes et mobiles.

En codage de musique, les recherches menées depuis quelques années sur ce sujet ont donné naissance à différents codeurs : la technique SBR utilisée dans le MP3Pro, la technique Plus-V5 de chez VLSI, et la technique PAT qui a été développée pendant cette thèse.

Sujets abordés

L'étude menée dans la première partie de cette thèse nous a permis de constater que, sur la majorité des signaux musicaux, le contenu haute-fréquence fin est très proche perceptivement du contenu basse-fréquence. Nous avons également vu l'importance de la notion de l'enveloppe spectrale dans la perception des sons.

A partir de ces deux constatations, une technique efficace d'extension de bande des signaux musicaux a été implémentée, implémentation reposant sur un module d'extension de la structure fine du spectre et sur un module d'estimation, de transmission et d'ajustement de l'enveloppe spectrale en haute-fréquence.

Concernant le module d'extension de la structure fine du spectre, différentes méthodes connues dans la littérature ont été explorées au chapitre 4. L'étude s'est portée plus particulièrement sur deux techniques utilisées depuis une vingtaine d'années pour l'extension des signaux de parole [MAK 79] : les distorsions non-linéaires et les translations de spectre.

La technique basée sur les distorsions non-linéaire s'est avérée prometteuse en terme de complexité et de qualité de restitution des signaux audio composés d'une seule fréquence fondamentale (parole mono-locuteur notamment). Cet outil reste toutefois difficile à contrôler dans le domaine fréquentiel sur les signaux plus complexes constitués de plusieurs fréquences fondamentales (c'est-à-dire sur les mélanges de parole et signaux musicaux en général).

La technique basée sur les translations de spectre dans le domaine fréquentiel a été retenue pour sa grande souplesse de mise en œuvre et pour ses performances en terme de qualité de restitution. Les translations de spectre ont en effet l'avantage de préserver la structure spectrale fine de la plupart des signaux musicaux. Liée à un blanchiment spectral variable des hautes fréquences synthétisées, la technique synthétise des signaux de bande passante élevée de qualité proche de celle des signaux originaux.

Nous avons exposé au chapitre 3 les techniques d'estimation et de transmission d'enveloppe en se concentrant sur deux techniques particulières : Celle basée sur la prédiction linéaire qui consiste à transmettre l'enveloppe sous la forme d'un filtre linéaire et celle basée sur une modélisation d'enveloppe par facteurs d'échelle dans le domaine fréquentiel qui consiste à transmettre l'enveloppe par des énergies en sous-bandes dans le domaine transformé.

Cette dernière technique offrant plus de degrés de liberté que la précédente, elle s'est avérée plus appropriée aux techniques d'enrichissement de spectre à bas débit (technique retenue dans le PAT et dans le SBR)

Techniques PAT développées

Afin de répondre aux différentes échéances de normalisations dans DRM et MPEG-4, un outil complet d'enrichissement de spectre a été développé en C++. Susceptible de fonctionner sur des codeurs cœurs utilisant différentes approches, deux techniques complètes ont fait l'objet d'un développement approfondi :

- L'une basée sur le codeur de parole de type ITU G-729. La solution G-729+PAT étend les signaux de parole de la bande téléphonique à la bande élargie pour des débits variant entre 8,4 et 14 kbit/s (G-729+PAT).
- L'autre basée sur le codeur de musique MPEG-4 AAC. La solution AAC+PAT permet par exemple de coder des signaux pleine-bande à des débits de 24 kbit/s monophonique et 48 kbit/s stéréophonique.

La technique PAT développée offre une grande souplesse d'utilisation. Elle fonctionne en effet avec des signaux d'entrée de fréquences d'échantillonnage diverse (8, 22.05, et 24 kHz) et peut délivrer en sortie des signaux de fréquences variées (16, 32, 44.1 ou 48 kHz). Elle s'adapte également quelle que soit la bande passante du signal à bande-limitée. Associée au codeur MPEG-4 AAC, elle tolère par exemple des signaux codés de bande passante comprise entre 3.5 et 9 kHz.

La solution développée requiert enfin un débit faible, de l'ordre de 2 kbit/s. Les techniques d'extension de bande offrent ainsi un gain en débit considérable. Rappelons que pour le codeur AAC, 28 kbit/s sont nécessaires à la synthèse des hautes fréquences comprises entre 6 et 15 kHz (Tableau 1.1).

Qualité

Les tests formels réalisés dans le cadre de la normalisation dans MPEG-4, et présentés au paragraphe 5.4 ont démontré l'efficacité des techniques d'enrichissement de spectre puisque concernant la technique PAT développée :

- La qualité des signaux monophoniques codés en G-729+PAT à 14 kbit/s est proche de celle codée par le CELP MPEG-4 seul en bande élargie à 24 kbit/s (57,5 % du débit de la référence).
- La qualité des signaux monophoniques codés en AAC+PAT à 24 kbit/s est équivalente à un son codé à 30 kbit/s en AAC seul, et que la qualité des signaux stéréophoniques codés en AAC+PAT à 48 kbit/s est statistiquement équivalente à celle des sons codés à 60 kbit/s en AAC seul (25 % du débit de la référence).

L'intérêt d'une telle technique d'extension de bande pour la réduction de débit est ainsi démontré.

Perspectives

Les principales perspectives concernent essentiellement le module d'extension de la structure fine. La technique développée est actuellement limitée en ce sens qu'elle ne permet pas de synthétiser des tonales isolées en haute-fréquence, et pose donc des problèmes pour l'extension de bande des signaux inharmoniques. Une approche paramétrique bas débit serait à explorer afin de générer toute la panoplie des signaux numériques (transmission de la position des tonales isolées). Une seconde perspective concernerait le module de blanchiment du spectre basse-fréquence. Afin de contrôler plus finement le rapport tonales à bruit des hautes fréquences synthétisées, il serait en effet intéressant de pouvoir blanchir indépendamment les différentes parties du spectre haute-fréquence (solutions adoptées dans le SBR).

Un second domaine d'étude concernerait l'adaptation de la technique à un codeur cœur particulier. Nous nous sommes attachés tout au long de cette thèse à développer une technique générique d'enrichissement de spectre susceptible de fonctionner avec différents codeur cœur (AAC et G-729 notamment). Il serait maintenant intéressant d'étudier les adaptations requises pour adapter la technique à ces codeurs particuliers. Cette adaptation offrirait des avantages certains en terme de qualité de restitution.

Valorisation

La technique issue de ces trois années de thèse a fait l'objet de deux participations à des projets de normalisation mondiaux.

La première version du codeur AAC+PAT a fait l'objet d'une proposition de normalisation dans DRM en novembre 2000 (contribution [DRM 00a] et [DRM 00b])

La seconde version du PAT a fait l'objet d'une proposition de normalisation en novembre 2001, dans MPEG-4 (contribution [ISO 01]).

Les codeurs/décodeurs G-729+PAT et AAC+PAT sont implémentés dans des projets internes à France Télécom R&D et TDF.

Les travaux de cette thèse ont également fait l'objet d'une présentation à Coresa en novembre 2001 [COL 01] et un article sur l'extension de bande est en cours de rédaction [COL 03]

Notons enfin que le travail réalisé dans le domaine de l'extension de bande au cours de cette thèse nous a mené à déposer quatre brevets.

6.1. Bibliographie du chapitre

- [ISO 01] ISO/IEC JTC1/SC29/WG11 MPEG01/M7325
Evaluation of the performance of the FT R&D Proposal to the Call for Proposal
J.B. RAULT & P. COLLEN, July 2001, Sydney
- [DRM 00a] Digital Radio Mondiale DRM TC SC 056
Subjective tests on the Perceptual Audio Transposition system (PAT) : an alternative to the SBR system.
P. COLLEN, P. PHILIPPE, et J.C. RAULT , Erlangen, Mars 2000
- [DRM 00b] Digital Radio Mondiale DRM TC SC 071
Complexity of PAT decoder, France Télécom R&D
P. COLLEN, P. PHILIPPE, et J.C. RAULT , Décembre 2000
- [MAK 79] J. MAKHOUL and M. BEROUTI
High frequency regeneration in speech coding systems
IEEE, ICASSP, pp 428-431, Avril 1979
- [COL 01] P. COLLEN et J.B RAULT
Le PAT : un outil d'élargissement de la bande audio pour le codage bas débit
CORESA (Compression et représentation des signaux audiovisuels), Dijon, novembre 2001
- [COL 03] P. COLLEN et P.PHILIPPE
Bandwidth extension tools applied to high quality audio coding
A soumettre à l'IEEE, Transactions on Speech and Audio Processing

CHAPITRE 7

ANNEXES

TRANSFORMEES TEMPS/FREQUENCE.....	144
CONVERSION LPC/LSP.....	146
CONVERSION LSP-LPC	148
SYNTAXE DU TRAIN BINAIRE PAT	149

ANNEXE A

TRANSFORMEES TEMPS/FREQUENCE

Transformée de Fourier discrète (DFT)

La transformée temps/fréquence la plus connue est la transformée de Fourier discrète. Elle s'écrit :

$$X(k) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x(n) e^{-j2\pi kn/N}, \quad 0 \leq k < N \quad (\text{A.1})$$

où N représente le nombre d'échantillons de la transformée.

La DFT étant une transformée complexe, il est aisé d'en extraire les informations de modules et de phase.

Cette transformée est utile pour l'estimation de la puissance spectrale des signaux. Les discontinuités temporelles aux limites des bords pouvant entraîner une concentration moindre dans le domaine transformé, le signal est le plus souvent pondéré par une fenêtre aux bords adoucis (par exemple une fenêtre de Hanning).

Le chevauchement des fenêtres imposé par ce fenêtrage, induit un surcoût de dans le domaine transformé, car l'échantillonnage n'est alors plus critique (le nombre de points fréquentiels est plus important que le nombre de point en temps).

Transformée en cosinus discrète modifiée (MDCT)

L'échantillonnage critique étant important pour les applications de codage audio, la MDCT a été introduite en 1986 [PRI 86]. Elle allie à la fois échantillonnage critique et recouvrement entre trames consécutives. La reconstruction parfaite est assurée, et la concentration d'énergie en fréquence y est meilleure qu'avec la DFT.

La MDCT, également appelée banc de filtres à annulation de repliement dans le domaine temporel (TDAC Time Domain Aliasing Cancellation), est une transformation temps-fréquence avec recouvrement de 50%. A chaque bloc, le calcul de la transformée se calcule sur N points, la progression n'étant que de M=N/2 échantillons.

Les composantes fréquentielles de la MDCT, notées X(k) d'un bloc x(n) s'écrivent ([PRI 86]) :

$$X(k) = \sqrt{\frac{2}{M}} \sum_{n=0}^{N-1} h(n) \cos\left(\left(n + \frac{M+1}{2}\right)\left(k + \frac{1}{2}\right)\frac{\pi}{M}\right) x(n), \quad 0 \leq k < M \quad (\text{A.2})$$

Où h(n) représente la fenêtre d'analyse et M le nombre de sous-bandes. La longueur de chaque trame d'analyse vaut N=2M.

Et l'équation de synthèse est donnée par :

$$y(n) = g(n) \sqrt{\frac{2}{M}} \sum_{k=0}^{M-1} X(k) \cos\left(\left(n + \frac{M+1}{2}\right)\left(k + \frac{1}{2}\right)\frac{\pi}{M}\right) \quad (\text{A.3})$$

Où $g(n)$ représente la fenêtre de synthèse.

Contrairement à la DFT, la MDCT est une transformée réelle et il devient plus délicat de retrouver les informations de phase et de module qui se mélangent dans les coefficients MDCT.

Référence

- [PRI 86] J.P. PRINCEN & A.B. BRADLEY
Analysis/Synthesis filter bank design based on Time Domain Aliasing Cancellation
IEEE Trans. on ASSP, Vol. 34, pp 1153-1161, Octobre 1986

ANNEXE B

CONVERSION LPC/LSP

A partir du polynôme d'ordre P ([UIT 96])

$$A(z) = 1 - \sum_{k=1}^P a_k \cdot z^{-k} \quad (1)$$

on construit les deux polynômes réciproques d'ordre P+1

$$\begin{aligned} F_1'(z) &= A(z) + z^{-(P+1)} A(z^{-1}) \\ F_2'(z) &= A(z) - z^{-(P+1)} A(z^{-1}) \end{aligned} \quad (2)$$

Ces deux polynômes ont les propriétés suivantes :

- $F_1'(z)$ est symétrique et $F_2'(z)$ est antisymétrique (conjugués)
- Si toutes les racines de $A(z)$ sont à l'intérieur du cercle unité alors les racines de $F_1'(z)$ et de $F_2'(z)$ sont sur le cercle unité
- Les racines de $F_1'(z)$ et de $F_2'(z)$ apparaissent de façon alternée sur le cercle unité.

Les polynômes $F_1'(z)$ et $F_2'(z)$ possèdent respectivement les racines $z = -1$ ($\omega = \pi$) et $z = +1$ ($\omega = 0$). On élimine ces deux racines en définissant les deux nouveaux polynômes suivants :

$$\begin{aligned} F_1(z) &= \frac{F_1'(z)}{1 + z^{-1}} \\ F_2(z) &= \frac{F_2'(z)}{1 - z^{-1}} \end{aligned} \quad (3)$$

Chacun de ces polynômes possède P/2 racines conjuguées sur le cercle unité et on peut les écrire comme suit :

$$\begin{aligned} F_1(z) &= \prod_{i=1,3,\dots,(P-1)} (1 - 2 \cdot q_i \cdot z^{-1} + z^{-2}) \\ F_2(z) &= \prod_{i=2,4,\dots,P} (1 - 2 \cdot q_i \cdot z^{-1} + z^{-2}) \end{aligned} \quad (4)$$

avec $q_i = \cos(\omega_i)$.

Les coefficients LSF correspondent à la position angulaire ω_i , entre 0 et π , des racines de P(z) et de Q(z) et vérifient :

$$0 = \omega_0 < \omega_1 < \dots < \omega_P < \omega_{P+1} = \pi \quad (5)$$

Etant donné que les deux polynômes $F_1(z)$ et $F_2(z)$ sont symétriques, il suffit de calculer les P/2 premiers coefficients de chaque polynôme, au moyen des relations de récurrence suivantes :

$$f_1(i+1) = a_{i+1} + a_{10-i} - f_1(i) \quad i = 0, \dots, P/2-1 \quad (6)$$

$$f_2(i+1) = a_{i+1} + a_{10-i} - f_2(i) \quad i = 0, \dots, P/2-1$$

où $f_1(0) = f_2(0) = 1$. Les coefficients LSF sont trouvés par évaluations des polynômes $F_1(z)$ et $F_2(z)$ à 60 points équidistants entre 0 et π puis en vérifiant les changements de signes. Chacun de ces derniers correspond à l'existence d'une racine et l'intervalle du changement de signe est alors divisé par 4 afin d'affiner la recherche de la racine.

On se sert des polynômes de Chebycheff pour évaluer les deux polynômes $F_1(z)$ et $F_2(z)$. Grâce à cette méthode, les racines sont calculées directement dans le domaine des cosinus. Le polynôme $F_1(z)$ ou $F_2(z)$, évalué à la valeur $z = e^{j\omega}$, peut s'écrire comme suit :

$$F(\omega) = 2.e^{-j\omega P/2}.C(x) \quad , \text{ où } C(x) = T_{P/2}(x) + f(1)T_{P/2-1}(x) + \dots + f(P/2-1)T_1(x) + f(P/2)/2 \quad (7)$$

où le terme $T_m(x) = \cos(m\omega)$ est le polynôme de Chebycheff du nième ordre et où les $f(i)$, $i=1, \dots, P/2$, sont les coefficients du polynôme $F_1(z)$ ou $F_2(z)$, calculés par l'équation (10). Le polynôme $C(x)$ est évalué à une certaine valeur de $x = \cos(\omega)$ au moyen de la relation de récurrence suivante :

pour $k =$ de $P/2-1$ à 1

$$B_k = 2xb_{k+1} - b_{k+2} + f(P/2-k) \quad (8)$$

fin

$$C(x) = xb_1 - b_2 + f(P/2)/2$$

Avec les valeurs initiales $b_{P/2} = 1$ et $b_{P/2+1} = 0$.

Référence

- [UIT 96] Codage de la parole à 8kbit/s par prédiction linéaire avec excitation par séquences codées à structure algébrique conjuguée – G.729
Union internationale des télécommunications, Recommandation UIT-T G729, Mars 1996

CONVERSION LSP-LPC

Une fois quantifiés, les coefficients LSF sont reconvertis en coefficients LPC, a_i .

Pour ce faire, on détermine les coefficients des polynômes $F_1(z)$ et $F_2(z)$ par expansion des équations (9), sur la base des coefficients LSF quantifiés. A partir des coefficients q_i , on calcule les coefficients $f_1(i)$, $i=1, \dots, P/2$, par la relation de récurrence suivante :

Pour $i=1$ à $P/2$

$$f_1(i) = -2 \cdot q_{i-1} \cdot f_1(i-1) + 2f_1(i-2) \quad (9)$$

Pour $j =$ de $i-1$ à 1

$$f_1(j) = f_1(j) - 2q_{2i-1} \cdot f_1(j-1) + f_1(j-2)$$

avec comme valeurs initiales $f_1(0) = 1$ et $f_1(-1) = 0$. Les coefficients $f_2(i)$ sont calculés de la même manière, avec remplacement des q_{2i-1} par q_{2i} .

Une fois les coefficients $f_1(i)$ et $f_2(i)$ calculés, on multiplie les polynômes $F_1(z)$ et $F_2(z)$ par $(1+z^{-1})$ et $(1-z^{-1})$ respectivement, pour obtenir les polynômes $F_1'(z)$ et $F_2'(z)$, qui donnent les coefficients suivants :

$$\begin{aligned} f_1'(i) &= f_1(i) + f_1(i-1) & , \quad i = 1, \dots, P/2 \\ f_2'(i) &= f_2(i) - f_2(i-1) & , \quad i = 1, \dots, P/2 \end{aligned} \quad (10)$$

Finalement, on retrouve les coefficients LPC à partir des coefficients $f_1'(i)$ et $f_2'(i)$ comme suit :

$$a_i = \begin{cases} 0,5 \cdot f_1'(i) + 0,5 \cdot f_2'(i) & i = 1, \dots, P/2 \\ 0,5 \cdot f_1'(11-i) + 0,5 \cdot f_2'(11-i) & i = 1, \dots, P/2 \end{cases} \quad (11)$$

Les polynômes $F_1'(z)$ et $F_2'(z)$ étant symétrique et antisymétrique, cette équation est directement issue de la relation $A(z) = (F_1'(z) + F_2'(z))/2$.

ANNEXE C

SYNTAXE DU TRAIN BINAIRE PAT

Syntax	No. of bits	Mnemonic
<code>fill_element()</code>		
{		
<code>cnt = bs_count</code>	4	uimsbf
if (cnt == 15)		
<code>cnt += bs_esc_count - 1</code>	8	uimsbf
<code>bs_extension_type</code>	4	
<code>bs_fill_nibble</code>	4	
if (bs_extension_type == 2){		
<code>PAT_bitstream()</code>		
<code>bs_fill_bits</code>	8*(cnt-1)-tot	
}		
else		
for (i= 0; i<cnt-1; i++)		
<code>bs_fill_byte[i]</code>	8	uimsbf
}		

Table 7.1 : Syntax of fill_element() with PAT extension

Syntax	No. of bits	Mnemonic
<code>PAT_bitstream()</code>		
{		
<code>core_coder_bandwidth_present</code>	1	uimsbf
if (core_coder_bandwidth_present){		
<code>temp</code>	4	uimsbf
<code>core_coder_bandwidth=2*temp+18</code>		
}		
<code>cf_index_present</code>	1	
if(cf_index_present){		
<code>temp</code>	4	uimsbf
<code>Cf8Index=2*temp+18</code>		
}		
<code>PAT_mode</code>	2	uimsbf
for(sub_frame=0;sub_frame<n_subframe;sub_frame++)		
<code>PAT_subframe()</code>		
}		

Table 7.2 : Syntax of PAT_bitstream()

Syntax	No. of bits	Mnemonic
PAT_subframe() {		
window_type	2	uimsbf
quantization_type	1	uimsbf
lpc_order	2	uimsbf
if(window_type==SHORT) {		
if(quantization_type==0) {		
for(i=0;i<n_nrj_t;i++)		
nrj_ind_t[i]	6	uimsbf
for(i=0;i<n_nrj_f;i++)		
nrj_ind_f[i]	6	uimsbf
}		
else {		
nrj_ind_t[0]	6	uimsbf
for(i=1;i<n_nrj_t;i++)		
nrj_ind_t[i]=nrj_ind_t[i-1]+huff_dec(t_huff_short, bs_codeword);	1..10	bslbf
nrj_ind_f[0]	6	uimsbf
for(i=1;i<n_nrj_f;i++)		
nrj_ind_f[i]=nrj_ind_f[i-1]+huff_dec(f_huff_short, bs_codeword);	2..10	bslbf
}		
}		
else {		
if(quantization_type==0) {		
for(i=0;i<n_nrj_f;i++)		
nrj_ind_f[i]	6	uimsbf
}		
else {		
nrj_ind_f[0]= huff_dec(f_huff_long_0, bs_codeword)	4..14	bslbf
for(i=1;i<n_nrj_f;i++)		
nrj_ind_f[i]=nrj_ind_f[i-1]+huff_dec(f_huff_long, bs_codeword);	1..17	bslbf
}		
}		
}		

Table 7.3 : Syntax of PAT_subframe()