Codage à débit variable de la parole en bande élargie

Giuseppina Biundo Lotito

THÈSE SOUMISE À LA FACULTÉ DES SCIENCES
DE L'UNIVERSITÉ DE NEUCHÂTEL POUR L'OBTENTION
DU GRADE DE DOCTEUR ÈS SCIENCES

Copyright © 2003 Giuseppina Biundo Lotito

IMPRIMATUR POUR LA THESE

Codage à débit variable de la parole en bande élargie

de Mme Giuseppina Biundo Lotito

UNIVERSITE DE NEUCHATEL

FACULTE DES SCIENCES

La Faculté des sciences de l'Université de Neuchâtel, sur le rapport des membres du jury

MM. F. Pellandini (directeur de thèse),
P.-A. Farine, M. Ansorge et
M. Unser (EPF Lausanne)

autorise l'impression de la présente thèse.

Neuchâtel, le 12 juin 2003

Le doyen:

François Zwahlen

Résumé

Avec la diffusion des technologies multimédia et l'introduction des prochains services des 3ème et 4ème générations de la téléphonie mobile, le codage de la parole en bande élargie est devenu un domaine de recherche prédominant. L'intérêt de traiter un signal de parole en *bande élargie* est d'obtenir un signal de parole reconstitué plus net, plus naturel et plus intelligible que dans le cas de la *bande étroite*.

Ce rapport de thèse présente l'étude, la conception et l'implantation d'un codeur propriétaire de parole en bande élargie, fonctionnant à différents débits, utilisable aussi bien pour la téléphonie mobile que pour la transmission via le protocole Internet, ou encore pour le stockage de la parole. L'objectif est d'obtenir une bonne qualité du signal reconstruit et de viser un portage futur sur une plate-forme miniaturisée. Ainsi, différentes contraintes telles que la complexité d'implantation (matériel requis), la complexité d'exécution (nombre d'opérations de calculs à effectuer) et la consommation (puissance électrique requise) sont considérées, tout autant que le débit nécessaire à la transmission de l'information.

L'algorithme réalisé se base sur le codeur de parole en bande étroite G.729, encodant le signal par prédiction linéaire à excitation par séquences codées (CELP), dont le principe a été modifié pour la bande élargie. Cet algorithme a été développé en code ANSI C, avec une arithmétique en virgule flottante. Son implantation en virgule fixe n'a pas été réalisée dans le cadre de cette étude.

Ce rapport de thèse décrit l'algorithme CS-ACELP (Conjugate-Structure Algebraic-Code-Excited Linear-Prediction) de la famille d'algorithmes CELP. Il présente l'état de l'art dans le domaine du codage de la parole en bande élargie et motive le choix du codeur propriétaire développé. Il décrit plusieurs contributions scientifiques élaborées dans le cadre de ce travail, principalement liées à une forte réduction des bruits typiquement contenus dans le signal de parole en bande élargie, lorsqu'il est encodé par un algorithme de type CELP. Toutes les informations pour l'implantation du

codeur développé sont données. Les tests réalisés pour qualifier le codeur propriétaire et les résultats obtenus sont présentés. Finalement, sur la base du travail effectué, de futurs travaux sont suggérés.

Ce rapport comprend trois innovations ayant donné lieu à la déposition de trois brevets. Ces innovations sont l'introduction d'un pré-filtre pour le dictionnaire adaptatif, le contrôle de gain de l'excitation innovatrice (à court-terme) et l'introduction de deux filtres de pondération formantique différents pour l'extraction de l'excitation adaptative et de l'excitation innovatrice. Elles permettent une nette amélioration de la qualité du signal reconstruit. Ces innovations ont été validées et ont donné lieu à une implémentation particulière avec optimisation des paramètres.

Abstract

With the current widespread diffusion of multimedia technologies and the announced introduction of new services within the 3rd and 4th generations of mobile telephony systems, wideband speech coding became a prevailing research field. The interest of processing a *wideband* speech signal is to be able to produce a decoded signal that is cleaner, more natural, and more intelligible than that obtained by using a corresponding *narrowband* signal.

This thesis presents the study, the design, and the implementation of a proprietary wideband speech coder. This coder, which operates at various bit rates, is suitable for real-time speech communication either over wireless networks or over packet-switched networks such as the Internet. The coder is also suitable for speech storage applications.

The objective for the final design of the codec was to produce a good reconstructed speech signal with a low-complexity system that could be embedded in a miniature speech communication system. To this end, various design constraints were dealt with, including the implementation complexity (reduced hardware resources), the computational complexity (reduced number of arithmetic operations), the required electric power (minimum consumption), along with the constraint of producing a minimum bit-rate to convey the information.

The implemented algorithm is based on the narrowband speech coder G.729, which encodes the signal by using the Code-Excited Linear-Prediction (CELP) method. The core principle of the latter was modified to handle wideband speech. This modified algorithm was developed in ANSI C code, with floating point arithmetic. The fixed point implementation was however not covered in this work.

This thesis report describes the CS-ACELP (Conjugate-Structure Algebraic-Code-Excited Linear-Prediction) algorithm belonging to the CELP-type algorithms. The state-of-the-art in wideband speech coding is reported and the motivation which led to the choice of the developed proprietary coder are equally discussed. The report discusses in detail several scientific

contributions that were elaborated in the frame of the thesis, which principally contribute to reducing the noise components that are typically affecting wideband speech when encoded using CELP-type encoders. All the information for the implementation of the developed coder is provided, and the different tests carried out to assess the quality of the proprietary coder are presented along with the corresponding results. Finally, referring to the work done, several propositions are made regarding potential future research extension.

Three innovations were developed in the thesis, which led to three filed patents. These innovations consist in the insertion of a prefilter for the adaptive dictionary, in the control of the innovating (short-term) excitation gain, and in the incorporation of two different formantic weighting filters to extract the adaptive and innovating excitations. The quality of the reconstructed speech signal is noticeably improved using these original concepts. These innovations were validated, and they were implemented using a specific structure with optimized parameters.

A mes très chers parents.

Remerciements

Le travail de thèse présenté ici a été réalisé grâce à la contribution de plusieurs personnes : je leur en suis gré.

Je tiens d'abord à remercier mon professeur de thèse, le professeur Fausto Pellandini, pour m'avoir donné l'opportunité de travailler dans son groupe de recherche et pour la confiance qu'il m'a accordée, ainsi que pour avoir supervisé l'écriture de cette thèse. J'aimerais également remercier le professeur Michael Unser, le professeur Pierre-André Farine, et le docteur Michael Ansorge pour avoir accepté d'être co-examinateurs de thèse.

Michael Ansorge a directement participé à la recherche décrite dans ce rapport. Il a contribué à l'organisation de mon travail, en apportant des idées intéressantes et en examinant, révisant et améliorant nos publications. Je tiens à mentionner Sara Grassi, Benito Carnero, Alain Dufaux, et Giuseppe Zamuner qui ont partiellement travaillé avec moi, et avec lesquels j'ai partagé d'enrichissantes discussions. Je mentionne également tous mes collègues qui ont toujours été disponibles et m'ont apporté une aide précieuse notamment dans le domaine informatique.

Je remercie particulièrement notre secrétaire Claudine Faehndrich et notre ingénieur système Laurent Jeanrenaud, ainsi que Heinz Burri et Catherine Lehnherr.

Mes parents m'ont donné une aide précieuse au cours de mon travail de recherche et au long de la rédaction de ce rapport. Ils m'ont continuellement encouragée et soutenue moralement. Je leur en suis infiniment reconnaissante et je leur dédie cet ouvrage.

Finalement je tiens à remercier mon mari, toute ma famille et mes amis (en particulier Sonia et Fred) pour leur soutien.

Table des matières

Résumé ..... **v**

Abstract ..... **vii**

Remerciements ..... **x**

Table des matières ..... **xi**

Chapitre 1 Introduction **1**

1.1 Motivations 1
1.2 But de la recherche 3
1.3 Organisation du rapport de thèse 4
1.4 Contributions principales 4
1.5 Publications 5
1.6 Références 5

Chapitre 2 Bases théoriques **7**

2.1 Introduction 7
2.2 Signal de la parole 8
2.2.1 Signal de parole sous forme analogique et numérique 8
2.2.2 Processus de production de la parole 9
2.2.3 Signal de parole dans le domaine des fréquences 11
2.3 Modélisation du signal de parole 14
2.3.1 Signal de parole modélisé par prédiction linéaire 15
2.3.2 Principe général de l'analyse par synthèse et codeur CELP 18
2.3.3 Extraction des paramètres de prédiction linéaire à court-terme 21
2.3.4 Extraction des paramètres de prédiction linéaire à long-terme 24
2.3.5 Extraction des paramètres de l'excitation innovatrice 29
2.4 Appareil auditif et perception auditive du signal de parole 31

2.4.1 Appareil auditif 31
2.4.2 Perception auditive 33
2.5 Limites du codage LPAS 38
2.5.1 Pré-accentuation 39
2.6 Critères pour l'évaluation des performances d'un codeur de parole 40
2.7 Complexité des algorithmes 41
2.8 Résumé du chapitre et conclusions 41
2.9 Références 41

Chapitre 3 Codeur CS-ACELP 43

3.1 Introduction 43
3.2 Description générale du codeur 44
3.2.1 Principe de l'encodeur CS-ACELP 44
3.2.2 Principe du décodeur CS-ACELP 48
3.3 Analyse LPC 49
3.3.1 Expansion de la largeur de bande et fenêtre décalée 50
3.3.2 Généralités pour la quantification des coefficients LPC 51
3.3.3 Mesure des performances de la quantification des LPC 51
3.3.4 Représentation alternative des coefficients LPC 53
3.3.5 Interpolation des coefficients LPC 54
3.3.6 Paires de lignes spectrales (LSP) 55
3.3.7 Quantification des paramètres LSP 60
3.3.8 Extraction des paramètres LSP 62
3.3.9 Transformation des LSP en LPC 63
3.4 Extraction de l'excitation adaptative par prédiction à long-terme 64
3.4.1 Extraction de l'excitation adaptative dans la boucle : implantation du dictionnaire adaptatif 66
3.4.2 Quantification des paramètres LTP 69
3.5 Extraction de l'excitation innovatrice 70
3.5.1 Dictionnaire algébrique 74
3.5.2 Prédiction du gain β 77
3.6 Quantification vectorielle des gains 79
3.7 Mise à jour des mémoires des filtres de synthèse et de pondération 79
3.8 Post-filtrage 79
3.9 Délai de traitement d'un codeur 80
3.10 Résumé du chapitre et conclusions 81
3.11 Références 82

Chapitre 4 Etat de l'art 85

4.1 Introduction 85
4.2 Contraintes et performances requises pour le codeur WB–AMR de l'ETSI 86

4.3 Etat de l'art jusqu'en 1999 87
4.3.1 Résumé 87
4.3.2 Discussion 90
4.4 Evolution de l'état de l'art de 2000 à 2002 91
4.4.1 Nouveau standard ETSI : le WB-AMR 91
4.4.2 Résumé 94
4.4.3 Discussion 95
4.5 Références 96

Chapitre 5 Choix du codeur développé et principales contributions 97

5.1 Introduction 97
5.2 Etapes de développement 98
5.3 Contraintes 99
5.4 Choix du codeur à développer 101
5.4.1 Rapport débit - qualité 101
5.4.2 Complexité de calcul et délai de traitement 102
5.4.3 Choix du type de codeur 103
5.4.4 Choix de la structure de base 103
5.4.5 Choix de la durée des trames et des sous-trames 104
5.5 Mesures des performances et bases de données 106
5.5.1 Mesures des performances 106
5.5.2 Bases de données : entraînements et tests 106
5.6 Conception d'un dictionnaire de quantification pour les coefficients de prédiction linéaire 107
5.6.1 Méthodes de quantification des LSP et méthodes de tests 108
5.6.2 Exemples de quantificateurs spectraux 110
5.6.3 Expérimentation de différents schémas de quantification spectrale 112
5.6.4 Choix d'un schéma de quantification pour le codeur propriétaire 126
5.6.5 Conclusions de la section 129
5.7 Contributions à la réduction du bruit de quantification, d'un codeur de type ACELP pour la parole en bande élargie 130
5.7.1 Problèmes de reconstruction du signal, recherche de solutions et brevets 131
5.7.2 Conclusions de la section 143
5.8 Méthodes d'extraction de l'excitation adaptative 143
5.8.1 Analyse du problème et solutions 144
5.8.2 Conclusions de la section 147
5.9 Modes d'extraction de l'excitation innovatrice 148
5.10 Conclusions 148
5.11 Références 150

Chapitre 6 Description fonctionnelle du codeur P-MRWB-ACELP 153

6.1 Introduction 153
6.2 Encodeur 154
6.2.1 Pré-traitement 154
6.2.2 Analyse par prédiction linéaire et quantification 154
6.2.3 Analyse du délai tonal en boucle ouverte 157
6.2.4 Calcul des réponses impulsionnelles et des réponses à zéro 158
6.2.5 Résidu et cible pour l'extraction de l'excitation adaptative 159
6.2.6 Recherche du délai tonal 159
6.2.7 Calcul de l'excitation adaptative 161
6.2.8 Calcul du gain adaptatif 162
6.2.9 Procédure de recherche dans le dictionnaire innovateur 163
6.2.10 Quantification des gains 166
6.2.11 Mise à jour du dictionnaire adaptatif et des mémoires 167
6.2.12 Encodage des différents paramètres 168
6.3 Décodeur 169
6.3.1 Décodage des paramètres 169
6.3.2 Traitement des paramètres LSP 169
6.3.3 Reconstruction de l'excitation adaptative 171
6.3.4 Reconstruction de l'excitation innovatrice et décodage de son gain . 171
6.3.5 Construction de l'excitation totale et mise à jour du dictionnaire adaptatif 171
6.3.6 Calcul de la parole reconstruite 172
6.3.7 Post-traitement 172
6.4 Complexité algorithmique totale 173
6.5 Conclusions 175
6.6 Référence 175

Chapitre 7 Tests et résultats 177

7.1 Introduction 177
7.2 Tests 177
7.3 Résultats 179
7.3.1 Limites des tests auditifs 181
7.4 Futurs travaux et conclusions 184
7.5 Références 184

Chapitre 8 Conclusions générales 187

8.1 Publications et brevets 189
8.1.1 Publications 189
8.1.2 Brevets 190

Annexe A Extraction des paramètres LSP 191

A.1 Méthode de Kabal 191
A.2 Références 193

Annexe B Transformation de LSP en LPC 195

B.1 Méthode d'expansion directe 195
B.2 Méthode de Kabal 198
B.3 Références 200

Annexe C Catégories et sous-catégories des codeurs de l'état de l'art 201

C.1 Catégories et sous-catégories des codeurs de l'état de l'art jusqu'en 1999 201
C.1.1 Codeurs de type CELP 201
C.1.2 Codeurs de type SB-CELP 203
C.1.3 Codeur par transformée 204
C.1.4 Codeurs mixtes 205
C.1.5 Autre codeur 206
C.2 Catégories et sous-catégories des codeurs de l'état de l'art de 2000 à 2002 206
C.2.1 Codeur de type CELP 206
C.2.2 Codeur de type SB-CELP 206
C.2.3 Codeurs en sous-bandes, où la bande de fréquences inférieure est encodée à l'aide d'un algorithme pour la bande étroite 207
C.2.4 Codeurs en sous-bandes, où la bande de fréquences inférieure est encodée à l'aide d'un algorithme pour la bande étroite et où la bande de fréquences supérieure est encodée par transformée 208
C.2.5 Codeur de type MELP 209
C.2.6 Codeur par transformée 209
C.3 Références 210

Annexe D Résultats des tests auditifs 215

Annexe E Glossaire 219

Chapitre 1
Introduction

Ce rapport de thèse décrit l'élaboration d'un algorithme complet de codage et décodage du signal de parole en bande élargie. Cet algorithme permet une compression à un taux variable. Il a été développé dans l'optique d'obtenir la meilleure qualité de signal reconstruit tout en considérant une complexité de calcul et un débit aussi bas que possible.

1.1 Motivations

L'importance particulière du traitement de la parole dans un cadre très général, s'explique par la position privilégiée de la parole comme vecteur d'information dans notre société. Grâce au développement de la téléphonie mobile, il est désormais possible de communiquer depuis pratiquement n'importe quel endroit.

Les systèmes de codage de la parole, utilisés par la deuxième génération de téléphonie mobile (2G, en Europe GSM phase 2) [1-1], traitent des signaux de parole en bande étroite, appelés couramment signaux de parole dans la bande téléphonique. On qualifie de parole en bande étroite, les signaux de parole filtrés en temps continu dans la bande [300 – 3400 Hz], puis échantillonnés à la fréquence de 8 kHz. Le codage de parole en bande étroite permet d'obtenir un signal de parole reconstruit de qualité suffisante pour la téléphonie. Toutefois, celui-ci n'a pas toujours une consonance naturelle, et il apparaît parfois comme métallique, voire synthétique. Quelquefois, il est difficile d'en distinguer deux consonnes fricatives telles que le «s» ou le «f».

On parle d'un signal de parole en bande élargie, si le signal est filtré dans la bande [50-7000 Hz], puis échantillonné à la fréquence de 16 kHz. L'intérêt de traiter un signal de parole en bande élargie est d'obtenir un signal de parole reconstitué plus net, plus naturel, plus intelligible et plus « proche » que dans

le cas de la bande étroite. Le terme « intelligible » définit la capacité de comprendre le contenu du message transmis, alors que le terme « proche » se réfère à la sensation de présence du locuteur. L'extension de la bande fréquentielle dans les hautes fréquences, soit de 3400 à 7000 Hz, permet d'obtenir un signal de parole plus net et intelligible, et de mieux différencier les consonnes fricatives. De plus, l'extension dans les basses fréquences en abaissant la fréquence de coupure inférieure, soit de 300 à 50 Hz, permet d'obtenir un signal plus naturel et donnant une sensation de proximité. Un codeur de parole en bande élargie doit par conséquent : d'une part encoder avec précision les composantes en basse fréquence, importantes du point de vue perceptuel; d'autre part retenir suffisamment d'information en haute fréquence, afin de préserver la richesse et la fidélité du signal de parole original.

La diffusion des technologies multimédia ainsi que la proéminence de la vidéoconférence sont en croissance constante. Avec la naissance de ces nouvelles technologies apparaît l'importance d'utiliser des systèmes de codage de la parole permettant d'obtenir un son non seulement intelligible, mais également naturel. Ainsi, ces dernières années, le codage de la parole en bande élargie est devenu un domaine de recherche prédominant et d'importantes activités de standardisation ont été menées aussi bien par l'ETSI (European Telecommunications Standards Institute) que par l'ITU-T (International Telecommunications Union – Telecommunications Standards Sector). Ces activités ont été effectuées en vue d'intégrer le codage de la parole en bande élargie dans les services de la troisième génération de téléphonie mobile.

L'extension de la bande de fréquences implique une extension du débit de transmission nécessaire à l'encodage du signal de parole. Les principaux standards de téléphonie en bande étroite fonctionnent à un débit variant de 4.75 à 13 kbits/s pour le codage de la parole (codage de source). Ces débits correspondent à 13.0, 12.2 et 5.6 kbits/s pour les GSM FR (Full-Rate), EFR (Enhanced Full-Rate) et respectivement HR (Half-Rate) [1-2], 8 kbits/s pour le G.729 [1-2] et finalement 4.75 à 12.2 kbits/s pour les différents modes du NB-AMR (Narrow-Band Adaptive Multi-Rate) [1-3]. Les services de la deuxième génération (2G) de téléphonie mobile ne permettent pas de concevoir un débit de transmission de données supérieur à 14.4 kbits/s et un débit total supérieur à 22.8 kbits/s (GSM-FR). Le débit de transmission total comprend non seulement le codage de la source (parole), mais également le codage de canal.

Lors de la transmission du signal de parole par téléphonie mobile ou par le protocole Internet, la qualité du canal de transmission est variable, car fonction de divers paramètres tels que la géographie, le climat et

l'infrastructure mise à disposition par l'opérateur. Il est donc intéressant d'utiliser un débit adaptatif variable pour le codage de source, ce qui permet de protéger l'information par le codage de canal de façon adaptative, en fonction de la qualité du canal de transmission. Les systèmes à débits adaptatifs multiples, AMR (adaptative multi-rate), allouent les bits disponibles soit au codage de source soit au codage de canal, de façon adaptative.

Avec l'introduction des services de la génération 2.5 (GPRS General Packet Radio Service [1-4]) et de la 3ème génération (3G), des débits totaux nettement supérieurs à 22.8 kbits/s sont réalisables. On a par exemple pour le GPRS, 171.2 kbits/s au maximum pour la transmission des données [1-5]. Ces nouvelles générations permettent d'envisager la standardisation d'un codeur de parole en bande élargie pour la téléphonie mobile. Ainsi, un nouveau standard pour le codage de la parole en bande élargie, le WB-AMR (Wide-Band Adaptive Multi-Rate), a été choisi par l'ETSI en 2001. Ce nouveau standard fonctionne à des débits multiples, variant de 6.6 à 23.85 kbits/s pour le codage de source [1-6].

C'est dans le contexte de l'importance croissante du codage de la parole en bande élargie que s'inscrit l'étude et l'implantation menées ici. On cherche à obtenir un codeur propriétaire, de parole en bande élargie, qui fonctionne à différents débits et qui puisse être utilisé aussi bien pour la téléphonie mobile que pour la transmission via le protocole Internet ou encore pour le stockage de la parole.

1.2 But de la recherche

Dans le cadre de la recherche présentée ici, seul le codage (source) du signal de la parole en bande élargie est considéré : le codage de canal n'est pas traité. L'objectif principal de cette recherche est d'étudier et d'implanter un algorithme d'encodage et de décodage du signal à débit adaptatif, en virgule flottante sur logiciel informatique. Le but étant d'élaborer un algorithme permettant d'obtenir une bonne qualité du signal reconstruit, tout en visant un codeur à faible complexité, sans pour autant négliger le débit nécessaire à la transmission de l'information.

Le but final de la recherche est l'implantation de l'algorithme sur une plate-forme de traitement numérique du signal (DSP) permettant son utilisation en temps réel. La transcription en virgule fixe de l'algorithme développé ainsi que son implantation sur DSP ne sont toutefois pas traités dans le cadre de ce rapport. Ces objectifs ont été définis dans le cadre d'un projet de recherche [1-7].

L'élaboration complète d'un codeur de parole entièrement nouveau requiert un travail de très longue durée, généralement confié à une équipe et non à une ou deux personnes. L'élaboration d'un tel codeur dépasse le travail d'une seule thèse. Ici, l'algorithme à implanter se base sur un codeur existant, suivant les grandes lignes d'une famille d'algorithmes fréquemment utilisés dans le codage de la parole en bande étroite : la prédiction linéaire à excitation par séquences codées : CELP (Code-Excited Linear-Prediction). Les codeurs CELP font partie des codeurs à prédiction linéaire utilisant l'analyse par synthèse, LPAS (Linear Predictive Analysis-by-Synthesis).

1.3 Organisation du rapport de thèse

Ce rapport de thèse est articulé comme suit. Le Chapitre 2 introduit les concepts nécessaires à la compréhension de ce rapport. Il donne des définitions de base. Il décrit le signal de parole en présentant son mécanisme de production et la modélisation de celui-ci. Il introduit en outre le traitement du signal de la parole et montre l'importance de la perception auditive pour obtenir un bon codage.

Le Chapitre 3 décrit l'algorithme CS-ACELP (Conjugate-Structure Algebraic-Code-Excited Linear-Prediction) de la famille d'algorithmes CELP. Il porte l'accent sur une implantation algébrique de l'excitation d'un codeur CELP. Les concepts théoriques qu'il présente, permettent de décrire le codeur propriétaire développé dans le cadre de ce travail de thèse. Le Chapitre 4 décrit l'état de l'art dans le domaine du codage de la parole en bande élargie. Il est scindé en deux parties. La première concerne la littérature publiée jusqu'au début de ce travail de thèse. La seconde présente l'évolution de l'état de l'art de l'année 2000 à l'année 2002. Le Chapitre 5 décrit le choix du codeur propriétaire développé, et présente les principales contributions scientifiques de ce travail de thèse. Le Chapitre 6 fournit toutes les informations pour la réalisation ou l'implantation du codeur développé. Finalement, le Chapitre 7 décrit les tests réalisés pour qualifier le codeur propriétaire, présente les résultats obtenus et propose de futurs travaux.

Les conclusions générales sont données au Chapitre 8.

1.4 Contributions principales

Les principales contributions et innovations décrites dans cette thèse de doctorat sont :

(1) L'implantation d'un quantificateur vectoriel séparé sur deux étages, pour encoder les paramètres spectraux (LSP) [1-8], [1-9].

(2) L'introduction d'un pré-filtre pour le dictionnaire adaptatif, permettant de réduire le bruit en haute fréquence [1-10].

(3) Le contrôle de gain de l'excitation innovatrice (à court-terme) en fonction de l'importance de l'excitation adaptative (à long-terme) [1-11].

(4) L'introduction de deux filtres de pondération formantique différents, mais liés, permettant de réduire le bruit d'encodage en basse fréquence [1-12], [1-13].

Les innovations (2), (3) et (4) ont fait l'objet de trois brevets Européens [1-10], [1-11] et [1-12]. Ces brevets ont été déposés en juillet 2002.

1.5 Publications

Outre les trois brevets cités en Section 1.4, une partie du travail décrit dans ce rapport a déjà fait l'objet de publications scientifiques. Le papier présenté au *First COST 276 Workshop on Information and Knowledge Management for Integrated Media Communication*, en novembre 2001, décrit une première méthodologie pour établir la quantification vectorielle des coefficients de prédiction linéaire (LPC) [1-8]. Une seconde méthodologie pour concevoir cette même quantification a été présentée au *Third COST 276 Workshop on Information and Knowledge Management for Integrated Media Communication,* en octobre 2002 [1-9]. Finalement, la contribution (4), citée en Section 1.4, a été présentée au *Fourth COST 276 Workshop on Information and Knowledge Management for Integrated Media Communication,* en avril 2003 [1-13].

De plus, deux rapports internes, couvrant partiellement le Chapitre 5 et le Chapitre 6, ont été rédigés.

1.6 Références

[1-1] V. Garg et J. Wilkes, "An overview of wireless communications systems", Chapter 1, dans *Principles & applications of GSM*, pp. 1-15, Theodore S. Rappaport Series Editor, Prentice Hall PTR, 1999.

[1-2] L. Hanzo, F. Somerville et J. Woodard, "Standard forward-adaptive CELP codecs", Chapter 7, dans *Voice compression and communications*, pp. 207-278, IEEE Series on Digital & Mobile Communication, John Wiley & Sons, Inc., Publication, NY, USA, 2001.

[1-3] 3GPP TS 26.090 V3.1.0 (1999–12) document, dans ftp://ftp.3gpp.org/Specs/2000-09/R1999/26_series/ (14 Nov. 2001).

[1-4] J. Hämäläinene, "General packet radio service", Chapter 3, dans *GSM evolution towards 3rd generation systems*, pp. 65-80, édité par Z. Zvonar, P. Jung et K. Kammerlander, Kluwer Academic Publishers, 1999.

[1-5] http://www.mobilegprs.com/gprs.asp?link=1, (11 Jan. 2003).

[1-6] 3GPP TS 26.190 V5.0.0 (2001–03) document, dans ftp://ftp.3gpp.org/Specs/2001-09/Rel-5/26_series/ (14 Nov. 2001).

[1-7] "New Methods for Joint Wide-Band Speech and Channel Coding in Wireless Applications (JOWICOD)", projet de recherche CTI (Commission pour la technologie et l'innovation) 4238.1.

[1-8] G. Biundo, S. Grassi, M. Ansorge et F. Pellandini, "Spectral quantization for wideband speech coding", dans *Proc. of 1st COST 276 Workshop on information and knowledge management for integrated media communication* (CD-ROM), Leganés (Madrid), Espagne, Nov. 2001.

[1-9] G. Biundo, S. Grassi, M. Ansorge, F. Pellandini et P.-A. Farine, "Design techniques for spectral quantization in wideband speech coding", dans *Proc. of 3rd COST 276 Workshop on information and knowledge management for integrated media communication* (CD-ROM), Budapest, Hongrie, Oct. 2002.

[1-10] G. Biundo, M. Ansorge et B. Carnero, "Codeur de parole", brevet déposé EP 02 015 918.2.

[1-11] G. Biundo, M. Ansorge et B. Carnero, "Codeur de parole à gain réduit", brevet déposé EP 02 015 920.8.

[1-12] G. Biundo, M. Ansorge et B. Carnero, "Codeur de parole avec 2 filtres formantiques", brevet déposé EP 02 015 919.0.

[1-13] G. Biundo, M. Ansorge, F. Pellandini et P.-A. Farine, "Perceptual weighting for ACELP wideband speech coder", dans *Proc. of 4th COST 276 Workshop on information and knowledge management for integrated media communication*, pp. 105-110, Bordeaux, France, Mars-Avril 2003.

Chapitre 2
Bases théoriques

2.1 Introduction

Ce chapitre rappelle les principales caractéristiques du signal de parole. Il couvre sa production, sa modélisation, son traitement, mais également sa perception et ses principales applications.

Pour communiquer l'information, le locuteur produit un signal de parole sous la forme d'une onde de pression qui se déplace de la tête du locuteur à l'oreille de l'auditeur. Le signal de parole possède la particularité d'être à la fois produit et perçu instantanément par le cerveau du locuteur. La perception de la parole par le locuteur lui-même influence sa production. Le traitement de la parole est une science tenant compte du rôle important que joue le cerveau humain à la fois dans la production et dans la compréhension du signal de parole.

Le but suivi dans ce chapitre est l'introduction des concepts et des définitions utilisés dans le cadre du traitement du signal de parole pour la téléphonie mobile. En particulier, l'accent est mis sur le traitement du signal de parole en bande élargie (50 à 7000 Hz) puisque celui-ci fait l'objet de ce rapport de thèse. Les concepts théoriques plus spécifiques sont introduits au Chapitre 3.

Ce chapitre est articulé comme suit. La Section 2.2 décrit le signal de parole sous sa forme analogique, numérique et fréquentielle. Son processus de production et de perception y sont introduits. La Section 2.3 décrit la modélisation du signal de parole et en particulier sa modélisation basée sur l'analyse par synthèse. La Section 2.4 couvre la perception auditive du signal de parole par l'oreille humaine. Les limites du codage reposant sur l'analyse par synthèse sont introduites à la Section 2.5. Finalement, les critères de

mesure de la qualité d'un codeur de parole ainsi que la méthode d'évaluation de la complexité des algorithmes sont décrits dans les Sections 2.6 et 2.7.

2.2 Signal de la parole

Le signal de parole est caractérisé par ses niveaux linguistiques : acoustique, phonétique, phonologique, morphologique, lexical, syntaxique, sémantique et pragmatique [2-1]. Il se distingue des autres sons par des caractéristiques acoustiques qui ont leur origine dans ses mécanismes de production.

Cette section est divisée en 3 parties. La Sous-section 2.2.1 décrit les caractéristiques du signal de parole ainsi que sa transformation dans le domaine numérique. La Sous-section 2.2.2 introduit le processus de production de la parole. Finalement, la Sous-section 2.2.3 présente une analyse du signal de parole dans le domaine des fréquences.

2.2.1 Signal de parole sous forme analogique et numérique [2-1], [2-2]

Physiquement, le signal de parole est décrit par une variation de la pression de l'air émis par le système articulatoire. Il se déplace sous la forme d'une onde de pression. Pour traiter le signal de parole de façon numérique, cette onde de pression est transformée en signal électrique à l'aide d'un microphone, filtrée par un filtre passe-bande, puis numérisée. La numérisation se fait en deux temps. D'abord le signal électrique est échantillonné périodiquement; on obtient ainsi un signal discret. Ensuite, l'amplitude de chaque échantillon est quantifiée; on obtient ainsi un signal numérique.

L'échantillonnage transforme le signal analogique, $s(t)$, continu dans le temps, en un signal à temps discret : $s(nT_S)$. Ce nouveau signal est défini aux instants d'échantillonnage correspondant aux multiples entiers, n, de la période d'échantillonnage, $T_S = 1/F_S$, où F_S, est la fréquence d'échantillonnage. Pour satisfaire le théorème de Shannon, la fréquence d'échantillonnage doit être au moins deux fois plus grande que la fréquence la plus élevée contenue dans le signal traité. Pour traiter le signal de parole en bande élargie, le spectre du signal analogique est limité à l'intervalle situé entre 50 et 7000 Hz, puis ce signal est échantillonné à 16000 Hz.

Un convertisseur analogique / numérique est utilisé pour quantifier les échantillons de parole $s(nT_S)$. Chaque échantillon est représenté par un nombre sélectionné dans un ensemble de L valeurs. Le signal numérique résultant est noté $s(n)$. Le nombre de bits B, nécessaire à encoder de façon binaire un ensemble de L valeurs, vaut $B = \lceil \log_2 L \rceil$, où la fonction $\lceil \log_2 L \rceil$ correspond au nombre entier supérieur ou égal à $\log_2 L$. La quantification

d'un signal induit généralement une distorsion de celui-ci; cette distorsion, appelée bruit de quantification, est inversement proportionnelle à L.

Pour traiter le signal de parole, il est nécessaire de comprendre son processus de production, afin de le modéliser. La sous-section suivante décrit ce processus.

2.2.2 Processus de production de la parole

Le processus de production de la parole se résume en trois phases :

- La génération d'une énergie de ventilation ou source acoustique, utilisée pour mettre en mouvement les cordes vocales et / ou générer des bruits;
- La vibration des cordes vocales, qui donne naissance à tous les sons voisés, et / ou à l'apparition de bruits d'explosion ou de friction (sons non-voisés);
- La réalisation d'une gestuelle articulatoire au niveau du conduit vocal, des fosses nasales et des lèvres.

Ainsi, la parole résulte de l'excitation des cavités nasales et / ou orales par une ou deux sources acoustiques :

- La première source acoustique, ou source laryngienne, est quasi-périodique. Elle génère une onde de débit. Cette source se situe au niveau du larynx, à la base de la trachée vocale, où le flux d'air provenant des poumons est interrompu périodiquement par les cordes vocales, qui sont ainsi excitées et vibrent. Il en résulte des sons à caractère voisé, qui présentent une périodicité et donc une fréquence fondamentale F_0. Cette fréquence correspond au taux de vibration des cordes vocales. Elle est couramment appelée "pitch". Le spectre du signal de parole voisée contient les harmoniques de F_0. Le délai tonal T_0 est l'inverse de la fréquence fondamentale.
- La deuxième source acoustique est non-périodique. Elle peut s'ajouter ou se substituer à la première : elle se présente sous forme de bruits. Ce sont soit des bruits d'explosion, de friction ou des turbulences créées par l'air s'écoulant rapidement dans une constriction du conduit vocal (de la glotte aux lèvres), soit des bruits dus au brusque relâchement d'une occlusion. Cette source acoustique est à l'origine de sons non-voisés.

Ces sources excitent le conduit vocal, dont la disposition dépend de l'articulation prononcée par le locuteur. Elles peuvent évoluer rapidement dans le temps (dans une gamme d'une octave autour de 110 Hz pour les

hommes et de 210 Hz environ pour les femmes), et rendent ainsi le signal de parole évolutif.

Chaque nouvelle position du système articulatoire entraîne des modifications du spectre des sons émis. Ces modifications sont liées d'une part aux propriétés de résonance inhérente aux cavités de l'appareil vocal, et d'autre part aux caractéristiques du rayonnement au niveau des lèvres. Il existe de nombreuses différences anatomiques entre locuteurs et les stratégies articulatoires présentent de grosses variations intra- et inter-locuteurs. De plus, la position du larynx varie avec l'âge du locuteur : il s'abaisse progressivement jusqu'à la puberté. Chez la femme il est plus élevé, ce qui entraîne une diminution de la longueur du pharynx. Cette différence homme / femme a des conséquences lors de la production de sons voisés et en particulier sur la fréquence fondamentale, qui est fonction de la taille des cordes vocales. Cette fréquence se trouve typiquement entre 80 et 160 Hz pour un homme et entre 132 et 223 Hz pour une femme.

La parole résultant du processus de production peut être classée dans deux catégories générales :

- La parole voisée, caractérisée par une quasi-périodicité et par des segments de grande énergie, telle que les voyelles.
- La parole non-voisée, caractérisée par des segments non-périodiques, généralement de faible énergie, telle que les consonnes.

La Figure 2.1 illustre chacune de ces catégories. Elle montre la différence en évolution temporelle (amplitude normalisée) entre un signal de parole voisée et un signal de parole non-voisée. On constate principalement le caractère périodique, respectivement non-périodique, du signal voisé et du signal non-voisé. Une description détaillée des sous-catégories de parole voisée et non-voisée est donnée en [2-3].

Certaines parties du signal de parole correspondent à un mélange de sons voisés et non-voisés et donc à un mélange des deux types de sources acoustiques. Un tel mélange apparaît typiquement lors d'une période transitoire entre la prononciation d'un signal voisé et celle d'un signal non-voisé.

Le processus complet de production de la parole est très complexe. Il dépend de l'interaction d'un nombre élevé de paramètres. Pour réaliser un codage du signal numérique de parole, capable d'être implanté sur un DSP, il est nécessaire de recourir à une modélisation de ce processus afin d'en réduire la complexité. La modélisation auto-régressive est une méthode particulièrement adaptée au traitement du signal de parole. Cette modélisation est introduite à la Section 2.3.

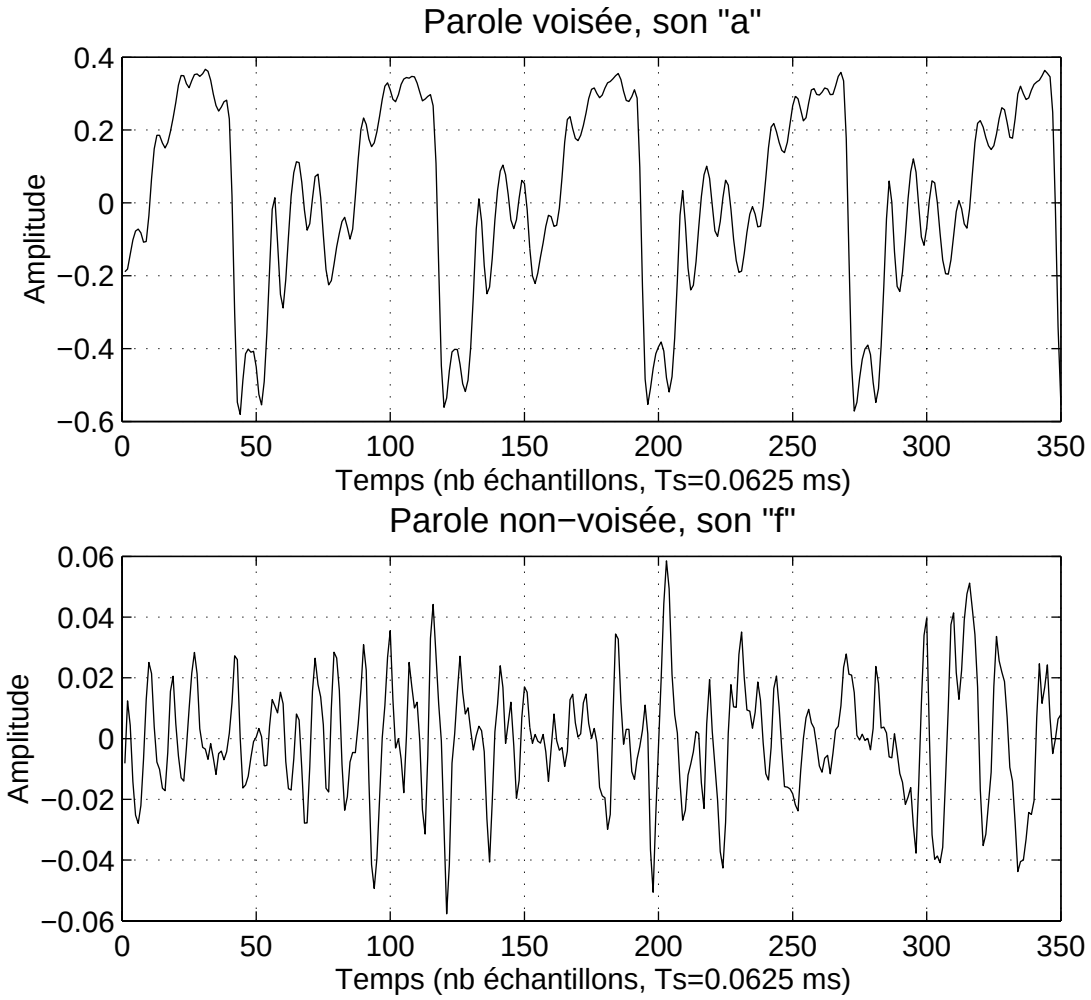

Figure 2.1 : Evolutions temporelles typiques du signal de parole voisée et non-voisée (Ts est la période d'échantillonnage).

Les paramètres les plus utiles à la modélisation de la production du signal vocal se trouvent dans le domaine spectral. Ces paramètres sont introduit à la Sous-section 2.2.3.

2.2.3 Signal de parole dans le domaine des fréquences

Les propriétés spectrales du signal vocal, et en particulier ses propriétés en amplitude spectrale, présentent un intérêt majeur si l'on considère la perception auditive de l'oreille humaine. En effet, l'oreille procède à une analyse spectrale de l'onde acoustique reçue. Elle accorde beaucoup plus d'importance à l'amplitude spectrale du signal qu'à sa phase ou qu'à son évolution dans un court intervalle de temps. En outre, l'analyse du signal de parole est plus facile et plus régulière dans le domaine fréquentiel.

Dans le domaine spectral, le processus de production de la parole se décompose en deux modèles : l'un relatif à la transmittance du conduit vocal

et au rayonnement des lèvres, qui agissent comme un filtre sur le spectre du signal; l'autre relatif à la source acoustique (impulsion glottique ou passage turbulent de l'air dans une constriction du conduit vocal).

Bien que le signal de parole ne soit pas stationnaire, on peut le considérer comme tel sur de courts intervalles de temps (10 à 30 ms). On parle de stationnarité à court-terme. Pour procéder à l'analyse du signal, un fenêtrage est nécessaire afin d'en isoler des séquences aux propriétés statistiques stationnaires dans le domaine spectral.

Une forme d'analyse spectrale couramment utilisée, est la transformée de Fourier. Pour un signal de parole dont on désire faire l'analyse à court-terme, elle est définie comme suit :

$$S_k(e^{j\omega}) = \sum_{n=-\infty}^{\infty} w(k-n)\cdot s(n)\cdot e^{-j\omega n}. \tag{2.1}$$

$w(n)$, $s(n)$ et ω sont respectivement la fenêtre à court-terme, le signal de parole numérique et la fréquence angulaire normalisée. Le choix de la longueur de la fenêtre, de sa forme, ainsi que de la longueur de recouvrement entre fenêtres successives, résulte d'un compromis entre la résolution temporelle et la résolution spectrale. La fenêtre couramment utilisée dans ce cas, est une fenêtre de Hamming dont la longueur est comprise entre 10 et 30 ms. Le recouvrement entre fenêtres successives vaut typiquement 5 à 10 ms.

La fréquence angulaire ω est liée à la fréquence réelle F par l'équation suivante :

$$\omega = 2\pi F / F_s, \tag{2.2}$$

où F_s est la fréquence d'échantillonnage du signal. Une variable couramment utilisée, pour le traitement du signal de parole, est la fréquence normalisée f liée à F et à ω par les relations suivantes :

$$f = F / F_s, \;\; f = \omega / 2\pi. \tag{2.3}$$

Le signal de parole est un signal réel. Son spectre numérique est périodique en ω et présente une symétrie paire par rapport à l'origine. Par conséquent, la gamme de fréquences intéressantes pour l'analyse spectrale est donnée par :

$$0 \le \omega \le \pi, \;\; 0 \le f \le 0.5, \;\text{ et }\; 0 \le F \le F_s / 2.$$

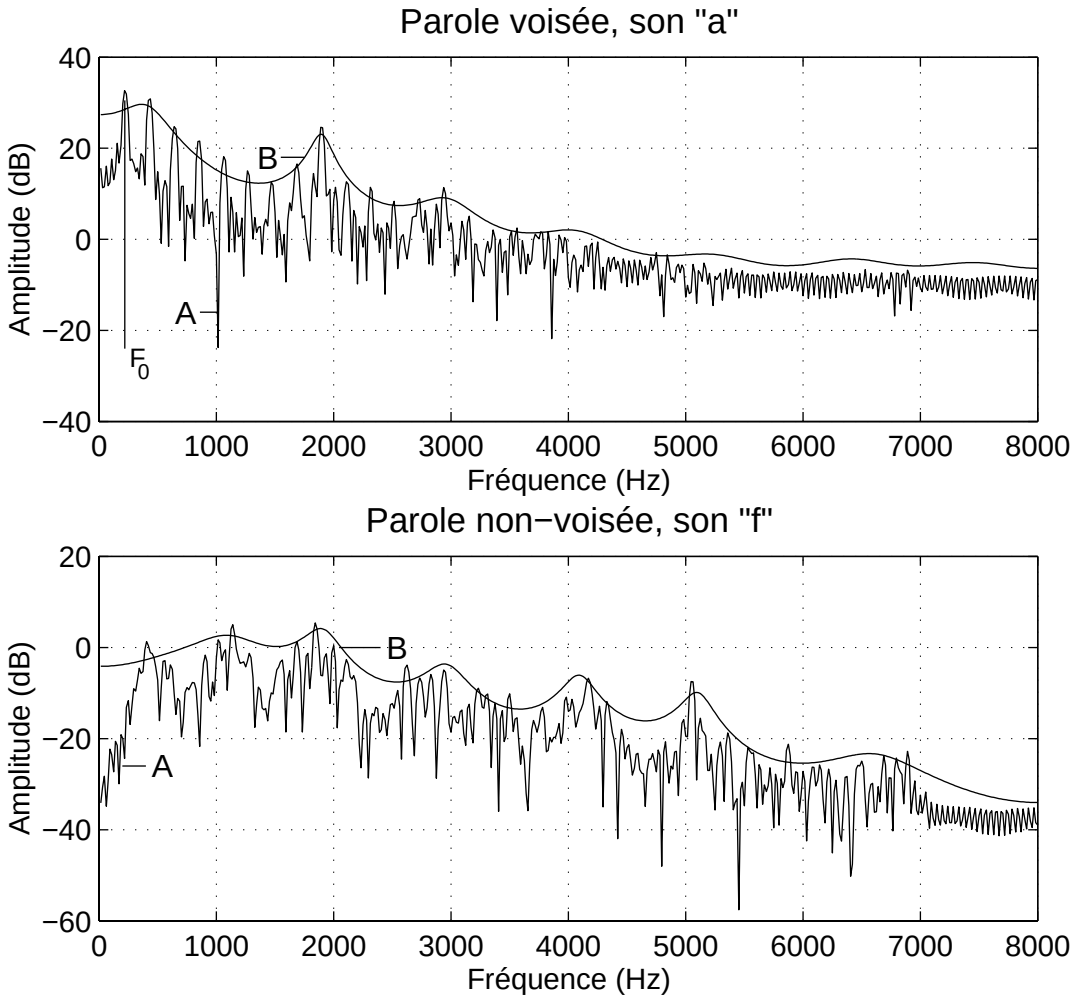

Figure 2.2 : Spectres de puissance (A) des signaux de parole voisée et non-voisée illustrés à la Figure 2.1, ainsi que leur enveloppe spectrale respective (B), pour une prédiction d'ordre 16.

L'enveloppe spectrale du signal de parole présente des pics qui correspondent aux fréquences des formants. L'estimation de la fréquence et de la sélectivité des principaux formants constitue une tâche essentielle pour de nombreuses applications du traitement du signal de parole. Les spectres de puissance des signaux de parole voisée et non-voisée illustrés à la Figure 2.1, ainsi que leur enveloppe spectrale respective, calculée pour une prédiction d'ordre 16 (cf. Sous-section 2.3.1), sont illustrés à la Figure 2.2.

Le spectre du son voisé montre une fréquence fondamentale F_0, correspondant à la fréquence de vibration des cordes vocales, ainsi que ses harmoniques. La plupart de l'énergie est concentrée dans les harmoniques de plus basses fréquences. L'extraction de la fréquence fondamentale est importante pour réaliser une bonne analyse / synthèse de la parole. Le spectre des sons non-voisés ne contient pas de structure harmonique.

La transformée de Fourier est une transformation inversible qui conserve toute l'information contenue dans le signal. Cependant, elle est une opération

coûteuse en terme de calcul. Une forme plus économique d'analyse spectrale de la parole est obtenue par prédiction linéaire. La Section 2.3 introduit la modélisation du signal de parole. La modélisation du signal de parole par prédiction linéaire est décrite au point 2.3.1.

2.3 Modélisation du signal de parole

La production de la parole se schématise par une opération de filtrage, où des sources acoustiques excitent le conduit vocal qui agit comme un filtre. Quelle que soit la source d'excitation, le conduit vocal modifie la distribution énergétique du spectre de la source d'excitation. Il amplifie certaines fréquences du son et en atténue d'autres. Il introduit ainsi des résonances (formants) et des anti-résonances (antiformants également appelés vallées spectrales). Si l'on représente le conduit vocal comme un filtre variant dans le temps, les résonances et les anti-résonances sont dues respectivement aux pôles et aux zéros de la réponse en fréquences du conduit vocal.

Lorsque la modélisation a pour but la compression du signal, on cherche à exploiter sa redondance propre, afin de ne transmettre (ou stocker) que l'information non prédictible. On réduit ainsi le débit d'information nécessaire à la transmission (ou au stockage).

Le signal de parole est redondant puisque les mécaniques des organes de la parole (conduit vocal et cordes vocales) sont limitées. Cette redondance s'observe dans le domaine des fréquences. En effet :

- Le spectre de la parole change relativement lentement dans le temps (excepté pendant des occlusives[1]).
- Les fréquences fondamentales successives sont généralement similaires.
- L'enveloppe spectrale du signal est relativement lisse; la plupart de l'énergie du signal est concentrée dans les basses fréquences.

La redondance contenue dans le signal de parole se traduit par une corrélation de ses échantillons. La corrélation à court-terme (entre échantillons voisins) se retrouve dans la structure formantique du spectre du signal. La corrélation à long-terme se retrouve dans la structure harmonique du spectre du signal. On peut exploiter la corrélation en utilisant une prédiction linéaire (cf. Sous-section 2.3.1).

[1] Les occlusives sont le résultat d'une pression sur une fermeture totale, quelque part le long de l'appareil vocal, suivie d'un relâchement rapide. Pendant la fermeture totale, l'onde produite a une amplitude acoustique très faible.

La composante du signal non prédictible correspond à la source acoustique, utilisée pour mettre en mouvement les cordes vocales et / ou générer des bruits. Cette composante ne contient aucune corrélation et ne peut être synthétisée par prédiction linéaire. Dans les codeurs prédictifs, elle correspond au résidu de prédiction. Sa quantification se fait en réalisant une analyse par synthèse. La méthode de codage utilisant l'analyse par synthèse (LPAS) est introduite à la Sous-section 2.3.2.

Dans le cas du signal de parole, la réduction du débit d'information à transmettre (ou à stocker) s'obtient également en utilisant les limites de perception de l'oreille humaine. En effet, celle-ci donne plus d'importance aux pôles spectraux qu'aux zéros. Cette propriété est appelée "effet de masque de l'oreille" [2-4]. Elle est traitée à la Section 2.4.

2.3.1 Signal de parole modélisé par prédiction linéaire

La prédiction linéaire permet d'exploiter les propriétés de corrélation des signaux de parole afin de les encoder efficacement. Comme le signal de parole est redondant, chacun de ses échantillons, $s(n)$, peut être approché par une combinaison linéaire des échantillons qui le précèdent dans l'espace temporel. La prédiction linéaire est utilisée, soit pour éliminer la redondance du signal de parole, soit pour modéliser le conduit vocal et le rayonnement des lèvres.

La redondance est retirée du signal de parole à l'aide de filtres de codage par prédiction linéaire (LPC). Le filtre de prédiction linéaire à court-terme élimine la redondance entre échantillons voisins. Ce filtre $A(z)$, appelé filtre d'analyse LPC, varie dans le temps. Il élimine la structure formantique du signal de parole.

La sortie de ce filtre, $x(n)$, est une erreur de prédiction de basse énergie, couramment appelée "le résidu de prédiction linéaire" ou "le signal d'excitation". L'inverse du filtre d'analyse LPC est le filtre de synthèse formantique $H_p(z)$. Il modélise le conduit vocal. Sa fonction de transfert décrit le spectre du signal de parole. Idéalement ce filtre contient aussi bien des pôles que des zéros. Cependant, pour réduire la complexité de l'analyse spectrale on suppose que c'est un filtre tous-pôles, dont la fonction de transfert est donnée par :

$$H_p(z) = \frac{S(z)}{X(z)} = \frac{1}{1 + \sum_{k=1}^{p} a_p(k) \cdot z^{-k}} = \frac{1}{A_p(z)}, \tag{2.4}$$

où $\{a_p(1),\ldots, a_p(p)\}$ sont les coefficients LPC, où p est l'ordre du filtre, et où $S(z)$ et $X(z)$ sont les transformées de Fourrier de $s(n)$ et respectivement de $x(n)$. Typiquement, un ordre de 10 et de 16 est respectivement utilisé pour la parole en bande étroite et en bande élargie. En utilisant de tels ordres, les résonances formantiques ainsi que l'enveloppe spectrale générale du signal sont correctement modélisées. Les spectres LPC d'ordre 16 des signaux de parole en bande élargie de la Figure 2.1, sont illustrés à la Figure 2.2.

Le filtre d'analyse LPC est donné par :

$$A_p(z) = 1 + \sum_{k=1}^{p} a_p(k) \cdot z^{-k}. \tag{2.5}$$

L'équation (2.4) transposée dans le domaine du temps devient :

$$x(n) = s(n) + \sum_{k=1}^{p} a_p(k) \cdot s(n-k) = s(n) - \hat{s}(n), \tag{2.6}$$

où $\hat{s}(n)$ est donné par :

$$\hat{s}(n) = -\sum_{k=1}^{p} a_p(k) \cdot s(n-k). \tag{2.7}$$

L'échantillon de parole courant $s(n)$ est ainsi prédit par une combinaison linéaire des p échantillons qui le précèdent : $\hat{s}(n)$. L'opération décrite par l'équation (2.6) est illustrée à la Figure 2.3.

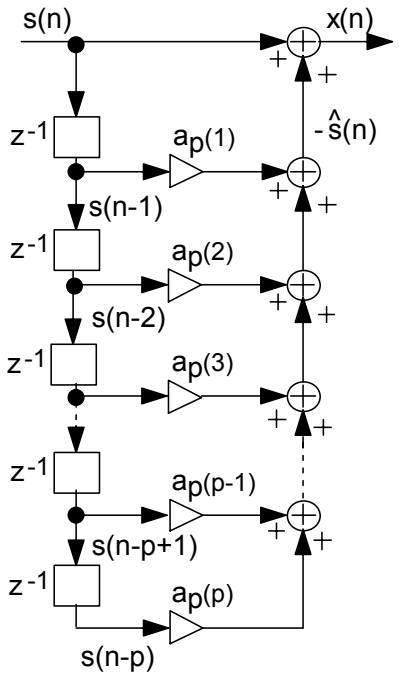

Figure 2.3 : Filtre d'analyse LPC, $A_p(z)$ appliqué au signal $s(n)$.

Le problème de l'analyse LPC d'ordre p est donc formulé ainsi : soit l'échantillon de parole courant $s(n)$, il faut déterminer les paramètres $\{a_p(1),\ldots,a_p(p)\}$ qui minimisent l'erreur de prédiction $x(n)$. Ce problème est traité à la Sous-section 2.3.3.

Il est clair que la prédiction n'est possible que si le signal $s(n)$ est auto-corrélé. En effet, un signal non auto-corrélé, tel un bruit blanc, ne peut être prédit. La matrice d'auto-corrélation joue donc un rôle important pour cette estimation.

Ayant un nombre de coefficients limités, le filtre d'analyse LPC ne peut pas éliminer toutes les composantes harmoniques du signal de parole dans ses parties fortement voisées. Cependant, l'oreille est très sensible aux bruits inter-harmoniques [2-5]. On remarque qu'une partie de la redondance reste apparente dans le résidu de prédiction LPC. Pour obtenir une prédiction plus fine, on peut utiliser les propriétés de corrélation à long-terme du signal voisé. Ainsi, un deuxième filtre de prédiction linéaire peut être employé afin d'éliminer la redondance périodique du signal, correspondant à des échantillons passés et éloignés. Ce deuxième filtre est communément appelé le prédicteur de « pitch », le prédicteur de délai tonal, ou le prédicteur à long-terme. Il exploite les propriétés de périodicité du signal.

On peut donc ajouter au filtre d'analyse LPC à court-terme, un filtre de prédiction à long-terme, LTP (Long Term Prediction). Ce nouveau filtre d'analyse est de la forme :

$$P(z) = 1 - \sum_{i=-m}^{m} G_i z^{-(T+i)}, \tag{2.8}$$

où T est une estimation, en nombre d'échantillons du délai tonal T_0. T_0 correspond à l'inverse de la période fondamentale F_0 du signal original. Les G_i sont les coefficients de prédiction à long-terme. Ici, le niveau de prédiction vaut (2*m+1). m prend typiquement la valeur 0, voire 1. La Figure 2.4 illustre le filtre d'analyse LTP, appliqué au signal $x(n)$, pour $m = 1$.

L'inverse du prédicteur de pitch est le filtre tonal $H_{LTP}(z)$, ou filtre de synthèse à long-terme. Il modélise l'effet des vibrations des cordes vocales. Sa fonction de transfert décrit la structure harmonique du signal de parole. Naturellement, le prédicteur de pitch est inexploitable dans le cas d'une excitation non-voisée.

Les Sous-sections 2.3.3 et 2.3.4 décrivent l'extraction des paramètres de prédiction linéaire à court-terme et respectivement à long-terme.

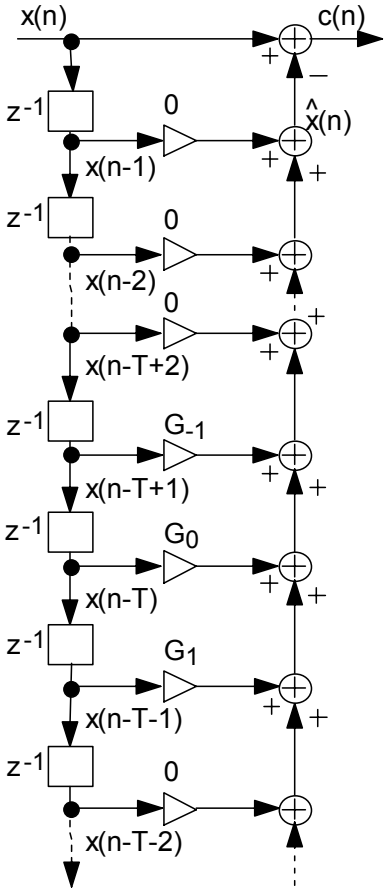

Figure 2.4 : Filtre d'analyse LTP, $P(z)$, appliqué au signal $x(n)$ pour $m = 1$.

2.3.2 Principe général de l'analyse par synthèse et codeur CELP

Les codeurs utilisant l'analyse par synthèse prévoient un décodeur local au niveau de l'encodeur. Ce décodeur local permet de calculer l'excitation nécessaire aux filtres de synthèse LPC et LTP, de façon à minimiser l'erreur commise entre la parole originale et la parole synthétisée ou reconstruite, en particulier dans le décodeur.

Le principe de l'analyse par synthèse repose sur une recherche des caractéristiques d'un vecteur de signal $\mathbf{s}$ par synthèses successives de vecteurs-candidats $\mathbf{s}_i$. Ces vecteurs-canditats sont reconstruits sur la base de vecteurs d'excitations contenus dans une base de données, également appelée dictionnaire, ou sont produits par un générateur d'excitations. Les synthèses successives se font grâce au décodeur local de l'encodeur. L'indice du vecteur-candidat permettant la meilleure reconstruction est transmis au décodeur. La meilleure reconstruction est celle qui minimise la distance $d(\mathbf{s}, \mathbf{s}_i)$ définie ainsi :

$$d(\mathbf{s},\mathbf{s}_i) = (\mathbf{s}-\mathbf{s}_i)^T \cdot \mathbf{Q} \cdot (\mathbf{s}-\mathbf{s}_i), \tag{2.9}$$

où $\mathbf{Q}$ est la matrice quadratique définie positive dont est issue la distance utilisée[1].

Les codeurs LPAS font intervenir le codage par prédiction linéaire et procèdent à un traitement du résidu de prédiction en utilisant l'analyse par synthèse. Un codeur LPAS contient deux ou trois éléments de base :

- Un filtre de synthèse LPC, $H_p(z)$, modélisant la corrélation à court-terme du signal.
- Un filtre de synthèse LTP $H_{LTP}(z)$, modélisant la corrélation à long-terme du signal. Cet élément est facultatif.
- Un signal d'excitation innovateur, modélisant le résidu de prédiction LPC et LTP. Ce signal est déterminé en utilisant l'analyse par synthèse. Parmi un ensemble d'excitations possibles, on choisit celle qui après avoir été filtrée par les filtres sus-mentionnés, produit la meilleure reconstruction du signal. Ce signal d'excitation est également appelé excitation à court-terme.

La Figure 2.5 illustre un encodeur LPAS. La Figure 2.6 présente le décodeur qui lui correspond. Le bloc A de la Figure 2.5 localise le décodeur local inclus dans l'encodeur.

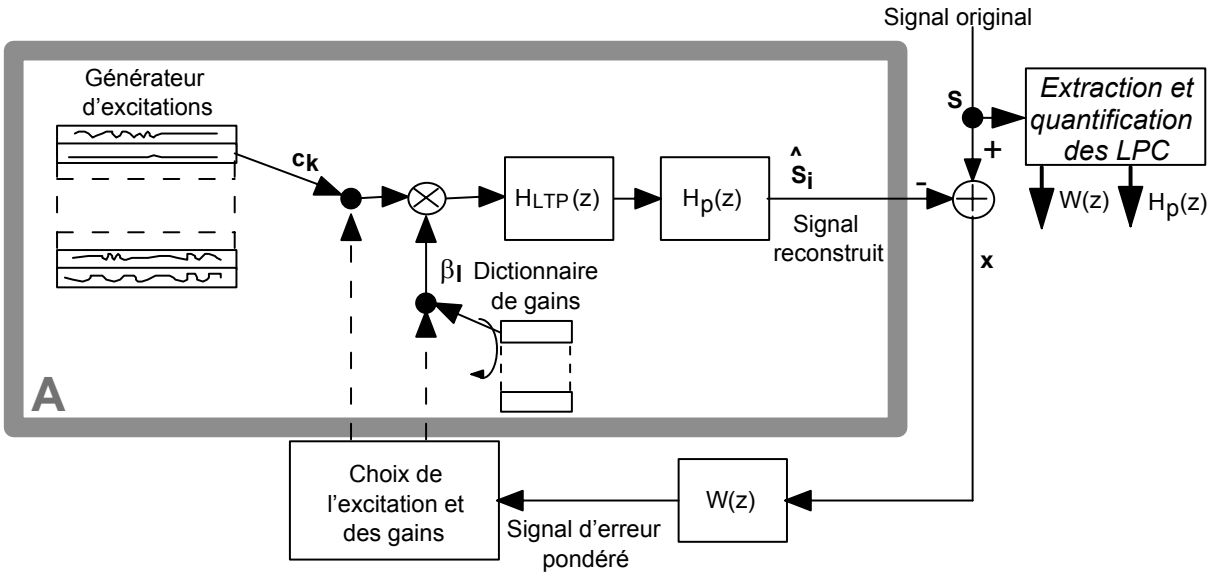

Figure 2.5 : Schéma d'un encodeur LPAS. Le bloc A situe le décodeur local au niveau de l'encodeur.

[1] Le symbole T apparaissant dans l'équation (2.9) dénote la transposition vectorielle.

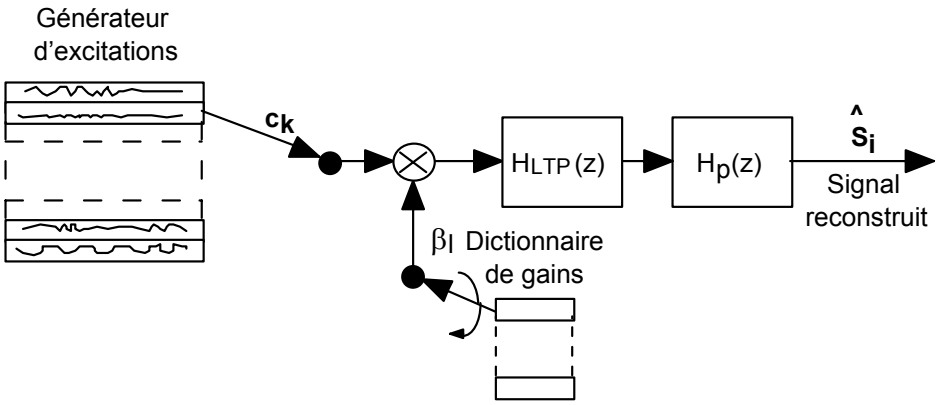

Figure 2.6 : Schéma d'un décodeur apparié au codeur de la Figure 2.5.

Pour tous les codeurs LPAS, il est jugé préférable de rafraîchir l'excitation plus fréquemment que les coefficients LPC et ce dans un rapport souvent choisi égal à 4 [2-6]. Le fait d'augmenter le taux de remise à jour des excitations permet de réduire la complexité de calcul en réduisant la longueur de l'excitation.

La mesure de distance $d(\mathbf{s}, \mathbf{s}_i)$ utilisée ici se base sur des critères de perception de l'oreille. Ces critères sont présentés en Section 2.4. Ils sont modélisés par un filtre $W(z)$, appelé filtre de pondération perceptuelle.

Pour réduire la complexité de calcul relative à l'extraction des paramètres du filtre de prédiction à long-terme, on peut remplacer ce filtre par un dictionnaire adaptatif. Ce dictionnaire contient des versions répétées des excitations des sous-trames précédentes. On appelle les mots de code qu'il contient, des excitations adaptatives puisque celles-ci varient en s'adaptant au signal d'entrée. Une description détaillée du dictionnaire adaptatif et de son utilisation est donnée à la Sous-section 2.3.4. Si l'on utilise un dictionnaire adaptatif à la place du filtre LTP, alors l'entrée du filtre de synthèse à court-terme sera composée de la somme pondérée de l'excitation adaptative et d'une excitation dite innovatrice (ou à court-terme), extraite d'un générateur d'excitation. Les poids servant à la pondération de la somme pré-citée sont extraits de deux dictionnaires de gains. Un codeur LPAS utilisant une telle implantation est illustré à la Figure 2.10.

Les codeurs CELP sont une famille de codeurs LPAS. Ils ont la particularité de coder le résidu de prédiction LPC et LTP à l'aide d'un quantificateur vectoriel appelé également dictionnaire d'excitations (innovateur ou à court-terme).

L'extraction de l'excitation optimale est une charge de calcul importante dans un codeur LPAS ou CELP. Cette charge dépend de la dimension et du nombre des excitations à tester. La charge de calcul est d'autant plus

importante dans un codeur en bande élargie, puisque par rapport au cas de la bande étroite, la fréquence d'échantillonnage a doublé. Pour éviter une forte augmentation du débit de transmission, des trames de parole ayant un nombre de points plus élevé, et correspondant à un temps plus long, doivent être utilisées pour extraire les paramètres nécessaires à la synthèse de la parole. Dans le cadre de la recherche présentée ici, et afin de réduire la complexité du codeur CELP pour la bande élargie, une excitation innovatrice de type algébrique est utilisée. L'implantation d'une telle excitation est décrite à la Sous-section 2.3.5.

2.3.3 Extraction des paramètres de prédiction linéaire à court-terme

On détermine la valeur des coefficients LPC en minimisant l'énergie, ε_p, de l'erreur de prédiction, $x(n)$ (équation (2.6)), selon la méthode des moindres carrés :

$$\varepsilon_p = \sum_{n=-\infty}^{+\infty} x^2(n) = \sum_{n=-\infty}^{+\infty} \left[s(n) + \sum_{k=1}^{p} a_p(k) \cdot s(n-k) \right]^2 . \qquad (2.10)$$

Cette somme ne peut être infinie. Elle doit être limitée en fenêtrant soit le signal original, soit le signal d'erreur. Les coefficients LPC sont ainsi déterminés soit par la méthode dite d'auto-corrélation, soit par la méthode dite de covariance. On ne s'intéresse ici qu'à la méthode d'auto-corrélation, puisqu'elle est plus efficace en termes de complexité de calcul et puisqu'elle produit toujours un filtre de synthèse stable [2-7]. Avec cette méthode, les coefficients LPC sont calculés en utilisant la méthode de récursion de Levinson-Durbin introduite au paragraphe suivant.

2.3.3.1 Récursion de Levinson-Durbin

Le signal de parole et en particulier son enveloppe spectrale sont considérés comme stationnaires sur un intervalle de temps compris entre 10 et 30 ms. Par conséquent, on effectue généralement l'analyse par prédiction linéaire sur des trames de parole $\{\mathbf{s}_1, \ldots, \mathbf{s}_N\}$ isolées par une fenêtre dont la longueur L correspond généralement à 10, 20 ou 30 ms de signal. La minimisation de l'énergie de l'erreur de prédiction ε_p (équation (2.10)), par rapport aux coefficients LPC, se base sur l'équation de Yule-Walker [2-8] :

$$\mathbf{R}_p \cdot \mathbf{a}_p = -\mathbf{r}_p , \qquad (2.11)$$

où[3] :

$$\mathbf{R}_p = \begin{bmatrix} r_0 & r_1 & r_2 & \cdots & r_{p-1} \\ r_1 & r_0 & r_1 & \cdots & r_{p-2} \\ r_2 & r_1 & r_0 & \cdots & r_{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{p-1} & r_{p-2} & r_{p-3} & \cdots & r_0 \end{bmatrix}, \qquad \begin{aligned} \mathbf{a}_p &= \begin{bmatrix} a_p(1) & \cdots & a_p(p) \end{bmatrix}^T, \\ \mathbf{r}_p &= \begin{bmatrix} r_1 & r_2 & \cdots & r_p \end{bmatrix}^T, \end{aligned} \tag{2.12}$$

et où r_k est le *k*-ième coefficient d'auto-corrélation du signal de parole fenêtré :

$$r_k = \sum_{n=k}^{L-1} w(n) \cdot s(n) \cdot w(n-k) \cdot s(n-k). \tag{2.13}$$

$\{w(n)\}$ est la fonction qui fenêtre les N échantillons d'une trame de signal. Sur la base de l'équation (2.11), les coefficients LPC sont donnés par :

$$\mathbf{a}_p = -\mathbf{R}_p^{-1} \cdot \mathbf{r}_p. \tag{2.14}$$

La matrice d'auto-corrélation $\mathbf{R}_p$ a une structure de Toepliz. La solution de l'équation (2.14) s'obtient avec la récursion de Levinson-Durbin [2-9] décrite ci-dessous. On pose :

$$\varepsilon_0 = r_0, \tag{2.15}$$

puis pour $1 \le m \le p,$ on effectue les équations (2.16) à (2.19)

$$a_m(0) = 1, \tag{2.16}$$

$$a_m(m) = k_m = \frac{\left[-r_m - \sum_{i=1}^{m-1} a_{m-1}(i) \cdot r_{m-i} \right]}{\varepsilon_{m-1}}, \tag{2.17}$$

$$a_m(i) = a_{m-1}(i) + k_m \cdot a_{m-1}(m-i), \qquad \text{pour} \quad 1 \le i \le m-1, \tag{2.18}$$

$$\varepsilon_m = \varepsilon_{m-1} \cdot (1 - k_m^2) \tag{2.19}$$

Les valeurs $\{k_m\}$ sont appelées coefficients de Parcor (partial correlation) ou coefficients de réflexion. Pour la parole en bande élargie, l'ordre de prédiction

[3] Le symbole T apparaissant dans l'équation (2.12) dénote la transposition vectorielle.

vaut typiquement $p = 16$. Avec un tel ordre, la complexité de calcul de la récursion de Levinson-Durbin est de 272 multiplications, 256 additions, et 16 divisions. Pour la suite de ce rapport, on considère un ordre p de 16.

La Figure 2.7 montre, dans le plan Z, la position des zéros du filtre d'analyse LPC d'ordre 16, pour un segment de 20 ms de la voyelle "a" illustrée à la Figure 2.1. Ces zéros correspondent aux pôles du filtre de synthèse LPC, dont le spectre est illustré à la Figure 2.8. Les formants sont les résonances, ou les pics aigus du spectre de puissance. Ils sont dus aux pôles qui se trouvent très proches du cercle unité. Plus les pôles sont proches du cercle unité, plus la largeur de bande des formants est étroite et l'amplitude augmente. Les coefficients LPC donnent une description attractive de l'enveloppe spectrale du signal, puisqu'ils décrivent les pics spectraux qui sont perceptuellement plus importants que les vallées spectrales [2-10].

Les LPC doivent être quantifiés avec soin pour leur transmission (ou stockage). En effet, une mauvaise quantification des LPC pourrait entraîner un filtre de synthèse à court-terme instable. La complexité de calcul liée à une quantification stable des LPC est très élevée. Pour cette raison, les LPC sont généralement quantifiés sous une représentation alternative. Cette nouvelle représentation est décrite à la Sous-section 3.3.4.

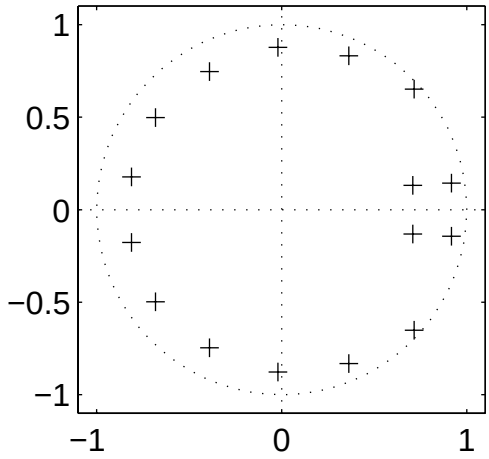

Figure 2.7 : Positions des zéros du filtre d'analyse LPC d'ordre 16, $A_{16}(z)$, pour un segment de 20 ms de la voyelle « a », illustrée à la Figure 2.1.

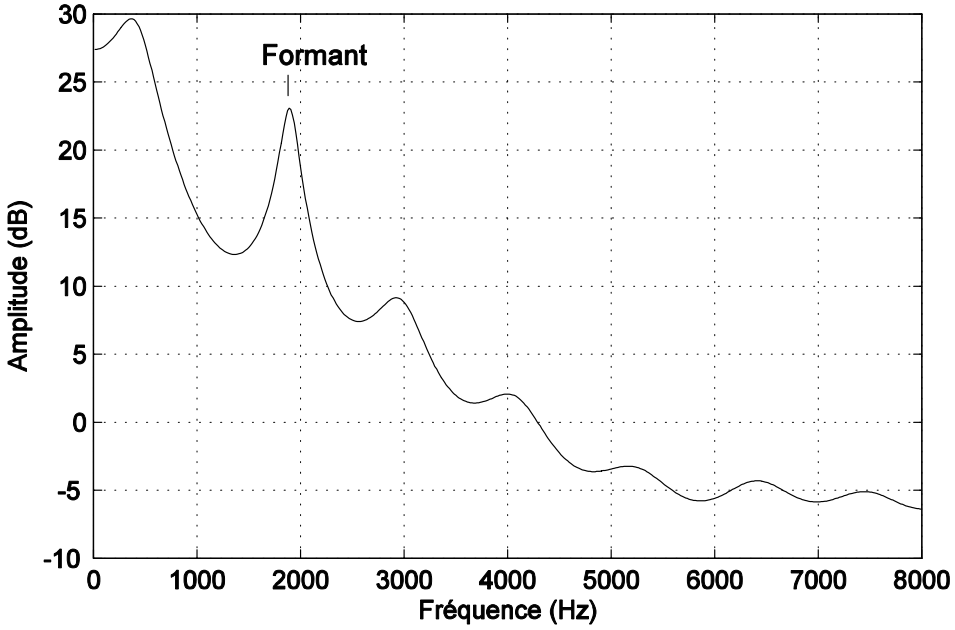

Figure 2.8 : Spectre de puissance LPC pour un segment de 20 ms de la voyelle « a », illustrée à la Figure 2.1.

2.3.4 Extraction des paramètres de prédiction linéaire à long-terme

La recherche du délai tonal peut se faire selon deux méthodes. La première consiste en une recherche en boucle ouverte. Dans ce cas, on estime le délai tonal sur la base du signal de parole passé et on calcule ensuite le gain. L'estimation du délai tonal se fait en corrélant le signal de parole présent et le signal passé. Le maximum de la corrélation permet d'obtenir une estimation du délai tonal ou un multiple de celui-ci. La seconde méthode, plus complexe mais plus efficace, consiste en une recherche en boucle fermée effectuée sur la base de l'analyse par synthèse. Ici, le résidu de prédiction est testé pour chaque délai tonal possible du filtre de synthèse LTP, $H_{LTP}(z)$.

Si l'on considère des voix d'enfants, de femmes et d'hommes, la fréquence fondamentale du signal de parole voisée se trouve comprise entre 50 et 600 Hz. Par conséquent, la valeur du délai tonal T_0, est comprise entre 27 et 320 échantillons pour la parole en bande élargie. Pour ne pas faire augmenter démesurément le débit de transmission, la mise à jour des paramètres de prédiction à long-terme et de l'excitation doit se faire environ toutes les 5 ms, soit pour des sous-trames de signal comprenant $N = 80$ échantillons à 16 kHz. Or le codeur LPAS illustré à la Figure 2.5 ne permet d'encoder qu'un délai supérieur ou égal à la longueur N de l'excitation innovatrice. En effet, le filtre de synthèse $H_{LTP}(z)$ ne peut contenir en mémoire que les valeurs des échantillons des excitations passées correspondant à un tel délai, puisque ce sont les dernières valeurs

synthétisées. Ainsi, pour encoder des valeurs de délai tonal T_0 inférieures à N, il faut répéter dans les mémoires correspondant à un délai inférieur à N, les T_0 derniers échantillons de l'excitation passée. Ceux-ci sont contenus dans les mémoires, qui correspondent à un délai compris entre N et $N+T_0-1$. Ce procédé est illustré à la Figure 2.9. On représente les mémoires du filtre $H_{LTP}(z)$ par un vecteur mémoire de longueur $5 \cdot N$. Les $4 \cdot N$ premières valeurs de ce vecteur, contiennent les valeurs des $4 \cdot N$ échantillons des excitations passées (20 ms). Les N dernières valeurs du vecteur mémoire, sont recalculées pour chaque valeur de T_0, en répétant les valeurs des T_0 derniers échantillons d'excitations synthétisées.

Dans le cas d'une recherche en boucle fermée, basée sur l'analyse par synthèse, l'excitation passée du filtre LPC à court-terme peut être utilisée comme un dictionnaire remis à jour au cours du traitement du signal. Le filtre de synthèse $H_{LTP}(z)$ peut être éliminé en ajoutant au codeur un dictionnaire contenant les excitations passées et dont l'indexation est liée à la période fondamentale. La Figure 2.10 illustre un codeur LPAS réalisé avec un tel dictionnaire. Ce dictionnaire est appelé dictionnaire adaptatif. Il représente le vecteur mémoire du filtre de synthèse LTP illustré à la Figure 2.9. Chaque indice du dictionnaire correspond à 80 échantillons du vecteur mémoire. L'indexation du dictionnaire est illustrée à la Figure 2.11. Naturellement, à chaque fois qu'une nouvelle excitation est calculée, ce dictionnaire est réactualisé par un décalage de N échantillons représentant le vecteur mémoire du filtre de synthèse LTP. Ce décalage est effectué vers la gauche. Pour chaque mot du dictionnaire correspondant à un délai T_0 inférieur à N, les N dernières valeurs du vecteur mémoire du filtre de synthèse LTP doivent être recalculées. L'indice du dictionnaire adaptatif et le gain qui lui est associé, sont remis à jour en accord avec la longueur N d'une sous-trame.

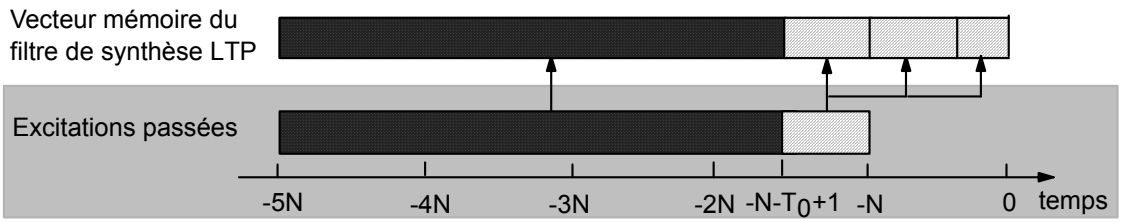

Figure 2.9 : Vecteur mémoire du filtre de synthèse LTP pour T_0 inférieur à N et excitations passées correspondantes. Le temps 0 correspond ici à la fin de l'excitation présente.

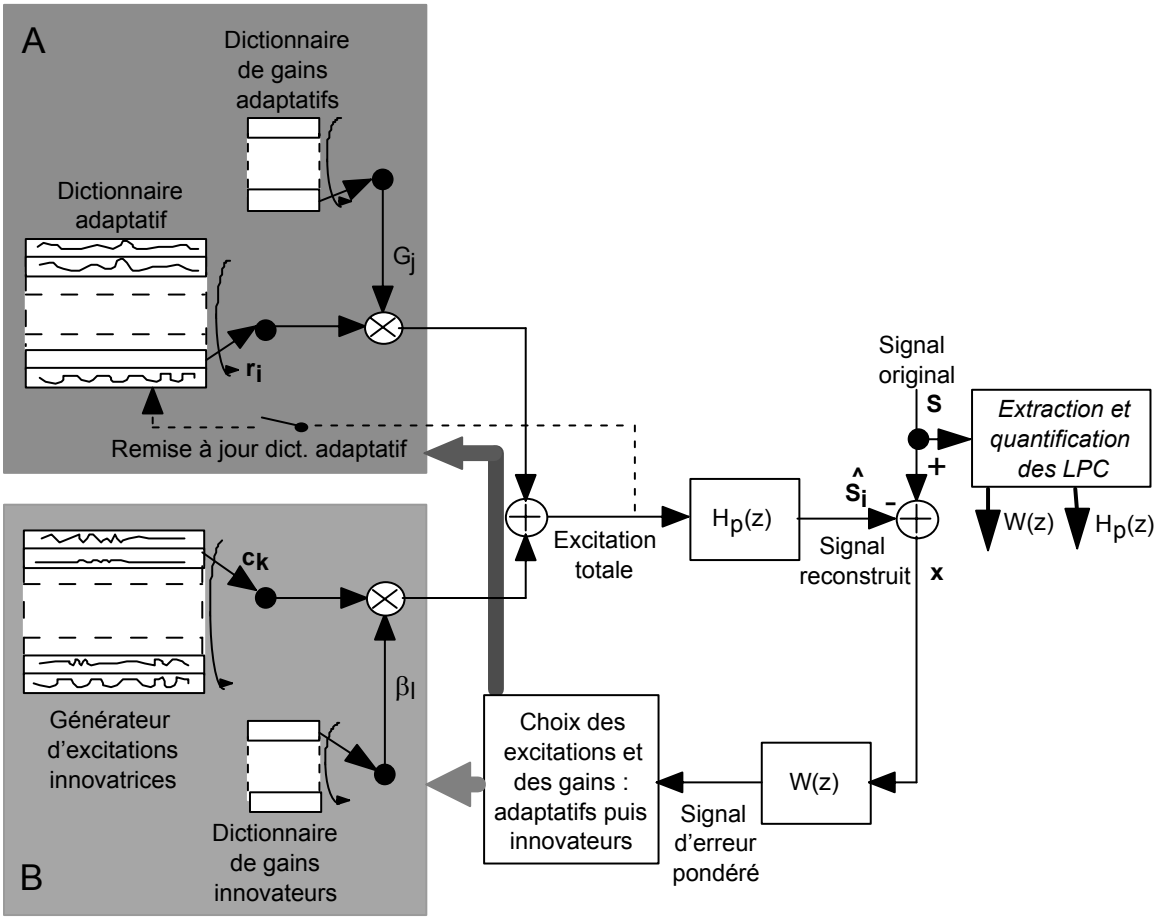

Figure 2.10 : Codeur LPAS où le filtre de synthèse LTP est remplacé par le dictionnaire adaptatif. On réalise le choix des excitations et des gains d'abord dans le bloc A, puis dans le bloc B.

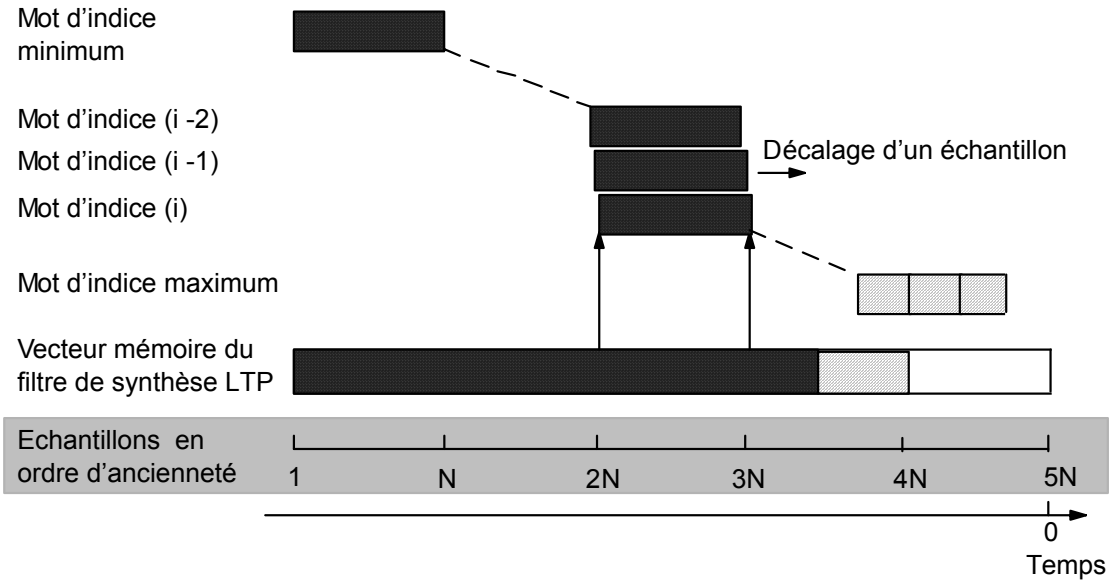

Figure 2.11 : Mots du dictionnaire adaptatif par rapport au vecteur mémoire du filtre de synthèse LTP.

Pour la suite de ce rapport, la contribution du dictionnaire adaptatif sera appelée excitation adaptative par opposition à l'excitation innovatrice, cette dernière étant établie univoquement lors de la conception de l'encodeur. Chaque excitation est pondérée par un gain, puis les excitations (adaptative et innovatrice) pondérées sont sommées et excitent le filtre de synthèse LPC, $H_p(z)$. La somme pondérée des excitations est appelée excitation totale. Les gains permettant la pondération des excitations sont extraits de deux dictionnaires distincts : le dictionnaire de gains adaptatifs et le dictionnaire des gains innovateurs. A la fin du traitement d'une sous-trame de parole, l'excitation totale sélectionnée est utilisée pour la remise à jour du dictionnaire adaptatif. Ce processus est illustré en Figure 2.10.

La Figure 2.12 illustre, pour une sous-trame donnée (numérotée st, $N = 80$), l'excitation sélectionnée dans le dictionnaire innovateur multipliée par son gain, l'excitation sélectionnée dans le dictionnaire adaptatif multipliée par son gain, l'excitation totale, ainsi que les excitations passées contenues dans le dictionnaire adaptatif, telles qu'illustrées à la Figure 2.9. Le temps 0 correspond ici à la fin de l'excitation présente. La Figure 2.13 illustre ces mêmes excitations pour la sous-trame suivante (numérotée $st+1$). Ces deux figures donnent un aperçu de l'évolution temporelle du dictionnaire adaptatif.

Dans un codeur LPAS, où le filtre de synthèse LTP est remplacé par un dictionnaire adaptatif, on choisit d'abord l'excitation adaptative dans le dictionnaire adaptatif ainsi que son gain, en considérant l'excitation innovatrice comme nulle. Puis, sur la base de ce choix on extrait l'excitation innovatrice. En effet, pendant les périodes de signal voisé, le gain de prédiction obtenu grâce au prédicteur à long-terme, ou dictionnaire adaptatif, est supérieur à celui obtenu par la génération d'une excitation innovatrice. Ainsi, un traitement inverse produirait plus de bruit non-périodique et les performances de la prédiction à long-terme s'en trouveraient dégradées.

Le filtre de prédiction à long-terme donné par l'équation (2.8), ne permet que l'implantation de valeurs entières du délai tonal. Toutefois, la valeur du délai tonal de la voix n'est pas toujours entière, et une approximation entière de celle-ci peut nuire à la qualité du signal reconstruit. Pour résoudre ce problème, on peut interpoler l'excitation du dictionnaire adaptatif, afin d'obtenir une approximation fractionnelle du délai tonal. Ce problème est traité à la Sous-section 3.4.1.

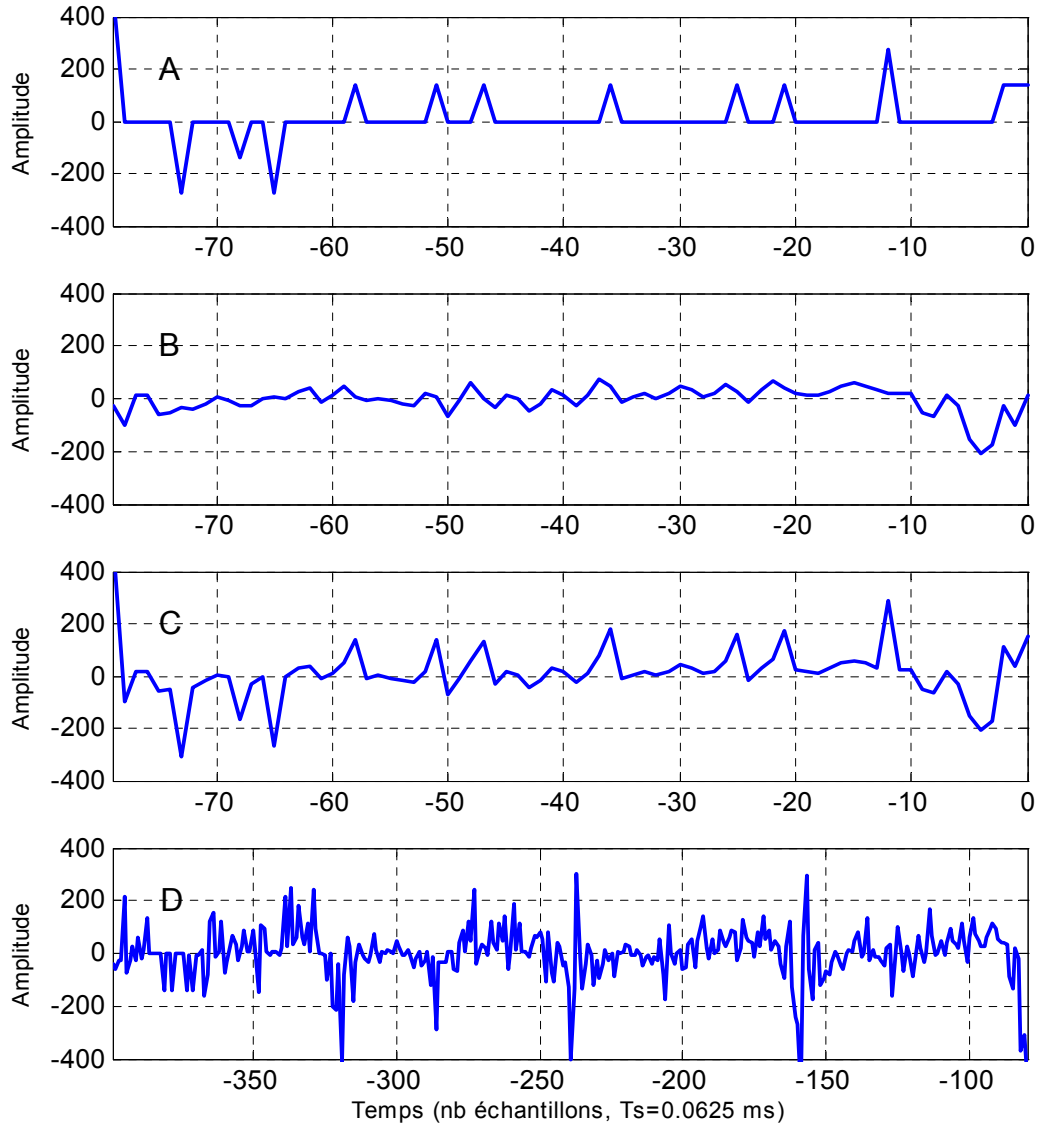

Figure 2.12 : Exemple d'excitations pour la sous-trame st ($N = 80$) : A) Excitation sélectionnée dans le dictionnaire innovateur multiplié par son gain. B) Excitation sélectionnée dans le dictionnaire adaptatif multiplié par son gain. C) Excitation totale. D) Excitations passées contenues dans le dictionnaire adaptatif. Le temps 0 correspond ici à la fin de l'excitation présente.

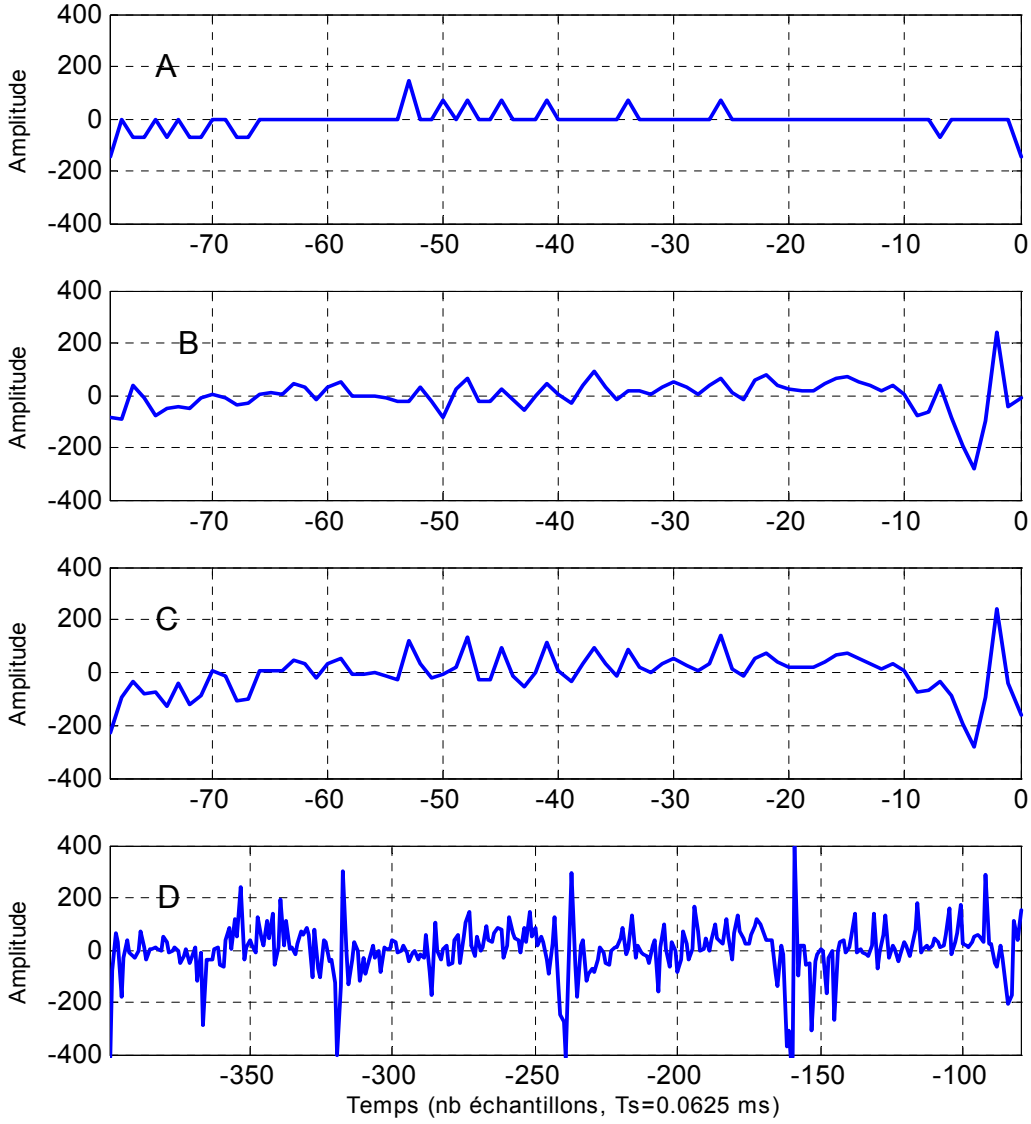

Figure 2.13 : Exemple d'excitations pour la sous-trame $st+1$ ($N=80$) : A) Excitation sélectionnée dans le dictionnaire innovateur multiplié par son gain. B) Excitation sélectionnée dans le dictionnaire adaptatif multiplié par son gain. C) Excitation totale. D) Excitations passées contenues dans le dictionnaire adaptatif. Le temps 0 correspond ici à la fin de l'excitation présente.

2.3.5 Extraction des paramètres de l'excitation innovatrice

L'excitation innovatrice représente le résidu de prédiction LPC et LTP. Elle est utilisée au niveau du décodeur pour exciter le filtre de synthèse à long-terme $H_{LTP}(z)$, suivi du filtre de synthèse à court-terme $H_p(z)$. Le premier codeur LPAS a été proposé par Atal et Remde en 1982 [2-11]. Ce codeur

générait une excitation sur la base d'un vecteur nul, auquel un certain nombre d'impulsions d'amplitude variable était ajouté. Ce codeur était appelé codeur LPC à multi-impulsions (MP-LPC). Pour simplifier la procédure de génération de l'excitation, les impulsions étaient introduites séquentiellement : la procédure d'analyse par synthèse tenait compte séquentiellement des impulsions déjà introduites. La position et l'amplitude de chacune des impulsions étaient transmises au décodeur.

Comme précité à la Section 1.2, les codeurs CELP font partie de la famille de codeur LPAS. Ils ont été introduits par Atal et Schroeder en 1984 [2-12]. Ils ont la particularité de coder le résidu de prédiction à l'aide d'un quantificateur vectoriel.

A l'origine le quantificateur CELP se présente sous la forme d'un dictionnaire contenant des vecteurs d'excitation stochastique fixes. L'encodeur sélectionne le vecteur qui produit la meilleure reconstruction du signal original, selon un critère perceptuel. L'indice du dictionnaire correspondant à ce vecteur ainsi que le gain de prédiction sont transmis au décodeur. Le décodeur possède une copie du dictionnaire et peut donc reproduire le signal synthétisé. Chaque élément des vecteurs d'excitation stochastique est généré par un nombre Gaussien aléatoire. Le dictionnaire présente donc un caractère non structuré peu adapté à une recherche efficace. Ainsi, une recherche exhaustive demande une complexité de calcul très élevée.

Dans le cadre du codage de la parole en bande élargie, une telle complexité est prohibitive. En effet, pour un tel codage, un dictionnaire d'excitation très grand est requis pour obtenir une bonne qualité de parole reconstruite. Les algorithmes CELP ordinaires n'arrivent pas à gérer de tels dictionnaires à cause d'une complexité de calcul excessive.

Depuis 1984, de nombreuses recherches ont permis de réduire la complexité des codeurs CELP, d'en accroître la robustesse face aux erreurs du canal de transmission et d'améliorer la qualité du signal reconstruit. Une partie de ces recherches est présentée en [2-13]. Cet effort de recherche a permis d'améliorer la qualité du signal d'excitation tout en réduisant la complexité de calcul liée à son extraction.

En 1987, Adoul et *al*. proposent l'utilisation de dictionnaires contenant des excitations conçues comme des vecteurs de codes ayant une structure en treillis. On appelle de tels dictionnaires, des "dictionnaires algébriques" [2-14]. Adoul et *al*. proposent en outre d'utiliser des vecteurs de codes de même énergie pour tout le dictionnaire algébrique. Ces codes sont générés à partir des codes correcteurs d'erreurs standards, où les symboles 1 et 0 sont respectivement remplacés par +1 et –1. L'utilisation des dictionnaires

algébriques donne naissance à la famille de codeur ACELP (Algebraic CELP).

La structure algébrique du dictionnaire donne plusieurs avantages. Une telle structure ne requiert pas de stockage puisque le vecteur d'excitation est généré en temps réel. Cette structure est robuste contre les erreurs de transmission, puisqu'une simple erreur de transmission ne corrompt qu'une position dans le vecteur d'excitation. De plus, la complexité de calcul pour réaliser l'analyse par synthèse est fortement réduite, ce qui permet une implantation du codeur en temps réel, également pour un codage en bande élargie. En effet, comme la plupart des composantes du vecteur d'excitation sont nulles, on peut éviter de réaliser un bon nombre de multiplications et / ou d'additions. Finalement, une telle structure permet d'utiliser une procédure de recherche efficace, avec laquelle seule une petite portion du dictionnaire est testée. Cette procédure est expliquée à la Sous-section 3.5.1.

Pour la suite de ce rapport, on ne traitera que des codeurs de type ACELP.

2.4 Appareil auditif et perception auditive du signal de parole

Dans le cadre du traitement de la parole, une bonne connaissance des mécanismes de l'audition et des propriétés perceptuelles de l'oreille est aussi importante qu'une maîtrise des mécanismes de production. En effet, l'évaluation finale d'un système de codage de la parole est soumise à l'appréciation de l'oreille humaine. Tout ce qui peut être mesuré acoustiquement n'est pas nécessairement perçu par celle-ci. Il faut donc exploiter les propriétés de perception de l'oreille pour améliorer la qualité perçue du signal synthétisé. La Sous-section 2.4.1 décrit l'appareil auditif qui réceptionne le signal de parole et le transmet au cerveau. Elle est suivie par la Sous-section 2.4.2 qui décrit la perception auditive.

2.4.1 Appareil auditif [2-15]

L'appareil auditif comporte trois parties : l'oreille externe, moyenne et interne. Il est illustré à la Figure 2.14. Les ondes sonores sont recueillies par cet appareil. Elles sont analysées dans l'oreille interne qui transmet au cerveau l'influx nerveux qui en résulte.

Le pavillon et le conduit auditif forment l'oreille externe. Celle-ci atténue les bruits parasites et guide les ondes sonores. Le conduit auditif relie le pavillon de l'oreille au tympan qui ferme son extrémité. L'oreille moyenne comprend le tympan, la trompe d'Eustache et une chaîne d'osselets : le marteau, l'enclume et l'étrier. Le tympan est une membrane élastique dont les

mouvements sont transmis à la chaîne des osselets situés dans la "caisse du tympan". Celle-ci est une cavité communiquant avec l'extérieur par la trompe d'Eustache, qui assure l'équilibre des pressions des deux côtés du tympan. La chaîne des osselets forme un bloc oscillant, transmettant les mouvements du tympan à l'étrier. Elle agit ainsi à la manière d'un piston sur la fenêtre ovale de la cochlée et, par-là, communique les vibrations sonores à l'oreille interne.

L'oreille interne comprend la cochlée et le vestibule. La cochlée est une partie auditive constituée par un canal membraneux neuro-sensoriel enroulé à la façon d'un limaçon. On le dénomme hélicotréma. Sur toute la longueur de la cochlée court la membrane basilaire, qui va en s'élargissant de la base au sommet de la spirale, appelé apex. De la même manière, sa rigidité va en s'amoindrissant. Le récepteur cochléaire, ou organe de Corti, forme l'appareil cellulo-membraneux assurant la transformation de l'énergie acoustique en potentiels neuro-sensoriels (transduction). Ces potentiels cheminent par le nerf cochléaire, dont les fibres se rassemblent dans l'axe de l'hélicotréma pour aller rejoindre le nerf vestibulaire dans le conduit auditif interne. Le vestibule comprend trois canaux semi-circulaires, sensibles au déplacement dans les trois plans de l'espace. Ces canaux aboutissent dans l'utricule qui communique avec le saccule. L'utricule et le saccule contrôlent la position de la tête, respectivement dans les plans horizontal et vertical.

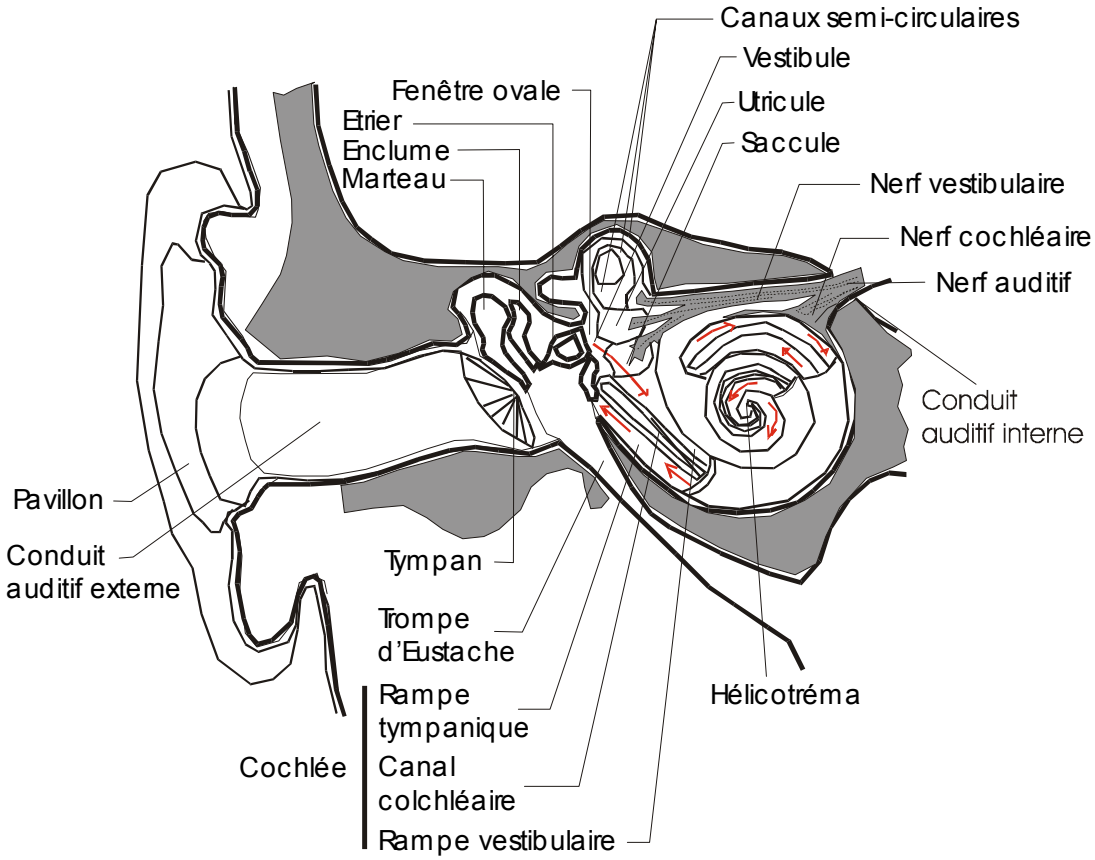

Figure 2.14 : L'appareil auditif (figure adaptée de [2-16]).

2.4.2 Perception auditive

On entend un son lorsque des vibrations de l'air ambiant atteignent notre tympan et le mettent en mouvement dans des conditions d'amplitude et de fréquence telles, que cette stimulation mécanique transmise par l'oreille moyenne à l'oreille interne, y provoque un phénomène bio-électrique. Commence alors le traitement de l'information produite par ce phénomène bio-électrique, traitement qui se poursuit à travers différents relais jusqu'au cortex cérébral. Il en résulte la perception du son.

Par sa géométrie et par la nature de ses parois, l'oreille externe ne transmet pas uniformément toutes les fréquences à l'oreille moyenne. Du fait de sa forme anatomique, le pavillon favorise et amplifie de quelques décibels les fréquences proches des fréquences allant de 155 à 7000 Hz, avec une résonance à 2000 Hz [2-17]. Le premier mode de résonance du conduit auditif est situé près de 3000 Hz, ce qui accroît la sensibilité du système auditif dans cette gamme de fréquences [2-1]. Au niveau de l'oreille moyenne, les vibrations du tympan sont retransmises à la chaîne des osselets. Celle-ci les amplifie et les communique aux structures de l'oreille interne. La meilleure communication à l'oreille interne se situe entre 1000 et 2000 Hz.

La théorie de l'onde se propageant le long de la cochlée, de la base vers l'apex, a été établie par Hurst en 1894 [2-18] : une onde transversale, perpendiculaire au plan de la membrane basilaire, change d'amplitude au cours de son déplacement, augmentant peu à peu pour atteindre son maximum avant de décroître rapidement. Une onde de fréquence basse atteindra son amplitude maximale près de l'apex; en revanche, une onde de plus haute fréquence trouvera son point d'amplitude maximale près de la base. Autrement dit, au fur et à mesure que l'on monte dans l'échelle des fréquences, la position de ce maximum d'amplitude se déplace de l'apex vers la base de la membrane. Dans la région médiane, la fréquence préférentielle est de 1600 Hz. Ce phénomène est dû à la variation de la rigidité des fibres basilaires de la base au sommet. Ainsi, la membrane basilaire agit en analyseur fréquentiel mécanique représenté par une série de filtres passe-bande spatialement répartis. En outre, cette membrane est capable d'effectuer une analyse spectrale divisant un son complexe en ses composantes fréquentielles. La transduction au niveau de la cochlée est illustrée à la Figure 2.15.

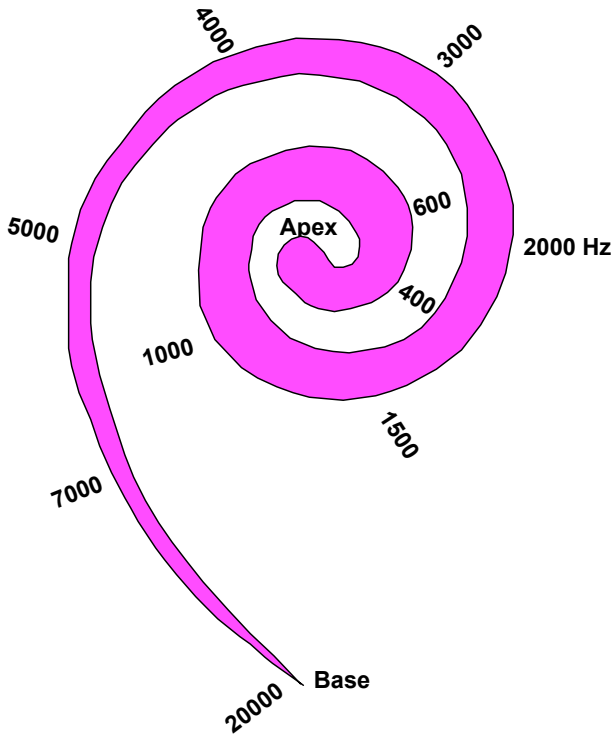

Figure 2.15 : Cochlée de la base à l'apex et correspondance spatiale avec la fréquence des vibrations sonores perçues (figure adaptée de [2-15]).

La perception auditive résulte du traitement de l'information sonore parvenue au cerveau. Comme introduit ci-dessus, le processus de transmission des informations sonores par l'oreille vers le cerveau est très complexe. Actuellement, la manière dont le cerveau traite ces informations sonores n'est pas résolue. Par conséquent, la perception auditive est généralement étudiée dans le cadre de la psycho-acoustique.

La psycho-acoustique a pour objet l'étude expérimentale des relations quantitatives entre les stimulis acoustiques mesurables physiquement et les réponses de l'ensemble du système auditif : sensations et perceptions auditives. Les aspects perceptuels de l'oreille sont plus évidents s'ils sont représentés dans le domaine fréquentiel et non pas dans le domaine temporel. En effet, l'oreille procède à une analyse spectrale de l'onde acoustique reçue.

L'une des premières observations de la psycho-acoustique est qu'il n'y a pas de relation bi-univoque entre les paramètres physiques des sons et les sensations qu'ils produisent. Par exemple, si une augmentation de la fréquence d'une vibration sinusoïdale entraîne principalement une augmentation de la hauteur du son perçu, elle peut aussi donner lieu à une variation de l'intensité perçue [2-4].

L'intensité minimale qui permet la perception d'un son à une fréquence donnée est appelée seuil d'audition. Ce seuil augmente considérablement pour les fréquences inférieures à 1 kHz ainsi que pour celles supérieures à 5 kHz. Le seuil de douleur désigne, pour une fréquence donnée, l'intensité maximale

au-delà de laquelle l'écoute devient douloureuse. Le champ auditif se trouve entre le seuil d'audition et le seuil de douleur. Le seuil d'audition, le seuil de douleur ainsi que le champ auditif (aire audible) sont illustrés à la Figure 2.16 pour un cas typique.

Le traitement par la cochlée des vibrations de la membrane basilaire est fortement non-linéaire, et par conséquent, la perception de l'énergie d'un son à une fréquence donnée est fortement dépendante de l'énergie des sons aux autres fréquences. En présence d'un son, appelé le masquant, à une fréquence f_m, l'oreille ne perçoit pas les sons de plus faible puissance, situés à proximité de f_m dans le domaine spectral. On appelle cette propriété l'effet de masque de l'oreille. Le seuil d'audition des sons masqués est caractérisé par une courbe dite courbe de masquage. Si lors de la synthèse de la parole, un bruit de codage est généré au-dessous de cette courbe, il n'est pas perçu par l'auditeur. La Figure 2.17 représente le seuil d'audition de l'oreille humaine dans un milieu silencieux (courbe A) ainsi que le seuil d'audition d'un son quelconque (courbe B) en présence d'un signal d'assez forte amplitude, dont les fréquences sont proches de 500 Hz.

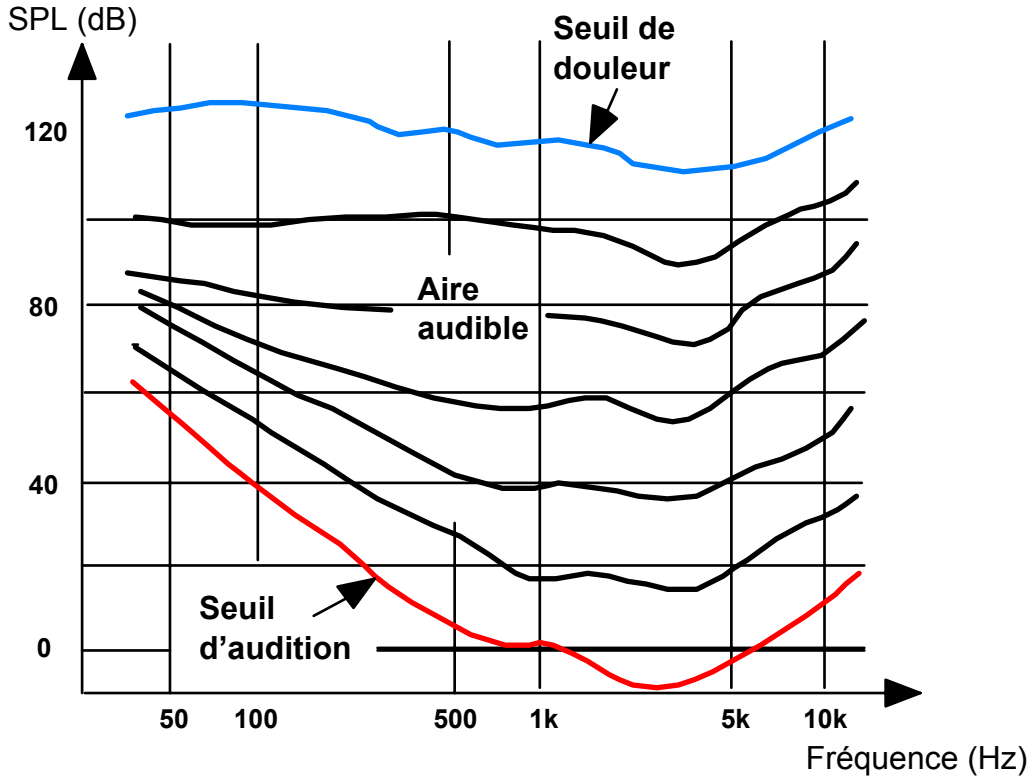

Figure 2.16 : Seuil d'audition, seuil de douleur, aire audible de l'oreille et courbes d'égale sensation sonore en écoute binaurale et en fonction du "Sound Pressure Level" SPL (cas typique, figure reprise de [2-19]).

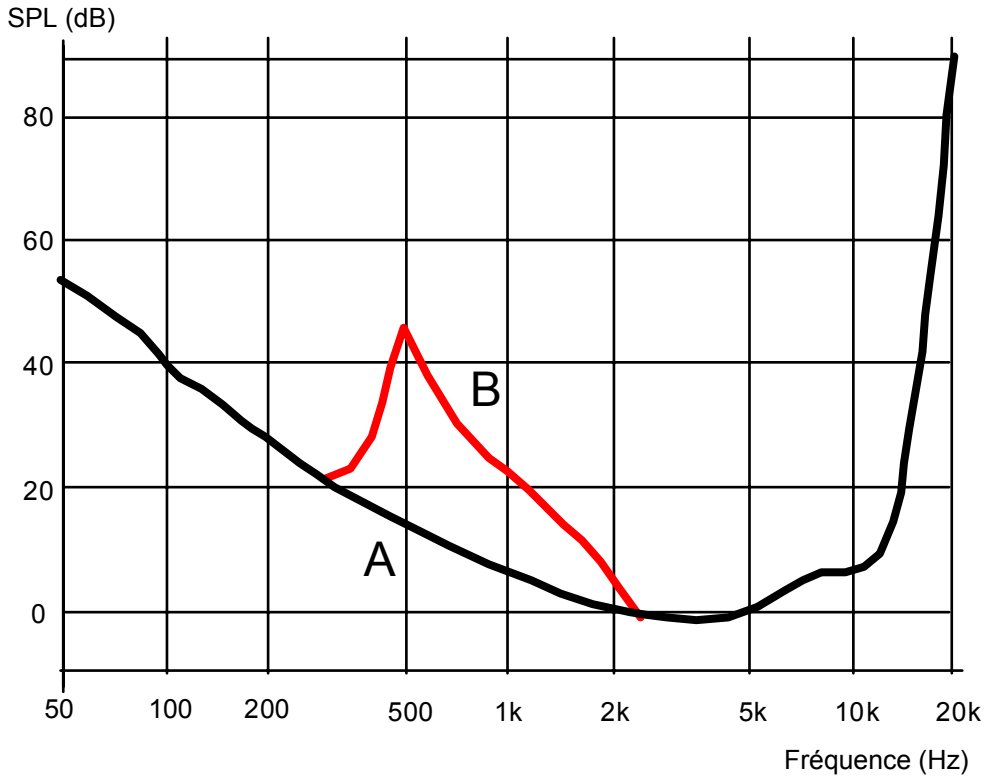

Figure 2.17 : Seuil d'audition de l'oreille humaine (cas typique) dans un milieu silencieux (courbe A) et en présence d'un signal de forte amplitude (70 dB SPL), dont les fréquences s'étalent autour de 500 Hz (courbe B) (figure reprise de [2-20]).

La propriété de masquage fréquentiel est exploitée pour améliorer la qualité de codage LPAS. Il faut porter une attention particulière aux amplitudes et fréquences des formants, qui jouent un rôle important dans la perception. Les voyelles se distinguent particulièrement par le fait qu'elles présentent trois fréquences formantiques principales. Les consonnes se différentient par les transitions de leurs formants dans le temps. Aussi bien pour les voyelles que pour les consonnes, les vallées entre les formants sont perceptuellement moins importantes que les formants.

En 1979, Atal et Schroeder [2-5] ont proposé de mettre en forme le bruit de quantification dans un codeur LPAS, à l'aide du filtre de pondération perceptuelle $W(z)$, introduit en Sous-section 2.3.2. Ce filtre permet d'utiliser une mesure de distorsion pondérée selon les propriétés de perception de l'oreille, afin de sélectionner dans les différents dictionnaires d'un codeur LPAS, les excitations permettant la meilleure reconstruction du signal de parole. En exploitant l'effet de masque, on accepte plus d'erreur dans les pics spectraux que dans les vallées. Comme le filtre de synthèse LPC décrit l'enveloppe du signal de parole, et donc ses pics et ses vallées, on peut utiliser un filtre $W(z)$, basé sur $H_p(z)$, de la forme suivante :

$$W(z) = \frac{A(z/\gamma_1)}{A(z/\gamma_2)}, \text{ avec}$$

$$A(z) = 1 + \sum_{k=1}^{p} a(k) z^{-k}, \text{ et} \qquad (2.20)$$

$$1 \geq \gamma_1 \geq \gamma_2 \geq 0.$$

Notons que si γ_1 vaut 1.0, alors le produit $W(z) \cdot H_p(z)$ se simplifie dans les formules de calcul de l'excitation optimale.

La Figure 2.18 montre la fonction de transfert du filtre $H_p(z)$ pour le signal de parole voisé et illustré à la Figure 2.1, ainsi que la fonction de transfert du filtre $W(z)$, dans le cas où $\gamma_1 = 1.0; \gamma_2 = 0.8$.

L'utilisation de la pondération perceptuelle peut également être utilisée au niveau du décodeur sous forme de post-filtrage. Le post-filtrage est réalisé pour améliorer la qualité de la parole reconstruite. Ceci est traité à la Section 3.8.

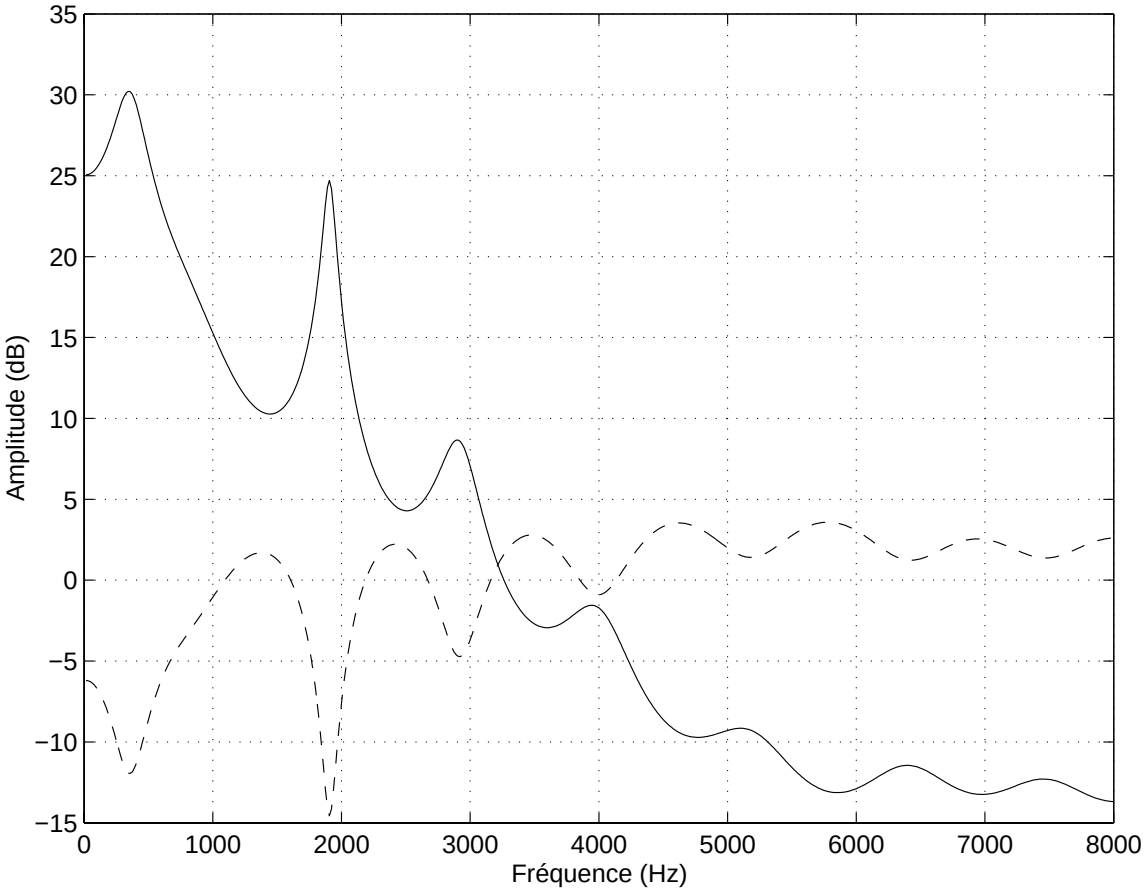

Figure 2.18 : Filtre de pondération perceptuelle $W(z)$ en trait discontinu comparé au filtre de synthèse LPC $H_p(z)$, en trait plein.

2.5 Limites du codage LPAS

Le codage LPAS permet de modéliser la production du signal de parole. Cependant ce modèle a des limites. Une partie d'entre elles est énumérée ci-dessous :

1. On peut observer que l'enveloppe spectrale du signal, et en particulier celle d'un son voisé, peut contenir de nombreuses résonances formantiques. Or, un filtre de synthèse LPC d'ordre limité p, ne peut modéliser qu'un nombre restreint de résonances formantiques, parfois inférieur au nombre effectif.

2. En utilisant un filtre de synthèse LPC, tel qu'il est donné par l'équation (2.4), on estime que le système de génération de la parole est un filtre tous-pôles. Cette estimation est inexacte si en plus des pôles on a quelques zéros dans la fonction de transfert du signal. Ceci est effectivement le cas lorsque le son émis est un son fricatif ou un son nasal. La modélisation d'un système ayant des pôles et des zéros pose un grand problème puisqu'il demande de résoudre des équations non-linéaires impliquant une complexité de calcul prohibitive.

3. Toute l'information spectrale d'un signal périodique est contenue dans ses harmoniques. Lorsque le signal de parole voisé a une fréquence fondamentale élevée, l'estimation de l'enveloppe spectrale du signal par analyse LPC peut poser problème. En effet, dans ce cas l'espace entre deux harmoniques est trop grand pour pourvoir une représentation adéquate de l'enveloppe spectrale par analyse LPC. Une solution à ce problème est l'utilisation d'une fenêtre dite "décalée". Cette solution est traitée à la Sous-section 3.3.1.

4. L'enveloppe spectrale du signal de parole a une pente spectrale (tilt). Les composantes du spectre de basse-fréquence ont une amplitude plus grande que celles de haute-fréquence. Il faut donc une grande précision de calcul pour décrire les formes spectrales sur tout le spectre. Ce problème est d'autant plus important pour un signal en bande élargie. Une solution est proposée à la Sous-section 2.5.1.

Notons encore que pour bien encoder les fricatives, il faudrait les échantillonner à 20 kHz. Or, ici on se contente d'un échantillonnage à 16 kHz.

2.5.1 Pré-accentuation

Pour réduire la dynamique du spectre du signal de parole, on peut utiliser une pré-accentuation du signal. La pré-accentuation consiste à faire passer le signal d'entrée dans un filtre de transmittance $H_{HP}(z)$, tel que :

$$H_{HP}(z) = 1 - \mu z^{-1}, \tag{2.21}$$

où le coefficient μ est compris entre 0 et 1.0. Ceci a pour effet d'accentuer l'amplitude des composantes en haute fréquence du spectre. Ainsi, celles-ci sont mieux prises en compte lors de la quantification par minimisation de l'erreur carrée. La Figure 2.19 illustre la réponse en fréquences du filtre $H_{HP}(z)$ pour diverses valeurs de μ. La valeur de μ devrait être calculée pour chaque trame de signal. Cependant, comme cette valeur n'est pas critique, on peut utiliser une valeur fixe qui est une estimation de celle-ci.

Pour éliminer l'effet de la pré-accentuation sur le signal de sortie, on utilise à la sortie du décodeur un filtre de désaccentuation de la forme : $1/H_{HP}(z)$.

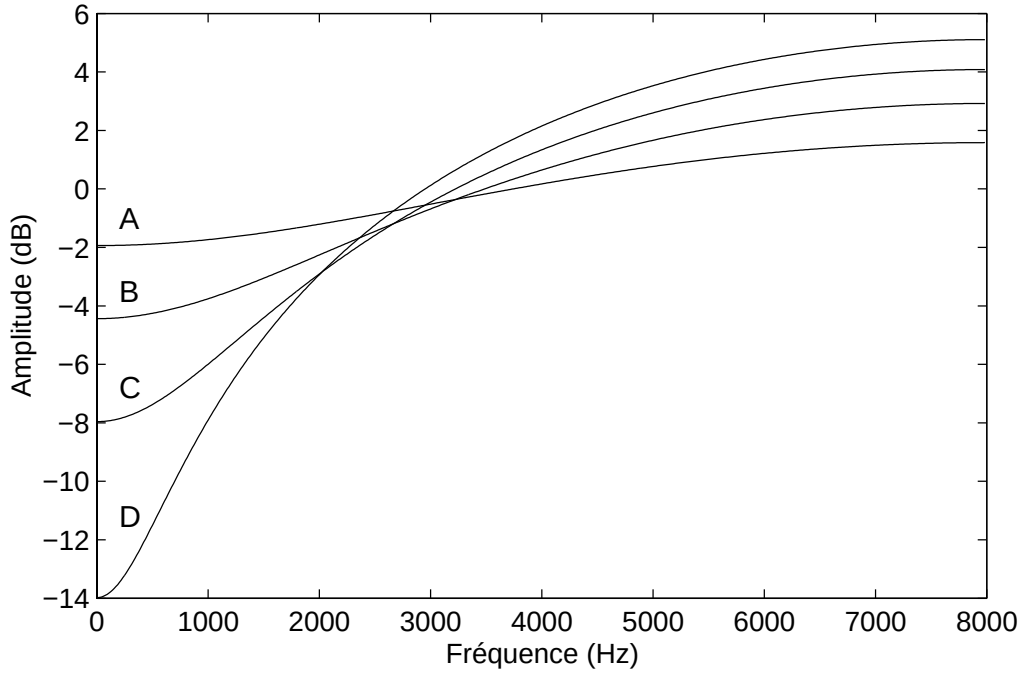

Figure 2.19 : Réponse en fréquences du filtre de pré-accentuation, $H_{HP}(z)$, pour $\mu = 0.2, 0.4, 0.6$ et 0.8 (courbes A, B, C et respectivement D), avec F_S=16 kHz.

2.6 Critères pour l'évaluation des performances d'un codeur de parole [2-21]

La mesure des performances d'un codeur de parole est une mesure critique à mettre en oeuvre. Une mesure objective standard de la qualité d'un codage est donnée par le rapport de la variance du signal d'entrée σ_s^2 et de la variance de l'erreur d'encodage σ_e^2. L'erreur d'encodage est la différence entre le signal d'entrée et le signal reconstruit. Le rapport précité est normalement appelé le "rapport signal sur bruit" (SNR); il se mesure en décibels et vaut :

$$SNR = 10\log_{10}(\sigma_s^2 / \sigma_e^2) \ \text{[dB]}. \qquad (2.22)$$

Une mesure plus adéquate, tenant compte du caractère non-stationnaire de la parole est définie en mesurant la moyenne des SNR sur des segments de parole.

Les codeurs LPAS ne sont pas des codeurs d'ondes, mais se basent sur des caractéristiques de la perception. Les mesures objectives ne donnent pas toujours l'information correcte sur la qualité du codage car elles ne tiennent pas compte des propriétés perceptuelles de l'oreille. De ce fait, les performances des codeurs LPAS ne peuvent pas être mesurées par le SNR, qui dans ce cas ne peut donner qu'une indication. Pour ce type de codeur, une mesure perceptuelle est plus adéquate.

Afin de pouvoir comparer différents systèmes de codage, des tests subjectifs ont été développés à partir de tests d'écoute. La présentation dite en "double aveugle avec référence cachée" est utilisée ici [2-22] et [2-23]. L'auditeur entend trois séquences. La première est toujours la séquence originale. Puis viennent "à l'aveugle", soit la séquence encodée suivie de l'originale, ou inversement. Par écoutes successives, l'auditeur doit comparer la deuxième et la troisième séquence à la première et les noter en fonction des dégradations qu'il a constatées. Les notes vont de 1.0 à 5.0. La note 5.0 est celle de la première séquence entendue, c'est-à-dire celle de l'originale. Ici, l'échelle des notations est appelée DMOS (Degradation Mean Opinion Score) et est spécifiée dans le Tableau 2-1.

Note	***Evaluation des dégradations***
5.0	Non perceptibles
4.0	Perceptibles mais non gênantes
3.0	Légèrement gênantes
2.0	Gênantes
1.0	Très gênantes

Tableau 2-1 : Echelle de notation à 5 notes (DMOS) pour l'évaluation des dégradations engendrées par l'encodage du signal.

2.7 Complexité des algorithmes [2-4]

Le codage de la parole pour la téléphonie mobile, pour la transmission via le protocole Internet ou pour la vidéo-conférence, implique en général un traitement en temps réel par des processeurs spécialisés. Le critère usuel pour estimer la complexité est le nombre d'opérations (additions, multiplications, etc.) par seconde, ainsi que la quantité de ressources requises. Ici, pour déterminer la complexité du codeur développé, on calculera pour chaque bloc fonctionnel de l'algorithme développé, le nombre d'opérations nécessaires, ainsi que la taille totale de la mémoire.

2.8 Résumé du chapitre et conclusions

Ce chapitre a donné une brève introduction du domaine du traitement de la parole. Différents concepts et définitions de base, utilisés dans le cadre du traitement du signal de parole pour la téléphonie mobile, ont été décrits. L'accent a été mis sur le codage de la parole pour la bande élargie. L'importance du codage LPAS et en particulier du codage CELP a été démontrée.

Les concepts théoriques plus spécifiques sont introduits au Chapitre 3.

2.9 Références

[2-1] R. Boite, H. Bourlard, T. Dutoit, J. Hancq et H. Leich, "Introduction", Chapitre 1, dans *Traitement de la parole*, pp. 1-26, publié par les Presses polytechniques et universitaires romandes, 2000.

[2-2] Calliope, "Introduction", Chapitre I, dans *La parole et son traitement automatique*, pp. 2-15, édité par J.P. Tubach, Masson et CNET-ENST, Paris, 1989.

[2-3] L. R. Rabiner et R. W. Schafer, "Digital models for the speech signal", Chapter 3, dans *Digital processing of speech signals*, pp. 38-115, Prentice-Hall Signal Processing Series, Alan V. Oppenheim Series Editor, 1978.

[2-4] Calliope, "Perception auditive et perception de la parole", Chapitre V, dans *La parole et son traitement automatique*, pp. 147-214, édité par J.P. Tubach, Masson et CNET-ENST, Paris, 1989.

[2-5] B. S. Atal et M. R. Schroeder, "Predictive coding of speech signals and subjective error criteria", dans *IEEE Transactions on acoustic, speech and signal processing*, Vol. ASSP-27, n°3, pp. 247-257, 1979.

[2-6] M. Bouraoui, "Les codeurs à analyse-par-sythèse : LPAS", Chapitre 1, dans *Elaboration et implantation sur DSP d'un codeur bas débit de la parole de faible complexité. Apport des codes correcteurs parfaits*, pp. 11-27, Thèse présentée à l'Institut National Polytechnique de Grenoble, Mars, 1992.

[2-7] S. Grassi, "Digital speech processing", Chapter 2, dans *Optimized implementation of speech processing algorithms*, pp. 7-32, Thèse éditée par la Faculté des Sciences de l'Université de Neuchâtel, Neuchâtel, 1998.

[2-8] A. Papoulis, "Mean square estimation", Chapter 13, dans *Probability, random variables, and stochastic processes,* 2nd ed., pp. 407-479, McGraw-Hill International Editions, 1984.

[2-9] J. Proakis et D. Manolakis, "Linear prediction and optimum linear filters", in *Digital signal processing. Principles, algorithms, and applications,* 3rd ed., pp. 852-895, Prentice Hall International Editions, Inc., 1996.

[2-10] A. Kondoz, "LPC parameter quantization using LSFs", Chapter 4, dans *Digital speech. Coding for low bit rate communication systems,* édité par John Wiley & Sons, Chichester, Grande Bretagne, 1994.

[2-11] B. Atal et J. Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit rates", dans *Proc. IEEE Int. conf. acoustic, speech, signal processing 1982, ICASSP-82,* Vol. 1, pp. 614-617, Paris, France, 1982.

[2-12] B. Atal et M. Schroeder, "Stochastic coding of speech signals at very low bit rates", dans *Proceedings of International conference on communications*, pp. 1610-1613, Mai 1984.

[2-13] A. Gersho, "Advances in speech and audio compression", dans *Proceedings of the IEEE,* Vol. 82, N° 6, pp. 900-918, 1994.

[2-14] J.-P. Adoul, P. Mabilleau, M. Delprat, et S. Morissette, "Fast CELP coding based on albebraic codes", dans *Proc. IEEE Int. conf. acoustic, speech, signal processing 1987, ICASSP-87,* pp.1957-1960, Avril, 1987.

[2-15] http://pst.chez.tiscali.fr/svtiufm/sens.htm#ouie (8 Août 2002).

[2-16] http://www.chez.com/naton/anatomie.htm (8 Août 2002).

[2-17] http://www.centre-audition.com/Anat.html (21 Août 2002).

[2-18] http://mediatheque.ircam.fr/articles/textes/Hammountene93a/ (8 Août 2002).

[2-19] http://www.inrp.fr/Acces/JIPSP/phymus/m_lexiq/lexbc1.htm (8 Août 2002).

[2-20] http://fr.audiofanzine.com/apprendre/dossiers/?idossier=20&page=3 (8 Août 2002).

[2-21] R. Boite, H. Bourlard, T. Dutoit, J. Hancq et H. Leich, "Codage", Chapitre 4, dans *Traitement de la parole*, publié par les Presses polytechniques et universitaires romandes, Lausanne, 2000.

[2-22] B. Gosselin, "Représentation de l'information", Chapitre 1, dans *Codage de l'information, représentation de l'information et quantification des signaux (Notes de cours)*, Faculté Polytehcnique de Mons, 2000, en http://www.tcts.fpms.ac.be/cours/1005-08/codage/repinfo.pdf (14 Fév. 2003).

[2-23] T. Grusec, L. Thibault et R. J. Beaton, "Sensitive methodologies for the subjective evaluation of high quality coding systems", dans *Proc. AES UK DSP Conf.*, pp. 62-76, Londres, Sept. 1992.

Chapitre 3
Codeur CS-ACELP

3.1 Introduction

Ce chapitre décrit un algorithme général pour le codage de la parole au moyen de la prédiction linéaire à excitation par séquences codées à structure algébrique conjuguée. Cet algorithme est appelé CS-ACELP (Conjugate-Structure Algebraic-Code-Excited Linear-Prediction). La terminologie CS, ou structure conjuguée, se réfère à l'utilisation de deux vecteurs d'excitation : l'un est extrait du dictionnaire adaptatif, l'autre du dictionnaire innovateur de type algébrique. Les concepts théoriques présentés au cours de ce chapitre vont permettre de décrire le codeur propriétaire, développé dans le cadre de ce travail de thèse.

Ce chapitre est articulé comme suit. La Section 3.2 donne une description générale de l'algorithme CS-ACELP. Les Sections 3.3, 3.4 et 3.5 décrivent les fondements théoriques des trois algorithmes principaux utilisés par l'encodeur. Il s'agit respectivement de l'analyse LPC, de l'extraction de l'excitation adaptative et de l'extraction de l'excitation innovatrice de type algébrique. La Section 3.6 décrit la quantification conjointe des gains par quantification vectorielle. Les Sections 3.7 et 3.8 introduisent la remise à jour d'une trame à l'autre des filtres de l'encodeur, et le post-filtrage utilisé au niveau du décodeur. Finalement la Section 3.9 traite le délai de traitement d'un encodeur.

Les modifications de l'algorithme général présenté ici, apportées dans le cadre de ce travail de thèse ayant pour application un codeur en bande élargie, sont présentées au Chapitre 5.

3.2 Description générale du codeur

Le codeur CS-ACELP est fondé sur la technique de codage prédictif d'analyse par synthèse, et plus particulièrement sur le codage prédictif linéaire à excitation par code (CELP). Ce codeur opère sur des trames de parole, considérées comme spectralement stationnaires, dont la durée varie d'une implantation à l'autre entre 10 et 30 ms. Pour un échantillonnage à 16 kHz, cela correspond à des trames de 160 à 480 échantillons. Pour chaque trame de signal à encoder, le signal est analysé et les paramètres nécessaires à sa transmission sont extraits. Ces paramètres sont :

- Les coefficients du filtre de prédiction linéaire (LPC);
- L'index et le gain du dictionnaire adaptatif;
- Les positions et les amplitudes des impulsions du vecteur d'excitation algébrique, ainsi que le gain correspondant.

Pour une bonne qualité de signal reconstruit, les paramètres relatifs aux dictionnaires adaptatif et algébrique sont généralement extraits plus fréquemment que les coefficients LPC, à savoir environ 2 à 4 fois par trame, en fonction de la longueur de celle-ci. Une trame de parole est donc divisée en différentes sous-trames non-recouvrantes.

Les paramètres cités ci-dessus sont extraits dans l'encodeur et transmis au décodeur. Le décodeur les utilise pour reconstituer l'excitation totale et fixer les paramètres du filtre de synthèse LPC, $H_p(z)$ donné par l'équation (2.4). L'excitation est filtrée par $H_p(z)$ pour produire la parole reconstruite. L'ordre de ce filtre dépend du signal. Il vaut généralement 16 pour la parole en bande élargie et 10 pour la parole en bande étroite. Le signal vocal reconstruit est souvent amélioré par un post-filtrage qui rehausse la qualité auditive du signal de sortie.

Les Sous-sections 3.2.1 et 3.2.2 décrivent les principes des algorithmes d'encodage et de décodage d'un codeur CS-ACELP.

3.2.1 Principe de l'encodeur CS-ACELP

Les principaux blocs fonctionnels de l'encodeur CS-ACELP sont représentés à la Figure 3.1. Si l'on compare cet encodeur au codeur LPAS illustré à la Figure 2.10, on constate une implantation différente du filtre de pondération perceptuel et du filtre de synthèse LPC. Ces différences n'apportent aucune modification fondamentale au fonctionnement de l'encodeur. L'implantation

selon la Figure 3.1 est utile pour l'extraction des paramètres et pour faciliter la suite de la discussion.

A l'entrée de l'encodeur CS-ACELP, le signal de parole $s(n)$ est passé dans un filtre passe-haut, puis dans un bloc de normalisation. Il en sort un nouveau signal $s_{pt}(n)$. Ces deux opérations constituent le bloc de pré-traitement du signal. Le filtre passe-haut élimine les composantes parasites en basse fréquence du signal. Le bloc de normalisation évite un dépassement de la dynamique maximale du signal, en cas de traitement en virgule fixe. Le pré-traitement du signal peut contenir également une pré-accentuation du spectre du signal, permettant d'en réduire la dynamique (cf. Sous-section 2.5.1).

A la sortie du bloc de pré-traitement, le signal $s_{pt}(n)$ est utilisé pour l'analyse prédictive linéaire, permettant d'extraire les coefficients de prédiction linéaire LPC. Cette analyse s'effectue pour chaque trame de signal. Afin de quantifier les coefficients LPC de façon efficace et stable, ceux-ci sont transformés en une représentation alternative, plus adéquate, telle que les coefficients appelés "paires de lignes spectrales" (LSP : line spectrum pairs) [3-1]. Les coefficients LSP sont quantifiés, soit individuellement par une quantification scalaire, soit de manière groupée par une quantification vectorielle (VQ). Pour réduire le débit de transmission, la quantification vectorielle est plus adaptée. Diverses techniques de quantification vectorielle sont applicables, telles que la quantification prédictive, la quantification par étage (MSVQ) ou encore la quantification séparée en sous-vecteurs (SVQ). Une combinaison de plusieurs techniques est également possible.

Les paramètres LSP codés sont transmis au décodeur. Au niveau de l'encodeur, ils sont transformés en LPC quantifiés pour l'implantation du filtre de synthèse à court-terme. Le signal d'excitation est alors choisi au moyen d'une procédure de recherche basée sur l'analyse par synthèse. Cette procédure est réalisée pour chacune des sous-trames de signal, puisque l'excitation varie plus rapidement que l'enveloppe spectrale du signal. Pour une transition plus douce du filtre de synthèse entre une trame et l'autre, les coefficients LPC sont modifiés pour chacune des sous-trames de signal à l'aide d'une interpolation. Pour assurer la stabilité du filtre de synthèse, on interpole les LSP quantifiés et non pas les LPC.

On extrait les excitations, en minimisant l'erreur pondérée perceptuellement entre le signal original et le signal reconstruit. L'erreur est donc passée dans un filtre de pondération perceptuelle $W(z)$. Les coefficients de ce filtre sont déduits des coefficients LPC interpolés mais non-quantifiés, puisqu'ainsi ils sont plus appropriés pour décrire le caractère perceptuel du signal original.

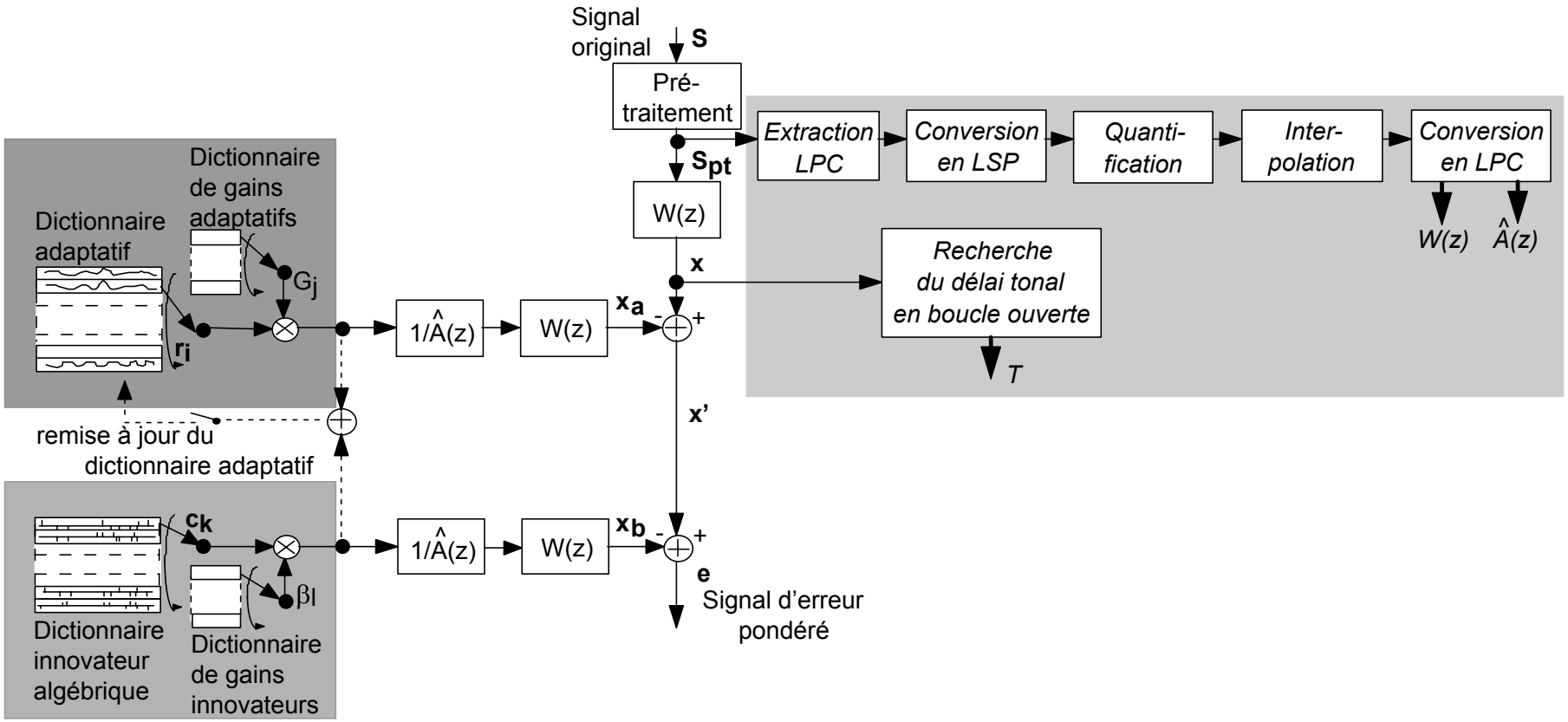

Figure 3.1 : Principe de l'encodeur CS-ACELP.

L'excitation adaptative contient la redondance à long-terme du signal (cf. Sous-section 2.3.1). Elle est extraite avant l'excitation innovatrice, qui théoriquement ne devrait plus contenir aucune redondance. Pour réduire la complexité de calcul relative à son extraction, il est possible d'estimer le délai tonal en boucle ouverte. Cette estimation se fait soit une fois par trame, soit une fois par sous-trame ou une fois toutes les deux ou trois sous-trames en fonction de la longueur de celles-ci et du type de signal traité. Rappelons que pour un codage de la parole en bande élargie, il est nécessaire de très bien encoder les basses fréquences. L'estimation en boucle ouverte[4] se base ici sur le signal cible $x(n)$, qui correspond au signal de parole pré-traité, puis passé dans le filtre de pondération perceptuelle $W(z)$.

La recherche en boucle fermée est réalisée en fonction du délai tonal T, estimé en boucle ouverte. Les excitations du dictionnaire adaptatif, correspondant à des délais tonals proches de T (compris dans l'intervalle $[T-\Delta T, T+\Delta T]$), sont filtrées par le filtre de synthèse pondéré $W(z)/\hat{A}(z)$. La sortie de ce filtre est appelée contribution adaptative $x_a(n)$. $1/\hat{A}(z)$ représente le filtre de synthèse, dont les coefficients LPC sont quantifiés. Le délai tonal entier T_1, compris dans l'intervalle $[T-\Delta T, T+\Delta T]$ et minimisant l'erreur pondérée perceptuellement (erreur quadratique), est retenu. Ici, l'erreur pondérée perceptuellement correspond à la différence entre les signaux $x(n)$ et $x_a(n)$. Pour une meilleure quantification du signal, on peut tester des délais tonals fractionnels T_1' proches de T_1. Pour ce faire, on utilise une interpolation linéaire.

Le délai tonal (entier ou fractionnel) est encodé et transmis au décodeur. Si la recherche en boucle ouverte du délai tonal T couvre plusieurs sous-trames, on peut recourir à l'encodage différentiel. Le délai tonal T_1' de la première sous-trame est encodé normalement. Pour la ou les sous-trame(s) suivante(s), on réalise un encodage différentiel par rapport à T_1. Ce procédé permet de réduire le débit de transmission.

Le signal cible $x(n)$ est mis à jour par soustraction de la contribution adaptative $x_a(n)$. Le nouveau signal cible ainsi obtenu, $x'(n)$, est utilisé pour explorer le dictionnaire innovateur (fixe) afin d'en extraire l'excitation innovatrice optimale. Les mots de code contenus dans le dictionnaire innovateur sont des vecteurs qui ne sont composés que de M éléments (valeurs) non-nuls. Ces éléments peuvent prendre la valeur +1 ou -1. Naturellement plus M est grand, plus la qualité du signal reconstruit est bonne, mais plus le débit et la complexité de calcul sont élevés. Les gains correspondant aux excitations adaptative et fixe sont quantifiés, soit indépendamment (dans ce cas le gain adaptatif est quantifié avant la mise à

[4] L'estimation du délai tonal en boucle ouverte se fait selon les équations (3.22) à (3.27), en remplaçant le signal résiduel de prédiction à court-terme par le signal cible $x(n)$.

jour du signal cible $x'(n)$), soit vectoriellement en minimisant l'erreur pondérée totale de reconstruction (erreur quadratique).

Finalement, les mémoires des filtres ainsi que le dictionnaire adaptatif sont remis à jour au moyen des signaux d'excitation extraits. Le dictionnaire adaptatif est remis à jour en sommant les deux excitations multipliées par leur gain respectif.

Les indices paramétriques correspondant aux LSP quantifiés, aux indices de dictionnaires et à leur gain respectif, sont transmis au décodeur sous la forme d'un flux binaire.

3.2.2 Principe du décodeur CS-ACELP

Le principe du décodeur CS-ACELP est représenté à la Figure 3.2 ci-dessous.

Les indices paramétriques reçus de l'encodeur sont décodés pour obtenir les paramètres de synthèse d'une trame de signal. Les coefficients LSP sont interpolés et convertis en coefficients LPC. Pour chacune des sous-trames, la reconstruction du signal se fait ainsi :

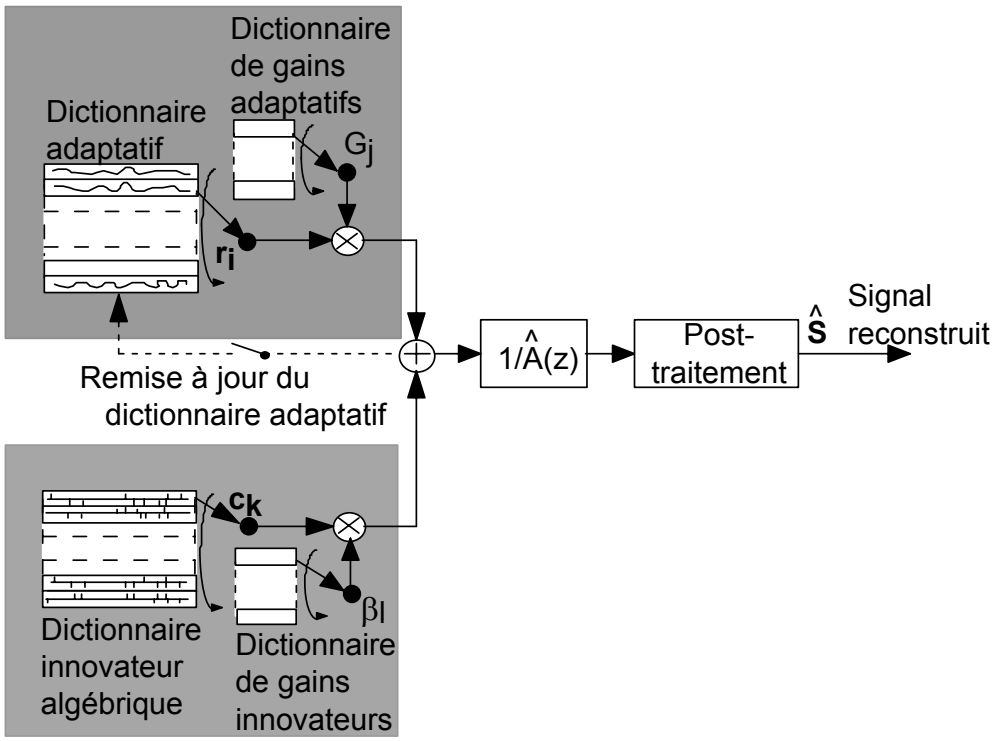

Figure 3.2 : Principe du décodeur CS-ACELP.

- L'excitation totale est construite par combinaison des excitations adaptative et fixe, extraites des dictionnaires, et pondérées par multiplication avec leur gain respectif;
- Le signal vocal est reconstruit par filtrage de l'excitation totale dans le filtre de synthèse à court-terme $1/\hat{A}(z)$;
- Le signal vocal reconstruit est envoyé dans un bloc de post-traitement. La sortie de ce bloc est le signal de parole reconstruit $\hat{s}(n)$.

Le bloc de post-traitement comprend généralement un post-filtrage adaptatif, utilisant la sortie du filtre de synthèse à court-terme, suivi d'un filtrage passe-haut. Si le pré-traitement au niveau de l'encodeur contient un bloc de pré-accentuation du spectre, alors le post-traitement contient un bloc de désaccentuation du spectre.

3.3 Analyse LPC

Dans les applications de traitement du signal de parole, le codage par prédiction linéaire (LPC) est couramment utilisé pour représenter l'enveloppe du spectre de puissance à court-terme du signal.

Si p est l'ordre de l'analyse LPC, l'échantillon $s(n)$ est prédit par une combinaison linéaire des p échantillons qui le précèdent. Soit le signal prédit $\hat{s}(n)$, et les coefficients LPC d'ordre p, $\{a_p(1),\ldots,a_p(p)\}$, alors :

$$\hat{s}(n) = -\sum_{k=1}^{p} a_p(k)\cdot s(n-k). \tag{3.1}$$

Le calcul des coefficients LPC a été présenté au point 2.3.3.1. La valeur $\hat{s}(n)$ soustraite de $s(n)$ donne le signal résiduel $x(n)$:

$$x(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{p} a_p(k)\cdot s(n-k). \tag{3.2}$$

La variance de $x(n)$ est réduite par rapport à celle de $s(n)$.

Si l'on considère la transformée en *z* de l'équation (3.2), on obtient :

$$X(z) = A_p(z)\cdot S(z), \tag{3.3}$$

où $X(z)$, $S(z)$ et $A_p(z)$ sont respectivement les transformées en *z* de l'erreur de prédiction (signal résiduel), du signal de parole, et du filtre d'analyse LPC. Le filtre $A_p(z)$ est donné par l'équation (2.5). Le filtre de synthèse à court-terme $H_p(z)$, inverse du filtre d'analyse, est donné par l'équation (2.4).

L'équation (3.3) est la base du modèle LPC. Inversement, le modèle de synthèse LPC consiste en une source d'excitation $X(z)$, entrant dans un filtre de mise en forme spectrale $H_p(z)$, résultant en le signal de parole synthétisé $S(z)$. On a :

$$S(z) = H_p(z) \cdot X(z). \tag{3.4}$$

La quantification de $X(z)$ et de $H_p(z)$ est choisie de sorte à rendre le signal $S(z)$ quantifié aussi proche que possible du signal orignal.

3.3.1 Expansion de la largeur de bande et fenêtre décalée [3-1]

Lorsque le signal de parole voisé a une fréquence fondamentale élevée, l'estimation de l'enveloppe spectrale du signal par analyse LPC peut poser problème. En effet, l'essentiel de l'information spectrale d'un signal de parole périodique est contenu dans ses harmoniques. Or, si le signal a une fréquence fondamentale très élevée, l'espace entre deux harmoniques est trop grand pour permettre une représentation adéquate de l'enveloppe spectrale par analyse LPC. Cette inadéquation entraîne un signal de parole synthétisé à consonance métallique. Pour résoudre ce problème on peut utiliser soit une expansion de la largeur de bande des formants, soit une fenêtre dite "décalée".

L'expansion de la largeur de bande des formants est utilisée pour rendre l'espace entre les harmoniques adéquat à la modélisation du signal. Dans la pratique, cela se fait par une multiplication des coefficients LPC $a_p(k)$, par un facteur d'écartement γ^k inférieur à 1.0. Soit B_i, la largeur de bande associée au i^{ème} pôle du filtre de synthèse LPC défini ainsi :

$$B_i = -\frac{F_s}{\pi} \cdot \ln(\tilde{r}_i), \tag{3.5}$$

où $\tilde{r}_i$ est le rayon du pôle considéré, et où F_s est la fréquence d'échantillonnage du signal. Une multiplication de ce rayon par γ étend la largeur de bande B_i associée à ce pôle à $B_i + \Delta B$, où

$$\Delta B = -\frac{F_s}{\pi} \cdot \ln(\gamma). \tag{3.6}$$

Le filtre de synthèse LPC résultant, $H'_p(z)$, est ainsi donné par :

$$H'_p(z) = H_p(z/\gamma) = \frac{1}{1 + \sum_{k=1}^{p} \left[a_p(k) \cdot \gamma^k \right] \cdot z^{-k}}. \tag{3.7}$$

L'utilisation d'une fenêtre dite "décalée" est une autre méthode pour résoudre le problème évoqué ci-dessus. Elle consiste à multiplier les coefficients d'autocorrélation r_k, donnés par l'équation (2.13), par une fenêtre généralement choisie de forme gaussienne. Ceci correspond à convoluer le spectre de puissance du signal avec une forme Gaussienne, afin d'obtenir l'expansion des formants du spectre du signal.

Notons que l'expansion de la largeur de bande diminue la sensibilité de l'enveloppe spectrale du signal reconstruit par rapport aux erreurs de quantification des paramètres LPC.

3.3.2 Généralités pour la quantification des coefficients LPC [3-1]

Il est important de quantifier les LPC, d'une part sans introduire une distorsion spectrale audible et d'autre part en prenant garde à ne pas rendre le filtre de synthèse $H_p(z)$ instable. Afin de quantifier les coefficients LPC de façon efficace et stable, ceux-ci sont transformés en une représentation alternative. La transformation des coefficients LPC en paires de lignes spectrales (LSP) est une solution adéquate et couramment utilisée dans le codage de la parole. Elle est introduite à la Sous-section 3.3.4.

Les coefficients LSP sont quantifiés, soit par une quantification scalaire, soit par une quantification vectorielle. Cette dernière permet de réduire le débit de transmission. Toutefois, avec une telle quantification il faut prendre garde à ne pas introduire une complexité de calcul trop importante. Pour cette raison, des méthodes de quantifications vectorielles sous-optimales, telles que la quantification par étage ou encore la quantification séparée en sous-vecteurs, sont utilisées.

La méthode usuelle permettant de mesurer les performances de la quantification des LPC est introduite à la Sous-section 3.3.3.

3.3.3 Mesure des performances de la quantification des LPC [3-1]

Les performances de la quantification des LPC sont usuellement évaluées en calculant la racine carrée de la distorsion spectrale SD_n de la nième trame du signal. Cette distorsion est donnée par l'équation ci-dessous :

$$SD_n = \sqrt{\frac{100}{(f_2 - f_1)} \int_{f_1}^{f_2} \left[\log_{10} \frac{S_n(f)}{Sq_n(f)} \right]^2 df} \quad [\text{dB}], \tag{3.8}$$

où $S_n(f)$ et $Sq_n(f)$ sont respectivement le spectre original et le spectre quantifié des filtres de synthèse LPC $A_n(z)$ et $\hat{A}_n(z)$. n est l'indice de la trame traitée. Ces spectres sont évalués à l'aide de la transformée de Fourier. Ils sont donnés par :

$$S_n(f) = \frac{1}{\left| A_n(z = e^{j\frac{2\pi f}{Fs}}) \right|^2}, \qquad Sq_n(f) = \frac{1}{\left| \hat{A}_n(z = e^{j\frac{2\pi f}{Fs}}) \right|^2}. \tag{3.9}$$

f est la fréquence en Hz et f_1 et f_2 sont les bornes fréquentielles de mesure. Celles-ci correspondent usuellement aux limites de la bande de fréquences du signal.

Pour mesurer les performances du quantificateur, on calcule la moyenne de la distorsion spectrale sur M trames consécutives du signal :

$$\overline{SD} = \frac{1}{M} \sum_{n=1}^{M} SD_n. \tag{3.10}$$

Pour le codage de la parole en bande étroite, un critère objectif bien établi permet déterminer si la quantification est transparente. On parle de transparence si la quantification n'entraîne aucune distorsion audible. Ce critère de transparence est spécifié ainsi [3-1] :

- La distorsion spectrale moyenne doit être inférieure à 1 dB;
- Aucune trame ne doit entraîner une distorsion spectrale supérieure à 4 dB;
- Le nombre de trames pouvant entraîner une distorsion spectrale comprise entre 2 et 4 dB doit être inférieur à 2 %.

Ce critère de transparence est suffisant mais pas nécessaire. Les trames dont la distorsion spectrale est supérieure à 2 dB sont qualifiées d'« outliers» (imposteurs).

Pour la parole en bande élargie, il n'existe aucun critère de transparence couramment accepté. En [3-2], Guibe et *al.* proposent d'utiliser le même critère que pour la bande étroite, en posant comme bornes d'intégration, 0 et 7000 Hz. En [3-3], Ferhaoui et *al.* proposent le critère suivant, avec ces mêmes bornes d'intégration :

- La distorsion spectrale moyenne doit être inférieure à 1.6 dB;
- Aucune trame ne doit entraîner une distorsion spectrale supérieure à 5 dB;
- Le nombre de trames pouvant entraîner une distorsion spectrale comprise entre 3 et 5 dB doit être inférieur à 2 %.

3.3.4 Représentation alternative des coefficients LPC [3-4]

Pour quantifier les coefficients LPC, de sorte à ne pas introduire de distorsion spectrale audible et en prenant garde à ne pas rendre le filtre de synthèse $H_p(z)$ instable, on recourt à une représentation alternative des coefficients LPC. Cette représentation doit avoir les diverses propriétés énumérées ci-après. La représentation alternative doit avoir un relation bijective avec les coefficients LPC. Elle doit également être spectralement moins sensible que la représentation LPC. En outre, elle doit assurer une vérification peu complexe de la stabilité du filtre de synthèse après quantification. Parmi les représentations alternatives des LPC, les plus appropriées sont les coefficients de réflexion (RC), de rapports de surface logarithmique (Log Area Ratio (LAR)), de sinus inverse (IS), de paires de lignes spectrales (LSP) [3-1] et de paires d'immittances spectrales (ISP) [3-5].

Les coefficients de réflexion[5] $\{k_m\}$, ont été introduits au point 2.3.3.1. Ces coefficients sont moins sensibles aux erreurs de quantification que les coefficients LPC. Ils sont limités en valeur absolue par un maximum de 1.0. Cette condition de limite est nécessaire et suffisante pour assurer la stabilité (non stricte) du filtre de synthèse LPC [3-1]. Bien que les coefficients RC soient plus adaptés à une quantification que les coefficients LPC, ils ont une sensibilité spectrale non uniforme. La sensibilité spectrale des coefficients RC augmente considérablement lorsque leur valeur absolue s'approche de 1.0. Ce problème peut être résolu en utilisant une transformation non-linéaire des coefficients RC proches de $|k_m| = 1.0$, conduisant aux coefficients LAR ou IS. Le majeur inconvénient de telles représentations est que la corrélation des paramètres LPC d'une trame à l'autre disparaît et ne peut être exploitée pour une quantification prédictive.

Les représentations des LPC en paires des lignes spectrales (LSP) et en paires d'immittances spectrales (ISP) possèdent toutes les propriétés énumérées au début de la présente sous-section. Ces représentations possèdent en outre de nombreuses propriétés additionnelles très intéressantes, permettant de réaliser un encodage efficace. Ainsi, la représentation LSP est très couramment utilisée dans le codage de la parole. Cette représentation, ainsi que ses propriétés sont décrites à la Sous-section 3.3.6.

[5] On peut montrer que si l'on modélise le conduit vocal par une succession de tubes de sections variables à travers lesquels se propage une onde de pression, alors chaque coefficient de réflexion correspond à la proportion de l'onde de pression réfléchie (par rapport à l'onde incidente) sur la jonction entre deux tubes successifs de sections différentes [3-6].

3.3.5 Interpolation des coefficients LPC

Pour des systèmes de codage de la parole, l'analyse LPC se fait généralement pour des trames de taille comprise entre 10 et 30 ms. Ce taux de remise à jour peut entraîner de brusques changements dans les paramètres des trames adjacentes et introduire des transitoires ou des petits bruits dans le signal de parole reconstruit. Pour résoudre ce problème, l'interpolation des paramètres LPC est couramment utilisée afin d'obtenir une variation plus lisse des filtres d'analyse et de synthèse LPC. Soit par exemple une trame de parole de 20 ms, et une sous-trame de 5 ms. On utilise l'interpolation des paramètres de prédiction linéaire entre trames adjacentes, afin d'obtenir un ensemble différent de paramètres pour chacune des sous-trames. Ainsi, l'interpolation permet de remettre à jours les paramètres du filtre de synthèse toutes les 5 ms, tout en ne les transmettant que toutes les 20 ms et ceci sans augmenter le débit.

Généralement une interpolation linéaire est utilisée. Pour assurer la stabilité du filtre de synthèse, (cf. Sous-section 3.3.4), l'interpolation ne se fait pas sur les LPC, mais sur leur représentation alternative [3-7]. En effet, l'interpolation sur des coefficients RC, LAR, IS, LSP ou ISP, produit toujours un filtre de synthèse stable, si les filtres basés sur les coefficients non-interpolés sont stables. Ainsi, la représentation utilisée pour la quantification est également retenue pour l'interpolation. Paliwal montre en [3-8] que les meilleures performances d'interpolation s'obtiennent avec les coefficients LSP. Paliwal ne considère pas les coefficients ISP.

Si $\mathbf{f}_n$ est le vecteur LSP quantifié de la trame présente et $\mathbf{f}_{n-1}$, le vecteur LSP quantifié de la trame passée, alors le vecteur LSP interpolé ($\mathbf{sf}_{4(n-1)+k}$) pour la sous-trame k de la trame n est donné par :

$$\mathbf{sf}_{4(n-1)+k} = \delta_k \mathbf{f}_{n-1} + (1-\delta_k)\mathbf{f}_n, \tag{3.11}$$

où δ_k est un nombre rationnel compris entre 0 et 1.0. δ_k diminue graduellement avec l'index de la sous-trame. Pour l'exemple cité ci-dessus, un bon choix de la valeur δ_k est 0.75, 0.5, 0.25 et 0 pour respectivement k=1, 2, 3, 4. En utilisant ces valeurs, les vecteurs de LSP interpolés valent :

$$\begin{aligned}
\mathbf{sf}_{4n-3} &= 0.75\mathbf{f}_{n-1} + 0.25\mathbf{f}_n, \\
\mathbf{sf}_{4n-2} &= 0.5\mathbf{f}_{n-1} + 0.5\mathbf{f}_n, \\
\mathbf{sf}_{4n-1} &= 0.25\mathbf{f}_{n-1} + 0.75\mathbf{f}_n, \\
\mathbf{sf}_{4n} &= \mathbf{f}_n.
\end{aligned} \tag{3.12}$$

Avec cette solution, les paramètres LSP calculés pour une trame sont attribués à la dernière sous-trame de celle-ci. Dans ce cas, il est bon d'utiliser une

fenêtre $w(n)$ (cf. équation (2.1)) dont le maximum se trouve sur la dernière sous-trame.

3.3.6 Paires de lignes spectrales (LSP)

La représentation des LPC sous forme de LSP a été introduite par Itakura en [3-9]. Dans le codage de la parole, cette représentation est largement utilisée pour ses intéressantes propriétés. En effet, les coefficients LSP ont une dynamique limitée, ils présentent des propriétés de corrélation inter- et intra-trame(s), et ils permettent une vérification simple de la stabilité du filtre de synthèse [3-10]. De plus, ces coefficients se prêtent à une interpolation, menant à une variation spectrale plus lisse d'une trame à l'autre du signal.

Les propriétés de corrélation inter- et intra-trame(s) sont utilisées pour réduire le débit d'encodage. Les propriétés inter-trames permettent d'utiliser une quantification prédictive. Les propriétés intra-trame, permettent d'exploiter efficacement la quantification vectorielle.

De plus, les paramètres LSP présentent d'autres propriétés intéressantes : ils représentent le signal dans le domaine fréquentiel, leur sensibilité spectrale est localisée et leur relation avec les pics de l'enveloppe spectrale (formants) est immédiate. En effet, plus les LSP consécutifs sont rapprochés, et plus la largeur de bande des pics formantiques qu'ils caractérisent est étroite.

La représentation LSP est utilisée dans presque tous les standards de codage de la parole pour la bande étroite, ainsi que dans de très nombreuses publications sur les codeurs CELP [3-4].

La représentation ISP est très proche de la représentation LSP. Elle est utilisée dans le nouveau standard WB-AMR de l'ETSI.

Le point 3.3.6.1 définit les paramètres LSP et ISP. Le point 3.3.6.3 décrit les propriétés des LSP énumérées ci-dessus.

3.3.6.1 Définition des paramètres LSP

Les coefficients LSP sont obtenus à partir du filtre d'analyse LPC, $A_p(z)$, donné par l'équation (2.5). Un polynôme symétrique $P(z)$ et un polynôme anti-symétrique $Q(z)$ sont formés en additionnant et respectivement en soustrayant à $A_p(z)$ la fonction, $z^{-(p+1)}A_p(z^{-1})$. Ainsi, si p est pair, $P(z)$ et $Q(z)$ ont respectivement un zéro en $z=-1$ et en $z=1$, et peuvent être exprimés ainsi :

$$P(z) = A_p(z) + z^{-(p+1)} A_p(z^{-1}) = (1 + z^{-1}) \cdot P'(z),$$
$$Q(z) = A_p(z) - z^{-(p+1)} A_p(z^{-1}) = (1 - z^{-1}) \cdot Q'(z). \quad (3.13)$$

$P'(z)$ et $Q'(z)$ sont des polynômes symétriques aux propriétés suivantes [3-11] :

- Si les racines $r_{a,i}$ de $A_p(z)$ sont dans le cercle unité ouvert, i.e. $|r_{a,i}| < 1.0$, alors les racines $r_{p,i}$ et $r_{q,i}$ de $P'(z)$ et $Q'(z)$ se trouvent sur le cercle unité, où elles sont entrelacées et ordonnées en commençant par une racine de $P'(z)$. Ainsi, si l'on exprime les racines de ces polynômes par leur position angulaire $\omega_{p,i}$ et $\omega_{q,i}$, on aura :

$$0 < \omega_{p,0} < \omega_{q,0} < \omega_{p,1} < \omega_{q,1} \ldots, < 2\pi. \quad (3.14)$$

- A l'inverse, si les racines de $P'(z)$ et $Q'(z)$ se trouvent sur le cercle unité et si elles satisfont l'équation (3.14), alors les racines de $A_p(z)$ sont dans le cercle unité ouvert et par conséquent le filtre de synthèse LPC correspondant est strictement stable.

La première propriété est qualifiée de "théorème d'analyse" ou propriété d'ordre. La seconde propriété est qualifiée de "théorème de synthèse". Elle est utilisée pour assurer la stabilité du filtre de synthèse LPC $H_p(z)$ quantifié.

Comme les racines de $P'(z)$ et $Q'(z)$ se trouvent sur le cercle unité, ces polynômes sont totalement définis par les positions angulaires $\omega_{p,i}$ et $\omega_{q,i}$ de leurs racines. De plus, comme leurs coefficients sont réels, leurs racines apparaissent par paires conjuguées complexes. Ainsi, seules les racines se trouvant sur la partie supérieure du cercle unité situé dans le plan z, sont nécessaires pour entièrement caractériser ces deux polynômes. Par conséquent :

les LSP se définissent comme les positions angulaires des racines de $P'(z)$ et $Q'(z)$ se trouvant sur le demi-cercle supérieur du plan z.

Pour la suite, les LSP seront dénotés par $\{\omega_i\}$ et $\{f_i\}$, dans le domaine des fréquences angulaires et respectivement dans celui des fréquences normalisées. $\{\omega_i\}$ et $\{f_i\}$ sont liés par la relation :

$$f_i = \omega_i / (2\pi). \quad (3.15)$$

Dans le domaine des fréquences normalisées, la propriété d'ordre s'exprime ainsi :

$$0 < f_1 < f_2 < \ldots < f_p < 1.0. \quad (3.16)$$

La Figure 2.7 illustre la position des zéros du filtre d'analyse LPC d'ordre 16, $A_{16}(z)$, pour un segment de 20 ms de la voyelle "a", illustrée à la Figure

2.1. Il s'agit là de parole en bande élargie. Les zéros des polynômes $P'(z)$ et $Q'(z)$ correspondants sont illustrés à la Figure 3.3. Pour ce même segment de parole, la Figure 3.4 illustre le spectre de puissance LPC et la position des LSP qui lui sont associés.

Le point 3.3.6.3 décrit les différentes propriétés des LSP énumérées ci-dessus.

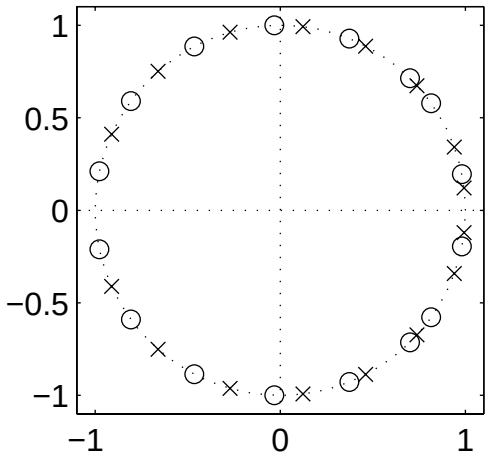

Figure 3.3 : Position des zéros des polynômes $P'(z)$ ('x') et $Q'(z)$ ('o') pour le segment de 20 ms, de la voyelle « a », illustrée à la Figure 2.1.

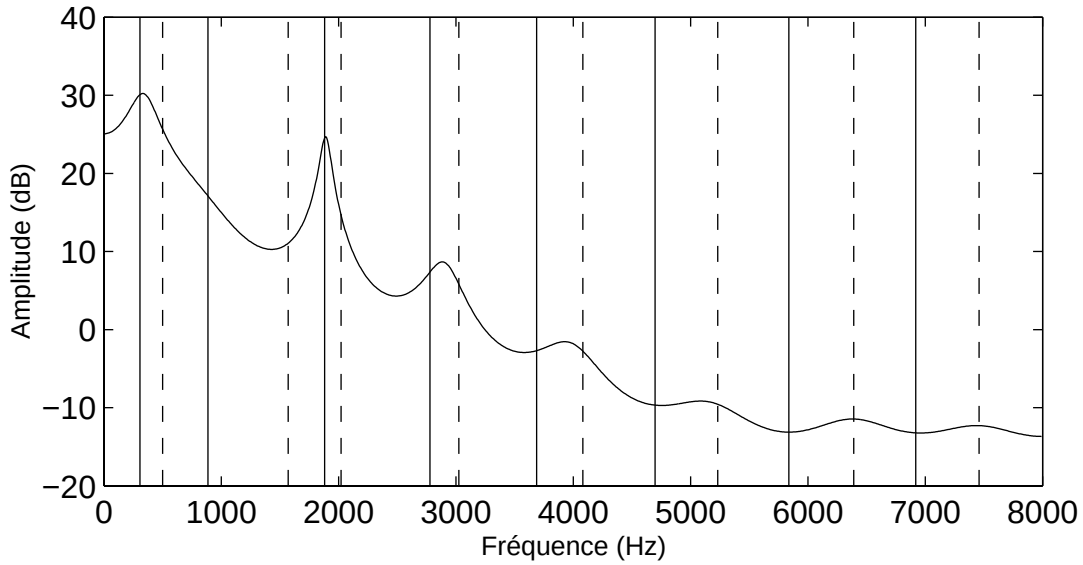

Figure 3.4 : Spectre de puissance LPC et position des paramètres LSP correspondants, pour le segment de 20 ms de la voyelle « a » illustrée à la Figure 2.1. Les traits pleins et les traits discontinus verticaux représentent respectivement les fréquences des LSP spécifiées par les zéros de $P'(z)$ et $Q'(z)$.

3.3.6.2 Définition des paramètres ISP

Les paires d'immittances spectrales ISP ont des propriétés de quantification à peine supérieures à celles des LSP, ainsi qu'une complexité de calcul légèrement réduite. La représentation ISP est utilisée dans le nouveau standard de l'ETSI pour la quantification de la parole en bande élargie.

Les coefficients ISP sont obtenus à partir du filtre d'analyse LPC, $A_p(z)$, donné par l'équation (2.5). Un polynôme symétrique $P(z)$ et un polynôme anti-symétrique $Q(z)$ sont formés en additionnant et respectivement en soustrayant à $A_p(z)$ la fonction $z^{-(p)}A_p(z^{-1})$. Ainsi, si p est pair, $P(z)$ et $Q(z)$ peuvent être exprimés ainsi :

$$\begin{aligned} P(z) &= A_p(z) + z^{-p}A_p(z^{-1}) = (1+a_p)\cdot P'(z), \\ Q(z) &= A_p(z) - z^{-p}A_p(z^{-1}) = (1-a_p)\cdot(1-z^{-2})\cdot Q'(z). \end{aligned} \tag{3.17}$$

Les polynômes $P'(z)$ et $Q'(z)$ sont des polynômes symétriques. Ils définissent les ISP et sont similaires à ceux définissant les LSP. Les positions angulaires des racines de $P'(z)$ et $Q'(z)$, sur la partie supérieure du cercle unité dans le plan z, sont les $p-1$ premiers coefficients ISP. Les $p-1$ premiers coefficients ISP d'un système d'ordre p sont les LSP d'ordre $p-1$. Le p^{ème} coefficient ISP est dérivé du dernier coefficient a_p, ce dernier étant égal au dernier coefficient de Parcor k_p $(-1.0 < k_p < 1.0)$ [3-12].

Les coefficients ISP partagent les propriétés essentielles des coefficients LSP et notamment la propriété d'ordre. Cependant, celles-ci ne sont pas décrites ici, puisque pour l'implantation du codeur en bande élargie faisant l'objet de ce rapport de thèse, les coefficients LSP ont été retenus. En effet, depuis la parution du premier article décrivant les ISP en 1993 [3-5], aucun standard ne les a utilisés, à l'exception du WB-AMR publié récemment.

3.3.6.3 Propriétés des paramètres LSP

Les différentes propriétés des paramètres LSP sont décrites ci-dessous. Ces propriétés montrent que la représentation LSP est idéale pour la quantification des coefficients LPC.

Représentation dans le domaine fréquentiel

La représentation LSP est une représentation fréquentielle. Lors de la quantification, cette propriété des coefficients LSP permet d'exploiter les propriétés spectrales de l'oreille humaine. Comme la résolution de l'oreille humaine est plus forte dans les basses fréquences, les LSP qui correspondent

aux basses fréquences sont généralement quantifiées avec plus de précision que les autres LSP. Pour ce faire, si l'on utilise une quantification vectorielle en sous-vecteurs, on attribuera plus de bits de quantification aux sous-vecteurs correspondants aux basses-fréquences.

Corrélation intra- et inter-trame(s)

Une propriété très importante des paramètres LSP est leur ordre naturel, introduit par l'équation (3.16). Cette propriété est utilisée aussi bien pour garantir la stabilité du filtre de synthèse LPC quantifié, que pour réaliser rapidement le calcul des LSP eux-mêmes (cf. Sous-section 3.3.8). Cette propriété montre également que les LSP de la même trame de signal sont corrélés. La forte corrélation entre LSP voisins est appelée corrélation intra-trame. Elle est traitée par Kondoz en [3-10]. La quantification vectorielle des LSP exploite cette propriété.

La configuration de la trachée vocale change lentement dans le temps. De ce fait, il existe également une forte corrélation entre les LSP de trames successives. On appelle cette corrélation, la corrélation inter-trames. On exploite cette propriété en réalisant une quantification des LSP par prédiction, où seuls les résidus de prédiction sont encodés.

La propriété d'ordre ainsi que les corrélations intra- et inter-trame(s), sont illustrées à la Figure 3.5. Cette figure montre l'évolution de la trajectoire des LSP d'ordre 16, extraits d'un morceau de parole en bande élargie.

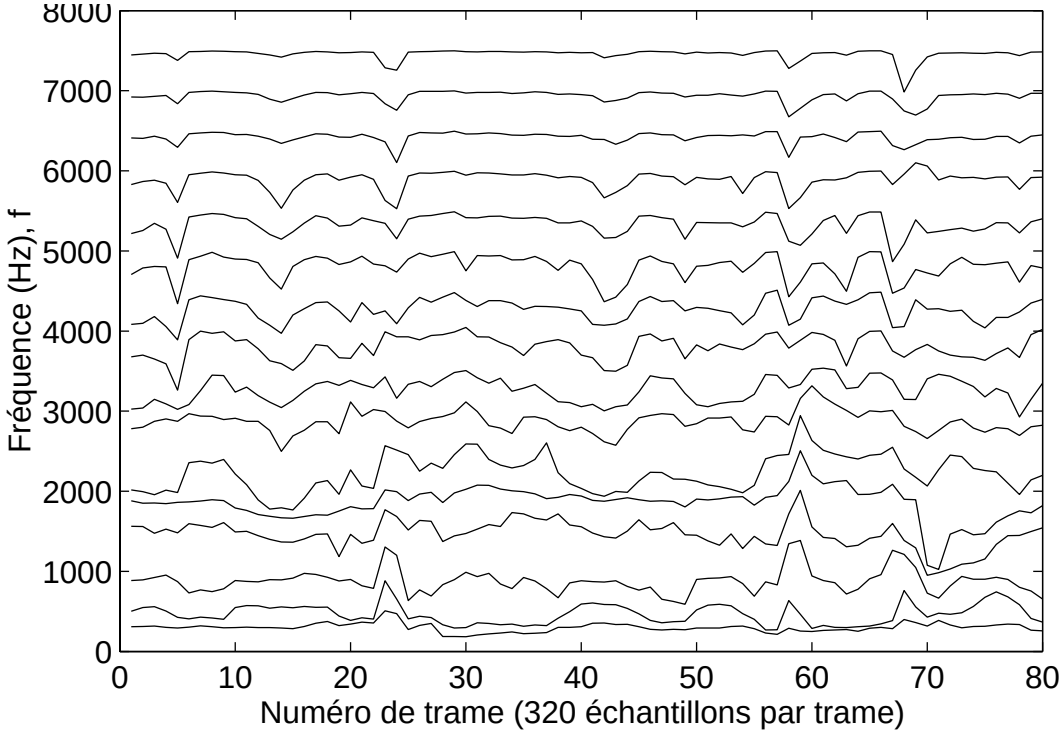

Figure 3.5 : Trajectoire des coefficients LSP pour la phrase « Alain fait du lin limpide pour Alain », prononcée par une femme.

Sensibilité spectrale localisée

La sensibilité spectrale de chaque LSP est localisée [3-13]. Ainsi, si une perturbation touche un LSP donné, elle n'influence que le spectre de puissance LPC dans la zone de fréquences très proche du LSP perturbé. Par conséquent, chaque LSP peut être quantifié individuellement, sans influencer la distorsion spectrale en dehors de la zone spectrale qui lui correspond. Les autres représentations alternatives des LPC, telles que les coefficients RC, LAR et IS n'offrent pas cette propriété.

Lien direct avec les formants de l'enveloppe spectrale

La Figure 3.4 montre qu'un groupe de 2 à 3 fréquences LSP caractérise la fréquence d'un formant. Ceci est particulièrement évident pour les premiers formants. La proximité des LSP successifs caractérise la largeur de bande du formant correspondant. Cette propriété est discutée en [3-14]. Elle peut s'exploiter pour la quantification des LSP, puisqu'elle est liée à la perception auditive. Pour cela, on peut utiliser une mesure de distance pondérée, lors de la recherche dans le (ou les) dictionnaire(s) de LSP quantifiés. La pondération est fonction de la proximité des LSP consécutifs, et permet d'attribuer plus ou moins d'importance aux formants lors de la quantification.

3.3.7 Quantification des paramètres LSP

Les paramètres LSP sont la représentation la plus utilisée pour la quantification du filtre de synthèse LPC. Cette sous-section décrit brièvement les différentes techniques de quantification possibles. La discussion est principalement axée sur le codage de la parole en bande étroite. Les différentes techniques de quantification vectorielle sont évaluées en termes de nombre de bits nécessaires pour atteindre une quantification transparente.

Quantification scalaire

La sensibilité spectrale localisée des LSP les rend idéaux pour une quantification scalaire. En effet, dans le cas d'une telle quantification, chaque LSP est quantifié séparément. Ainsi, une erreur de quantification sur un LSP donné n'a qu'une influence sur les fréquences qu'il caractérise, et n'a que très peu d'influence sur les LSP environnants. De plus, avec une quantification scalaire, chaque LSP est quantifié en utilisant une allocation de bits non-uniforme, propre à l'indice du coefficient LSP à quantifier. Ainsi, le nombre de niveaux de quantification est variable. Une quantification non-uniforme

donne de meilleurs résultats en termes de distorsion spectrale. Les quantificateurs sont couramment implantés à l'aide de l'algorithme de Lloyd [3-15].

Afin d'exploiter la corrélation inter-trames, il est utile de quantifier la différence entre un LSP et le LSP correspondant de la trame précédente. On appelle ce type de quantification, la quantification prédictive. De plus, afin d'exploiter la corrélation intra-trame, il est également possible de quantifier la différence entre LSP voisins dans la même trame. On appelle ce type de quantification, la quantification différentielle. Le fait d'utiliser une quantification différentielle ou prédictive rend la quantification plus sensible aux erreurs de canal.

Dans le cas de la bande étroite, où 10 coefficients LSP sont extraits par trame de signal, en exploitant la corrélation inter-trames il est pratiquement possible d'obtenir une quantification transparente avec 30 bits [3-1].

Bien que la quantification scalaire respecte la sensibilité spectrale localisée des LSP, la quantification vectorielle, présentée ci-dessous, exploite d'autres propriétés intéressantes des LSP et est souvent préférée à la quantification scalaire.

Quantification vectorielle

La quantification vectorielle exploite la corrélation intra-trame des paramètres LSP, et permet d'obtenir avec le même débit, une distorsion spectrale inférieure par rapport à la quantification scalaire.

Dans le cas de la bande étroite et par une estimation informelle, Paliwal suggère en [3-1] que la limite inférieure du nombre de bits nécessaires pour atteindre la transparence est 20, si une quantification vectorielle non-prédictive est utilisée. Naturellement ce nombre est plus élevé pour la bande élargie. Dans le cas de la bande étroite, avec 10 coefficients LSP et 20 bits de quantification, un quantificateur unique devrait contenir plus d'un million de vecteurs de code de dimension 10. Cela demanderait non seulement un nombre prohibitif de données d'entraînement et une taille de mémoire inconcevable, mais également une complexité de calcul trop élevée pour l'extraction des paramètres LSP quantifiés en temps réel.

La taille de la mémoire ainsi que la complexité de calcul peuvent être réduites en utilisant diverses méthodes de quantification vectorielle sous-optimales. Toutefois, ces méthodes abaissent les performances de la quantification pour un nombre de bits donné. Parmi les méthodes sous-optimales, on peut citer la quantification vectorielle séparée en sous-vecteurs (SVQ : Split Vector Quantization) [3-16], la quantification vectorielle sur

plusieurs étages de quantification (MSVQ : Multi-Stage Vector Quantization) [3-17], ou encore la quantification vectorielle structurée en arbre [3-18].

En [3-1], Paliwal teste les deux premières méthodes citées ci-dessus, en utilisant la même base de donnée de parole en bande étroite. A peine plus de 26 bits/trame sont nécessaires pour atteindre la transparence en utilisant une quantification sur plusieurs étages, alors que 26 bits/trame sont suffisants pour la quantification en sous-vecteurs. De plus, si une mesure de distance pondérée est utilisée pour la quantification, alors 25 bits/trame sont nécessaires pour la quantification multi-étages, alors que 24 sont suffisants pour la quantification en sous-vecteurs. La quantification en sous-vecteurs semble donc plus efficace que la quantification multi-étages.

La mesure de distance pondérée exploite la relation entre la distance séparant les LSP consécutifs et les formants de l'enveloppe spectrale. Plus de poids est donné aux LSP correspondant à des formants de largeur de bande étroite, qu'aux LSP correspondant à des formants de largeur de bande plus large. Finalement, de très petits poids sont attribués aux LSP correspondant à des vallées spectrales. L'utilisation d'une telle mesure peut permettre de réduire légèrement le débit.

Afin d'exploiter les propriétés d'inter-corrélation des LSP, on peut réaliser une quantification vectorielle prédictive, en implantant soit une prédiction auto-régressive (AR), soit une prédiction à moyenne glissante (moving average (MA)). Bien que la prédiction AR soit plus efficace que la prédiction MA, cette dernière est souvent retenue car elle est moins sensible aux erreurs de transmission du canal [3-19].

Alors que la quantification vectorielle permet de réduire considérablement le débit de transmission, la quantification scalaire a l'avantage d'être moins complexe en calcul, plus robuste contre la variation des locuteurs et des environnements. De plus, la quantification scalaire permet une protection plus efficace contre les erreurs de transmission du canal [3-20].

3.3.8 Extraction des paramètres LSP

Les paramètres LSP s'obtiennent en extrayant les racines des polynômes $P'(z)$ et $Q'(z)$ de l'équation (3.13). Les solutions des équations $P'(z) = 0$ et $Q'(z) = 0$, obtenues en utilisant une méthode numérique telle que la méthode de Newton-Raphson [3-21] sont trop coûteuses en termes de complexité de calcul. En effet, une telle méthode implique de résoudre un polynôme d'ordre 16 en utilisant une arithmétique compliquée.

D'autres méthodes, telles que celle proposée par Kabal en [3-22], sont plus adaptées pour une réalisation en temps réel. Ces méthodes sont

présentées en [3-4]. La méthode de Kabal, retenue pour l'implantation du codeur développé dans le cadre de ce travail de thèse, est décrite à l'Annexe A.

3.3.9 Transformation des LSP en LPC

La transformation des coefficients LSP en coefficients LPC est moins coûteuse en termes de complexité de calcul que la conversion de LPC en LSP. Sur la base des polynômes symétriques et antisymétriques $P(z)$ et $Q(z)$ donnés par l'équation (3.17), le filtre d'analyse LPC s'exprime ainsi :

$$A_{16}(z) = \frac{P(z) + Q(z)}{2}, \tag{3.18}$$

Les polynômes $P(z)$ et $Q(z)$ sont obtenus à partir des paramètres LSP $\{\omega_i\}$ en utilisant la relation suivante :

$$\begin{aligned} P(z) &= (1+z^{-1}) \prod_{i=1,3,5,7,9,11,13,15} \left[1 - 2\cos(\omega_i)\cdot z^{-1} + z^{-2}\right], \\ Q(z) &= (1+z^{-1}) \prod_{i=2,4,6,8,10,12,14,16} \left[1 - 2\cos(\omega_i)\cdot z^{-1} + z^{-2}\right]. \end{aligned} \tag{3.19}$$

Il est important de noter que si les paramètres LSP sont exprimés dans le domaine "x", où $x_i = \cos(\omega_i)$, comme c'est le cas lorsqu'on utilise la méthode de Kabal, alors cette transformation est facilitée puisqu'elle ne requiert aucun calcul et stockage des fonctions trigonométriques. Il existe diverses méthodes pour résoudre ces équations. La méthode d'expansion directe et la méthode de Kabal sont présentées ci-dessous. On retiendra la méthode de Kabal qui est peu coûteuse en termes de complexité de calcul.

Méthode d'expansion directe

Les polynômes $P(z)$ et $Q(z)$ s'obtiennent en multipliant les termes de l'équation (3.19). Les coefficients LPC se calculent alors au moyen de l'équation (3.18). Ce calcul est donnée à l'Annexe B.1 et a une complexité totale de 618 additions et 519 multiplications. Par rapport à la complexité calculée en [3-4] pour la même méthode mais avec un ordre p de 10, on a ici pratiquement 7 fois plus d'additions et 10 fois plus de multiplications !

Méthode de Kabal

En [3-22], Kabal et Ramachandran proposent une méthode de reconstruction qui utilise la représentation en séries de Tchebychev. Cette méthode est très efficace puisqu'elle tient compte de la symétrie des polynômes. Elle est décrite à l'Annexe B.2 et sa complexité de calcul totale est de 88 multiplications et de 143 additions. La complexité de calcul relative à la méthode de Kabal est ainsi fortement inférieure à celle relative à la méthode d'expansion directe. Par rapport à la complexité calculée en [3-4] pour la même méthode mais avec un ordre p de 10 (bande étroite), on a ici pratiquement 3 fois plus d'additions et 4.5 fois plus de multiplications !

3.4 Extraction de l'excitation adaptative par prédiction à long-terme

La prédiction à long-terme, ou la prédiction du "pitch", modélise la structure fine de l'enveloppe spectrale du signal de parole. Lorsque le signal original est voisé, le résidu de prédiction à court-terme contient encore de la redondance sous la forme de périodicité. Cette périodicité est relative à la période du "pitch". Elle est de l'ordre de 30 à 320 échantillons pour la parole en bande élargie. Si l'on ajoute la prédiction de pitch au résidu de prédiction et ceci sous la forme du filtre d'analyse à long-terme donné par l'équation (2.8), le nouveau signal résiduel est un signal de type bruité. Ce signal ressemble à du bruit blanc pour la parole en bande étroite, alors que pour la parole en bande élargie, ce bruit présente une pente spectrale.

La forme générale du filtre de synthèse à long-terme est donnée par :

$$\frac{1}{P(z)} = \frac{1}{1 - P_l(z)} = \frac{1}{1 - \sum_{k=-m_1}^{m_2} G_k z^{-(T+k)}}, \tag{3.20}$$

avec le prédicteur à long-terme $P_l(z)$ donné par :

$$P_l(z) = \sum_{k=-m_1}^{m_2} G_k z^{-(T+k)}. \tag{3.21}$$

T est le délai tonal entier ou l'un de ses multiples, alors que les G_k sont les coefficients de prédiction à long-terme. Généralement, on a soit $m_1 = m_2 = 0$, soit $m_1 = m_2 = 1$, ce qui correspond à une prédiction sur un et respectivement trois coefficients. La deuxième solution est plus précise que la première, mais elle nécessite un débit et une complexité de calcul plus élevés. De même, elle entraîne des difficultés pour assurer la stabilité du filtre de synthèse. On

utilisera la première solution et l'on introduira une interpolation des valeurs prédites afin d'augmenter la précision.

Les paramètres T et G_0 sont déterminés en minimisant sur un segment de N échantillons de signal, l'erreur carrée moyenne sur le signal résiduel de prédiction à court-terme $r(n)$, après prédiction à long-terme. Soit cette erreur carrée moyenne donnée par :

$$E = \sum_{n=0}^{N-1} e^2(n) = \sum_{n=0}^{N-1} \left[r(n) - G_0 r(n-T) \right]^2, \tag{3.22}$$

où le résidu de prédiction à long-terme vaut :

$$e(n) = r(n) - G_0 r(n-T). \tag{3.23}$$

En posant $\partial E / \partial G_0 = 0$ pour minimiser cette erreur, on trouve :

$$G_0 = \frac{\sum_{n=0}^{N-1} r(n) r(n-T)}{\sum_{n=0}^{N-1} \left[r(n-T) \right]^2}. \tag{3.24}$$

En substituant cette valeur dans l'équation (3.22), on obtient :

$$E = \sum_{n=0}^{N-1} r^2(n) - \frac{\left[\sum_{n=0}^{N-1} r(n) r(n-T) \right]^2}{\sum_{n=0}^{N-1} \left[r(n-T) \right]^2}. \tag{3.25}$$

Minimiser E signifie maximiser le second terme du côté droit de l'équation ci-dessus. Ce terme correspond à la corrélation normalisée entre le signal résiduel $r(n)$ et sa version décalée dans le temps. Il est calculé pour toutes les valeurs T comprises dans un intervalle de valeurs possibles. La valeur T qui maximise ce terme est retenue.

Soit l'énergie ε_T :

$$\varepsilon_T = \sum_{n=0}^{N-1} \left[r(n-T) \right]^2, \tag{3.26}$$

apparaissant à l'équation (3.25). Le calcul de ε_T peut être simplifié en l'exprimant en fonction de ε_{T-1}. Ainsi :

$$\varepsilon_T = \varepsilon_{T-1} + r^2(-T) - r^2(-T+N). \tag{3.27}$$

Ce calcul ne requiert que 4 opérations (2 additions et 2 multiplications).

Le gain G_0 s'obtient selon l'équation (3.24). Le filtre de synthèse LTP, $1/P(z)$, n'est stable que si $|G_0| \leq 1.0$. Cependant, une instabilité momentanée de ce filtre n'est pas dramatique. En effet, de par la nature temporelle du signal de parole, le gain G_0 ne peut être supérieur à 1.0 que pour un petit nombre de sous-trames consécutives. Ainsi, l'énergie du signal de sortie ne peut augmenter démesurément et créer un inconfort.

3.4.1 Extraction de l'excitation adaptative dans la boucle : implantation du dictionnaire adaptatif [3-23]

Le début de la Section 3.4 décrit l'extraction des paramètres de prédiction à long-terme en boucle ouverte, à l'extérieur de la boucle de minimisation de l'erreur, et directement sur le signal résiduel. Toutefois, une amélioration importante de la synthèse de la parole est apportée, si la recherche de ces paramètres se fait en boucle fermée, en tenant compte de la minimisation de l'erreur pondérée perceptuellement.

Dans le schéma de l'encodeur CS-ACELP illustré à la Figure 3.1, le filtre de synthèse LTP est remplacé par une excitation adaptative extraite d'un dictionnaire. Ce dictionnaire est appelé dictionnaire adaptatif. Il est remis à jour pour chacune des sous-trames du signal. Pour une telle implantation, l'extraction des paramètres LTP en boucle fermée se fait comme suit.

L'extraction des paramètres LTP est réalisée avant l'extraction de l'excitation innovatrice algébrique. Dans un premier temps, cette dernière est donc considérée comme nulle. On cherche à minimiser l'erreur pondérée perceptuellement entre la cible $x(n)$ et l'excitation adaptative filtrée par le filtre de synthèse pondéré perceptuellement : $x_a(n)$. Soit $r_\tau(n)$, l'excitation adaptative correspondante à un retard temporel τ et soit G, le gain correspondant, $x_a(n)$ vaut :

$$x_a(n) = G\sum_{i=0}^{n} r_\tau(i)h(n-i) + \hat{x}_a(n), \tag{3.28}$$

où $h(n)$ est la réponse impulsionelle du filtre $W(z)/\hat{A}(z)$ et où $\hat{x}_a(n)$ est la réponse à zéro de ce filtre, c'est-à-dire la sortie de ce filtre due à ses états initiaux. L'erreur pondérée entre le signal original et la parole reconstruite vaut donc :

$$e_a(n) = x(n) - x_a(n) = x'_a(n) - G\sum_{i=0}^{n} r_\tau(i)h(n-i), \tag{3.29}$$

où

$$x'_a(n) = x(n) - \hat{x}_a(n), \tag{3.30}$$

et où $x(n)$ est le signal cible correspondant au signal pré-traité $s_{pt}(n)$ filtré par $W(z)$ (cf. Sous-section 3.2.1).

Soit le vecteur d'excitation $u(n)$ contenant les valeurs d'excitations passées selon la Figure 2.9, on a :

$$r_\tau(n) = u(n-\tau) \tag{3.31}$$

et

$$e_a(n) = x'_a(n) - Gy_\tau(n), \tag{3.32}$$

où

$$y_k(n) = u(n-k) * h(n) = \sum_{i=0}^{n} u(i-k)h(n-i). \tag{3.33}$$

L'erreur carrée moyenne pondérée est donnée par :

$$E_a = \sum_{n=0}^{N-1} \left[x'_a(n) - Gy_\tau(n)\right]^2. \tag{3.34}$$

En posant $\partial E_a / \partial G = 0$ on obtient :

$$G = \frac{\sum_{n=0}^{N-1} x'_a(n) y_\tau(n)}{\sum_{n=0}^{N-1} \left[y_\tau(n)\right]^2}. \tag{3.35}$$

En insérant l'équation (3.35) dans l'équation (3.34), on obtient :

$$E_a = \sum_{n=0}^{N-1} \left[x'_a(n)\right]^2 - \frac{\left[\sum_{n=0}^{N-1} x'_a(n) y_\tau(n)\right]^2}{\sum_{n=0}^{N-1} \left[y_\tau(n)\right]^2}. \tag{3.36}$$

Le délai tonal T est le délai τ qui maximise le second terme de l'équation ci-dessus. Une fois T déterminé, G est calculé avec l'équation (3.35).

Comme précité, la qualité de la parole reconstruite est fortement améliorée si la recherche se fait en boucle fermée. Le désavantage de cette solution est la complexité de calcul nécessaire à l'extraction de la convolution de l'équation (3.33). Une procédure rapide pour calculer cette convolution $y_\tau(n)$ pour tous les délais possibles est de la calculer pour la première valeur et de la remettre ensuite à jour en utilisant la relation suivante :

$$\begin{aligned} y_i(0) &= u(-i)h(0), \\ y_i(n) &= u(-i)h(n) + y_{i-1}(n-1). \end{aligned} \tag{3.37}$$

Les excitations synthétisées passées, $u(n)$, sont contenues dans un registre de stockage à décalage adaptatif. Le délai est compris entre T_{max} et 0. T_{max} est la longueur du registre. Le contenu du registre est remis à jour pour chacune des sous-trames en introduisant N nouveaux échantillons et en décalant les anciens échantillons de N valeurs. On a donc :

$$u(n) \leftarrow u(n+N), \quad n = -T_{\max}, \ldots, -1. \tag{3.38}$$

Ce registre à décalage peut être représenté comme un dictionnaire adaptatif, dans lequel chaque mot de code $r_i(n)$ est obtenu en décalant d'un échantillon le mot de code précédant. Les mots de code sont donnés par :

$$r_i(n) = u(-i+n), \quad n = 0, \ldots, N-1; \quad i = N, \ldots, T_{\max}. \tag{3.39}$$

Pour chaque délai inférieur à la longueur de la sous-trame N, seules les premières i valeurs sont disponibles. Dans ce cas, les mots de code $r_i(n)$ sont reconstruits en répétant les valeurs disponibles jusqu'à compléter le mot. Ainsi pour $i < N$, on a

$$r_i(n) = \begin{cases} u(-i+n), & avec \quad n = 0, \ldots, i-1; \\ u(-2i+n), & avec \quad n = i, \ldots, 2i-1. \end{cases} \tag{3.40}$$

Une approche plus simple pour traiter les délais inférieurs à la longueur de la sous-trame est d'étendre le registre d'excitation par le résidu de prédiction à court-terme du signal en cours de traitement.

Les performances du prédicteur de délai tonal peuvent être augmentées en utilisant un délai tonal fractionnel. En effet, souvent le délai tonal ne coïncide pas avec un instant d'échantillonnage. Dans ce cas, le délai entier le plus proche du délai réel, ou un multiple de celui-ci, sera sélectionné. Afin de trouver le délai le plus proche du délai réel, il faut utiliser une résolution d'échantillonnage plus élevée [3-23]. Pour cela on interpole le signal d'excitation $u(n)$, afin d'obtenir des mots de code interpolés. Le facteur de sur-échantillonnage dépend de la résolution requise.

L'utilisation de délais tonals fractionnels introduit un accroissement de la complexité de calcul considérable. Afin d'éviter cela, la procédure suivante est utilisée :

- La fonction de correspondance $\Psi(\tau)$, n'est déterminée que pour des délais entiers. Cette fonction est la racine carrée du second terme de l'équation (3.36). Elle est donnée par :

$$\Psi(\tau) = \frac{\left| \sum_{n=0}^{N-1} x'_a(n) y_\tau(n) \right|}{\sqrt{\sum_{n=0}^{N-1} \left[y_\tau(n) \right]^2}}, \quad \tau = T_{\min}, \ldots, T_{\max}. \tag{3.41}$$

- L'interpolation n'est d'abord réalisée que sur la fonction de correspondance $\Psi(\tau)$. Si le maximum de la fonction de correspondance interpolée correspond à un délai non-entier, alors seulement l'excitation correspondante du registre est interpolée. On calcule ainsi le mot du dictionnaire correspondant à ce délai non-entier.
- Seul un petit nombre de points interpolés, de la fonction de correspondance, doivent être calculés. Ces points interpolés sont déterminés autour du délai entier qui maximise cette fonction.

Pour profiter des avantages de simplicité de la boucle ouverte et des performances de la boucle fermée, on choisit de réaliser une estimation du délai tonal en boucle ouverte et de réaliser ensuite une recherche en boucle fermée uniquement autour du délai déterminé en boucle ouverte.

3.4.2 Quantification des paramètres LTP

Les paramètres LTP sont le délai τ, entier ou fractionnel, et le gain G qui lui correspond. On les appelle l'indice et le gain du dictionnaire adaptatif. Ils sont quantifiés séparément.

Afin de réduire le débit de transmission, on peut utiliser un encodage différencié de l'indice du dictionnaire, et ceci pour une sous-trame sur deux. On estime le délai T_0 pour deux sous-trames groupées, puis on détermine autour de T_0, le délai de chacune des sous-trames T_1 et T_2. On encode T_1 de façon exacte, alors que pour T_2, seule la variation par rapport à T_1 est encodée. Afin de simplifier la procédure totale, on recherche T_0 en boucle ouverte, puis T_1 et T_2 en boucle fermée.

Le gain G présente une distribution non-uniforme. Ainsi un quantificateur non-uniforme de Lloyd-Max est implanté sur la base d'une grande base de données. Les gains proches de 1.0 correspondent à des segments de parole voisée. Les gains négatifs correspondent à des segments de parole non-voisée, ne contenant aucune périodicité. On peut limiter les valeurs du gain à l'intervalle [0, 1.2]. Un gain supérieur à l'unité peut entraîner une instabilité dans l'énergie du signal reconstruit (cf. Sous-section 3.4). Toutefois, une telle instabilité n'est pas dramatique puisque la valeur du

gain est remise à jour pour chacune des sous-trames. Le gain peut être quantifié soit avant l'extraction de l'excitation innovatrice, soit conjointement au gain de cette deuxième excitation. Dans le premier cas on utilisera une quantification scalaire, dans le second cas une quantification vectorielle.

3.5 Extraction de l'excitation innovatrice [3-23]

Les filtres d'analyse LPC et LTP (ou la réalisation de ce dernier sous la forme de dictionnaire) enlèvent la corrélation à court-terme et à long-terme du signal. Le signal cible restant, $x'(n)$, illustré à la Figure 3.1, ne contient pratiquement plus de redondance. Ce signal est de type "bruité". Dans le cas de la bande étroite, il a l'aspect d'un bruit blanc, alors que dans le cas de la bande élargie son spectre présente généralement une certaine pente.

La partie cruciale du modèle de codage de la parole reposant sur l'analyse par synthèse, est la détermination du vecteur d'excitation innovatrice. Ce vecteur excite aussi bien le filtre de synthèse LTP, ici remplacé par un dictionnaire, que le filtre de synthèse LPC. L'excitation innovatrice doit être implantée intelligemment, pour rendre le nombre de bits nécessaire à sa transmission aussi petit que possible tout en maintenant une complexité de traitement raisonnable.

Les premiers codeurs de type CELP modélisent l'excitation innovatrice $c(n)$, par un processus Gaussien à moyenne nulle, et dont le spectre de puissance varie lentement dans le temps. Ainsi l'excitation innovatrice est implantée sous la forme d'un grand dictionnaire stochastique. Actuellement, pour réduire la complexité de recherche de l'excitation ainsi que la taille de la mémoire nécessaire à l'implantation d'un dictionnaire stochastique, on utilise couramment un dictionnaire de type algébrique en lieu et place du dictionnaire stochastique. Ceci est d'autant plus important pour la bande élargie.

L'extraction de l'excitation innovatrice se fait par minimisation de l'erreur pondérée perceptuellement entre le signal cible $x'(n)$ et l'excitation innovatrice filtrée par le filtre de synthèse pondéré perceptuellement $x_b(n)$. Soit $c_k(n)$, l'excitation innovatrice testée, et β le gain qui lui correspond, alors $x_b(n)$ vaut :

$$x_b(n) = \beta \sum_{i=0}^{n} c_k(i) h(n-i) + \hat{x}_b(n), \tag{3.42}$$

où $h(n)$ est la réponse impulsionelle du filtre $W(z)/\hat{A}(z)$, et où $\hat{x}_b(n)$ est la réponse à zéro de ce filtre. La réponse à zéro d'un filtre correspond à la sortie

de ce filtre excité par ses seuls états initiaux. L'erreur pondérée $e(n)$ entre le signal original et la parole reconstruite vaut ainsi :

$$e(n) = x'(n) - x_b(n) = x'_b(n) - \beta \sum_{i=0}^{n} c_k(i) h(n-i), \tag{3.43}$$

où

$$x'(n) = x(n) - x_a(n), \tag{3.44}$$

avec $x_a(n)$ donnée par l'équation (3.28), et

$$x'_b(n) = x'(n) - \hat{x}_b(n). \tag{3.45}$$

$x(n)$ est le signal cible correspondant au signal pré-traité $s_{pt}(n)$ filtré par $W(z)$ (cf. Sous-section 3.2.1). Ainsi, l'erreur carrée moyenne pondérée E_b est donnée par :

$$E_b = \sum_{n=0}^{N-1} \left[e(n) \right]^2 = \sum_{n=0}^{N-1} \left[x'_b(n) - \beta c_k(n) * h(n) \right]^2. \tag{3.46}$$

En posant $\partial E_b / \partial \beta = 0,$ on obtient :

$$\beta = \frac{\sum_{n=0}^{N-1} x'_b(n) \left[c_k(n) * h(n) \right]}{\sum_{n=0}^{N-1} \left[c_k(n) * h(n) \right]^2}. \tag{3.47}$$

En remplaçant β dans l'équation (3.46) avec la valeur trouvée dans (3.47), E_b devient :

$$E_b = \sum_{n=0}^{N-1} \left[x'_b(n) \right]^2 - \frac{\left[\sum_{n=0}^{N-1} x'_b(n) \left[c_k(n) * h(n) \right] \right]^2}{\sum_{n=0}^{N-1} \left[c_k(n) * h(n) \right]^2}. \tag{3.48}$$

Les équations (3.47) et (3.48) peuvent être réécrites comme suit, sous forme matricielle, où l'on suppose que $\mathbf{x}_b$ et $\mathbf{c}_k$ sont des vecteurs-colonne :

$$\beta = \frac{\mathbf{x}'^T_b \mathbf{H} \mathbf{c}_k}{\mathbf{c}_k^T \mathbf{H}^T \mathbf{H} \mathbf{c}_k}, \tag{3.49}$$

et

$$E_b = \left\| \mathbf{x}'_b - \beta \mathbf{H}\mathbf{c}_k \right\|^2 = \mathbf{x}'^T_b \, \mathbf{x}'_b - \frac{\left(\mathbf{x}'^T_b \, \mathbf{H}\mathbf{c}_k \right)^2}{\mathbf{c}_k^T \mathbf{H}^T \mathbf{H}\mathbf{c}_k}. \tag{3.50}$$

On a :

$$\mathbf{x}'^T_b = \begin{pmatrix} x'_{b0} & x'_{b1} & \dots & x'_{bN-1} \end{pmatrix}, \tag{3.51}$$

$$\mathbf{c}^T = \begin{pmatrix} c_0 & c_1 & \dots & c_{N-1} \end{pmatrix}, \tag{3.52}$$

$\mathbf{H}$ étant la matrice de convolution de la réponse impulsionnelle $h(n)$, triangulaire sur la partie inférieure. Cette matrice est donnée par :

$$\mathbf{H} = \begin{pmatrix} h_0 & 0 & 0 & & 0 \\ h_1 & h_0 & 0 & & 0 \\ h_2 & h_1 & h_0 & & 0 \\ & & h_1 & \ddots & \vdots \\ h_{N-1} & h_{N-2} & h_{N-3} & \cdots & h_0 \end{pmatrix}. \tag{3.53}$$

Soit

$$\mathbf{\Theta} = \mathbf{H}^T \mathbf{H}, \tag{3.54}$$

alors $\mathbf{\Theta}$ est une matrice symétrique contenant la corrélation de la réponse impulsionnelle :

$$\phi(i,l) = \sum_{n=\max(i,l)}^{N-1} h(n-i)h(n-l), \quad i,l = 0,\dots,N-1. \tag{3.55}$$

Soit

$$\mathbf{\Psi}^T = \mathbf{x}'^T_b \, \mathbf{H}, \tag{3.56}$$

un vecteur dont les éléments valent :

$$\psi(i) = x'_b(i) * h(-i) = \sum_{n=i}^{N-1} x'_b(n)h(n-i), \quad i = 0,\dots,N-1, \tag{3.57}$$

alors, l'erreur carrée pondérée moyenne se minimise en maximisant le second terme de l'équation (3.50), donné par :

$$\mathbf{\Gamma}_k = \frac{(C_k)^2}{\xi_k} = \frac{\left(\mathbf{x}'^T_b \, \mathbf{H}\mathbf{c}_k \right)^2}{\mathbf{c}_k^T \mathbf{H}^T \mathbf{H}\mathbf{c}_k} = \frac{\left(\mathbf{\Psi}^T \mathbf{c}_k \right)^2}{\mathbf{c}_k^T \mathbf{\Theta}\mathbf{c}_k}. \tag{3.58}$$

C_k est la cross-corrélation entre $\mathbf{x}'_b$ et le mot de code filtré $\mathbf{Hc}_k$. ξ_k est l'énergie du mot de code $\mathbf{c}_k$ filtré. On a :

$$C_k = \sum_{n=0}^{N-1} x_b'(n)\left[c_k(n) * h(n)\right] = \sum_{n=0}^{N-1} \psi(n) c_k(n), \tag{3.59}$$

et

$$\begin{aligned} \xi_k = & \sum_{n=0}^{N-1} \left[c_k(n) * h(n)\right]^2 = \\ & \sum_{i=0}^{N-1} c_k^2(i)\phi(i,i) + 2\sum_{i=0}^{N-2}\sum_{l=i+1}^{N-1} c_k(i)c_k(l)\phi(i,l). \end{aligned} \tag{3.60}$$

$\phi(i,l)$ et $\psi(n)$ sont calculés à l'extérieur de la boucle de minimisation de l'erreur. Le terme $\mathbf{\Gamma}_k$ est sur cette base évalué pour tous les indexes k allant de 0 à $L-1$. L est la longueur du dictionnaire. Le mot de code dont l'index k maximise ce terme est choisi et le gain scalaire β est alors calculé, puis quantifié. La quantification de β se fait soit de façon scalaire, soit de façon vectorielle, associée à la quantification du gain G du dictionnaire adaptatif. Dans le cas d'une quantification vectorielle, on minimise le signal d'erreur pondéré $e(n)$ (cf. Figure 3.1), en prenant les mots des codes extraits des différents dictionnaires, et en testant toutes les combinaisons vectorielles possibles du dictionnaire de gains vectoriels.

La quantification du gain β peut se faire également en utilisant une prédiction MA sur la base de l'énergie de l'excitation innovatrice multipliée par son gain associé. Cette technique est introduite à la Sous-section 3.5.2.

La recherche exhaustive de l'excitation innovatrice est une procédure très coûteuse en termes de complexité de calcul. Elle est par conséquent difficile à implanter en temps réel. Il existe différentes méthodes permettant de simplifier cette recherche, sans affecter la qualité du signal reconstruit. Une méthode intéressante est la méthode d'auto-corrélation introduite en [3-23]. Cette méthode simplifie la représentation de l'erreur carrée moyenne pondérée et réduit la charge excessive de calcul, pour rechercher la séquence d'innovation optimale. Une autre méthode consiste à utiliser des dictionnaires innovateurs structurés permettant une recherche rapide. Un tel dictionnaire est le dictionnaire algébrique présenté à la sous-section suivante.

3.5.1 Dictionnaire algébrique

Le dictionnaire d'excitations innovatrices peut être peuplé de codes algébriques. S'ils sont très structurés, ces codes permettent d'utiliser un algorithme de recherche très efficace.

La structure basée sur une permutation d'impulsions simples entrelacées, appelée ISPP (Interleaved Single-Pulse Permutation), est décrite ici puisque c'est celle utilisée par l'algorithme implanté dans le cadre de ce travail de thèse.

Les vecteurs d'excitation réalisés par codes basés sur une permutation entrelacée, sont composés principalement de valeurs nulles et de quelques impulsions non-nulles. Les amplitudes des impulsions non-nulles sont fixées à +1.0 ou –1.0. On pré-définit des ensembles de positions entrelacées. Chaque impulsion peut être positionnée sur un des ensembles de positions, qui est généralement distinct des ensembles correspondant aux autres impulsions. Le code d'excitation est défini par les positions des impulsions non-nulles et par le signe de celles-ci. La recherche dans le dictionnaire innovateur consiste ici essentiellement en une recherche des positions optimales des impulsions non-nulles.

L'exemple suivant décrit une structure de dictionnaire algébrique pour un vecteur d'excitation de 80 échantillons, dont seuls cinq échantillons du vecteur sont non-nuls avec une amplitude valant +1.0 ou –1.0. Les 80 échantillons sont divisés en cinq ensembles de 16 échantillons. Le premier ensemble contient les échantillons 0,5,10,15, …,75; le second ensemble contient les échantillons 1,6,11, …,76, etc. On a donc cinq ensembles de 16 échantillons. Chaque ensemble est appelé "une piste". Si on dénote une piste par l'indice i et la position de l'échantillon dans une piste par l'indice l, on aura $i = 0,1,...,4;\ \ l(i) = 0,1,...,15,$ et la position m_i de la $i^{\text{ème}}$ impulsion sera spécifiée par :

$$m_i = i + 5 \cdot l(i), \quad i = 0,1,2,3,4, \quad l(i) = 0,1,...,15. \tag{3.61}$$

Les positions que chaque impulsion i peut prendre, sont données par le Tableau 3-1.

Impulsion	***Signe***	***Positions* (q ∈[0,1,…,15])**
0	$s_0 : \pm 1$	m_0 : 0+5q
1	$s_1 : \pm 1$	m_1 : 1+5q
2	$s_2 : \pm 1$	m_2 : 2+5q
3	$s_3 : \pm 1$	m_3 : 3+5q
4	$s_4 : \pm 1$	m_4 : 4+5q

Tableau 3-1 : Structure du dictionnaire algébrique proposé, avec 5 pistes et une impulsion par piste, pour une excitation de 80 échantillons.

La position d'un échantillon non-nul dans une piste donnée est ainsi quantifiée sur 4 bits. On obtient un dictionnaire de 20 bits au total, qui spécifie l'emplacement des cinq impulsions non-nulles. La piste est implicitement donnée par l'ordre d'encodage. L'encodage requiert 5 bits de plus pour spécifier les amplitudes des différentes impulsions.

La structure de dictionnaire présentée ci-dessus offre l'avantage considérable de ne pas requérir de stockage. Elle est robuste contre les erreurs de transmission. En effet, une erreur n'affecte qu'une impulsion et non pas le vecteur entier d'excitation. De plus, une telle structure de dictionnaire permet une recherche de minimisation de l'erreur très efficace. La cross-corrélation C_k de l'équation (3.59), devient ainsi :

$$C_k = \sum_{i=0}^{4} s_i \psi(m_i), \tag{3.62}$$

où $s_i = \pm 1.0$ est l'amplitude de m_i. De plus, l'énergie ξ_k du mot de code $\mathbf{c}_k$, donné par l'équation (3.60), devient :

$$\xi_k = \sum_{i=0}^{4} \phi(m_i, m_i) + 2 \sum_{i=0}^{3} \sum_{l=i+1}^{4} s_i s_l \phi(m_i, m_l). \tag{3.63}$$

Pour simplifier la procédure de recherche, les amplitudes des impulsions sont pré-déterminées en quantifiant le signal $\psi(n)$. On pose l'amplitude de l'impulsion en position n, égale au signe de $\psi(n)$ en cette position. Avant la recherche dans le dictionnaire, les étapes suivantes sont réalisées. D'abord le signal $\psi(n)$ est décomposé en deux parties : sa valeur absolue, dénotée $|\psi(n)|$ et son signe, dénoté $sign[\psi(n)]$:

$$\psi(n) = sign[\psi(n)] \cdot |\psi(n)| \tag{3.64}$$

Ainsi, la corrélation de l'équation (3.62) devient :

$$C_k = |\psi(m_0)| + |\psi(m_1)| + |\psi(m_2)| + |\psi(m_3)| + |\psi(m_4)|. \tag{3.65}$$

Ensuite, les éléments de la matrice $\mathbf{\Theta}$ de l'équation (3.54) sont modifiés comme suit, pour inclure l'information de signe. Soit:

$$\phi'(m_i, m_l) = sign[\psi(m_i)] sign[\psi(m_l)] \phi(m_i, m_l), \quad \begin{cases} m_i = 0, ..., 79, \\ m_l = m_i + 1, ..., 79. \end{cases} \tag{3.66}$$

Si l'on pose :

$$\phi'(m_i, m_i) = 0.5 \cdot \phi'(m_i, m_i), \quad m_i = 0, ..., 79, \tag{3.67}$$

alors, l'équation (3.63) devient :

$$\begin{aligned}\xi_k / 2 = \quad & \phi'(m_0, m_0) + \\ & \phi'(m_1, m_1) + \phi'(m_0, m_1) + \\ & \phi'(m_2, m_2) + \phi'(m_0, m_2) + \phi'(m_1, m_2) + \\ & \phi'(m_3, m_3) + \phi'(m_0, m_3) + \phi'(m_1, m_3) + \phi'(m_2, m_3) + \\ & \phi'(m_4, m_4) + \phi'(m_0, m_4) + \phi'(m_1, m_4) + \\ & \phi'(m_2, m_4) + \phi'(m_3, m_4).\end{aligned} \tag{3.68}$$

L'opération décrite par l'équation (3.67) permet de mettre à l'échelle les éléments de la diagonale principale de $\mathbf{\Theta}$ afin d'éliminer le facteur 2 de l'équation (3.63).

La recherche dans le dictionnaire innovateur est réalisée en cinq boucles emboîtées. Chaque boucle correspond au positionnement d'une impulsion. Ainsi à chaque boucle, la contribution d'une nouvelle impulsion est ajoutée. Cette nouvelle impulsion est prise en compte lors de la minimisation de l'erreur de la boucle suivante. Lors du traitement de la dernière boucle, la corrélation est remise à jour avec une addition et l'énergie est remise à jour avec cinq additions et une multiplication. Malgré l'efficacité de cette procédure de recherche, une recherche exhaustive devient rapidement compliquée. Pour réduire davantage la complexité, une stratégie de recherche a été développée, pour laquelle seul un petit sous-ensemble du dictionnaire est testé. Bien que cette méthode soit sous-optimale, les performances restent proches de celles obtenues avec une recherche exhaustive [3-24].

Avec l'exemple présenté ci-dessus et dans le but d'augmenter la qualité de reconstruction de la parole, on peut augmenter le nombre d'impulsions. Ceci se fait soit en augmentant le nombre de pistes, soit en attribuant plusieurs impulsions à une même piste. Dans le premier cas, il faut prendre garde à conserver un encodage optimal. Dans le deuxième cas, une fois les cinq premières impulsions attribuées, on revient sur la première piste pour ajouter la $6^{\text{ème}}$ impulsion, puis sur la deuxième pour ajouter la $7^{\text{ème}}$ impulsion et ainsi de suite. Dans le cas où l'on attribue plus d'une impulsion par piste, on utilise l'ordre d'encodage des différentes impulsions sur la même piste pour indiquer le ou les signe(s) des impulsions supplémentaires. Soit la première piste à laquelle on attribue deux impulsions m_0 et m_6, où $m_6 > m_0$; si les deux impulsions ont le même signe, on encodera m_0 puis m_6 et le signe de l'impulsion en m_0. Par contre si les signes sont opposés, on encodera m_6 puis m_0 et le signe de m_6. Ainsi l'ordre croissant ou décroissant d'encodage sur la même piste donne l'information de signe pour la deuxième impulsion de la piste. Les signes de m_6 et de m_0, en fonction de leur ordre d'encodage et du signe encodé (en gras), sont donnés au Tableau 3-2. On utilise cette même technique également si l'on dispose de plus de deux impulsions sur la même

piste. Si l'on a 4 impulsions par piste, seule la position relative des impulsions d'une piste suffit à encoder leur signe [3-25].

Ordre	**Signe de m_0**	**Signe de m_6**
m_0, m_6	+	+
m_6, m_0	-	+
m_0, m_6	-	-
m_6, m_0	+	-

Tableau 3-2 : Signe de m_6 en fonction de l'ordre d'encodage et du signe encodé (en gras) pour $m_0 < m_6$.

3.5.2 Prédiction du gain β [3-26]

Soit le vecteur $z(n)$, appelé contribution innovatrice, correspondant au vecteur d'excitation innovateur $c(n)$ mis à l'échelle par le gain qui lui correspond β, au temps n :

$$z(n) = \beta \cdot c(n). \tag{3.69}$$

Les racines carrées des variances associées à ces vecteurs au temps n valent $\sigma_z(n)$ et $\sigma_c(n)$. Elles sont liées par la relation :

$$\sigma_z(n) = \beta \sigma_c(n), \tag{3.70}$$

qui dans le domaine logarithmique devient :

$$\log_{10}\left[\sigma_z(n)\right] = \log_{10}\left[\beta\right] + \log_{10}\left[\sigma_c(n)\right]. \tag{3.71}$$

La philosophie du schéma de prédiction du gain β consiste à exploiter la corrélation entre sa valeur logarithmique présente et ses valeurs passées. Cette corrélation est une conséquence de la variation temporelle lente de l'enveloppe du signal de parole. On peut utiliser soit une prédiction de type AR, soit une prédiction de type MA.

Dans le codeur G.729 de l'ITU, le gain β est quantifié sur la base d'une prédiction de type MA d'ordre 4. La prédiction est réalisée sur la base des valeurs de l'énergie des vecteurs $z(n)$ des quatre sous-trames passées (dans le domaine logarithmique). On peut exprimer le gain optimum β comme suit :

$$\beta = \gamma \beta', \tag{3.72}$$

où β' et γ sont respectivement le gain prédit et un facteur de correction. C'est le facteur de correction qui est quantifié pour la transmission.

Soit l'énergie moyenne du mot de code du dictionnaire innovateur, en décibels :

$$E = 10\log_{10}\left(\frac{1}{N}\sum_{n=0}^{N} c(n)^2\right), \tag{3.73}$$

où N est la longueur d'une sous-trame de signal. Après normalisation du vecteur $c(n)$ avec le gain β, l'énergie de la contribution innovatrice $z(n)$ est donnée par l'expression $20\log_{10}\beta + E$. Soit $E^{(m)}$, cette même énergie après suppression de la moyenne des contributions innovatrices $\overline{E}$, à la sous-trame m, donnée par :

$$E^{(m)} = 20\log_{10}\beta + E - \overline{E}, \tag{3.74}$$

où $\overline{E}$ est calculée statistiquement. Le gain β peut être exprimé en fonction des énergies $E^{(m)}$, E et $\overline{E}$ par :

$$\beta = 10^{\left\lfloor E^{(m)} - E + \overline{E} \right\rfloor / 20}. \tag{3.75}$$

On détermine le gain prédit β' en prédisant par moyenne glissante (MA), l'énergie logarithmique de la contribution innovatrice actuelle, d'après l'énergie logarithmique des quatre précédentes contributions innovatrices. L'énergie prédite est donnée par :

$$\tilde{E}^{(m)} = \sum_{i=1}^{4} b_i \hat{U}^{(m-i)}, \tag{3.76}$$

où les b_i sont les coefficients de prédiction par moyenne glissante et où $\hat{U}^{(m)}$ est la version quantifiée de l'erreur de prédiction donnée par :

$$U^{(m)} = E^{(m)} - \tilde{E}^{(m)}. \tag{3.77}$$

On trouve le gain prédit β' en remplaçant $E^{(m)}$ par sa valeur prédite dans l'équation (3.75) :

$$\beta' = 10^{\left\lfloor \tilde{E}^{(m)} - E + \overline{E} \right\rfloor / 20}. \tag{3.78}$$

Le facteur de correction γ est associé, comme suit, à l'erreur sur la prédiction de gain :

$$U^{(m)} = E^{(m)} - \tilde{E}^{(m)} = 20\log(\gamma). \tag{3.79}$$

Dans le codeur G.729, le gain du dictionnaire adaptatif G et le facteur de correction γ sont quantifiés vectoriellement comme introduit à la Section 3.6.

3.6 Quantification vectorielle des gains

Pour réaliser une quantification vectorielle du gain du dictionnaire adaptatif G et du gain du dictionnaire innovateur β, on procède comme suit. Le dictionnaire de gains codés est exploré de façon à minimiser l'erreur quadratique pondérée D entre le signal de parole original et le signal de parole reconstruit. D est donnée par :

$$D = \mathbf{x}^t\mathbf{x} + G^2\mathbf{x}_a^t\mathbf{x}_a + \beta^2\mathbf{x}_b^t\mathbf{x}_b - 2G\mathbf{x}^t\mathbf{x}_a - 2\beta\mathbf{x}^t\mathbf{x}_b + 2G\beta\mathbf{x}_a^t\mathbf{x}_b, \tag{3.80}$$

où $x(n)$ est le signal cible, $x_a(n)$ est l'excitation adaptative filtrée par le filtre de synthèse pondéré perceptuellement, ou contribution adaptative, et $x_b(n)$ est l'excitation innovatrice filtrée par le filtre de synthèse pondéré perceptuellement, ou contribution innovatrice.

3.7 Mise à jour des mémoires des filtres de synthèse et de pondération

La mise à jour des états des filtres de synthèse et de pondération de l'encodeur est nécessaire pour calculer le signal cible de la sous-trame suivante. Une fois les deux gains quantifiés on obtient le vecteur d'excitation total $u(n)$ au moyen de la relation suivante :

$$u(n) = G' r_T(n) + \hat{\beta} c(n), \quad n = 0,....,N. \tag{3.81}$$

où G' est le gain quantifié du dictionnaire adaptatif, $\hat{\beta}$ est le gain quantifié du dictionnaire innovateur, $c(n)$ est l'excitation innovatrice et $r_T(n)$ est l'excitation adaptative correspondant à un délai tonal T (cf. Sous-section 3.4.1). On peut mettre à jour les états des filtres en faisant passer les différentes excitations multipliées par leur gain respectif dans les filtres $1/\hat{A}(z)$ et $W(z)$ qui leur correspondent.

3.8 Post-filtrage [3-23]

Le post-filtrage est réalisé pour rehausser la qualité auditive du signal de sortie. En effet, plus le débit de transmission est bas, plus la qualité du signal reconstruit est dégradée. Le post-filtrage permet d'accentuer les formants de la parole reconstruite. Un post-filtrage $F(z)$, couramment utilisé est donnée par :

$$F(z) = \frac{A(z/\lambda)}{A(z/\nu)} = \frac{1 - \sum_{k=1}^{p} a_k \lambda^k z^{-k}}{1 - \sum_{k=1}^{p} a_k \nu^k z^{-k}}, \tag{3.82}$$

où $0 \le \lambda \le \nu \le 1.0$, et où p et les coefficients a_k sont respectivement l'ordre et les coefficients du filtre de synthèse LPC. Le spectre du filtre $F(z)$ est contrôlé par les valeurs de λ et ν. Il faut prendre garde au fait que ce filtre présente une pente spectrale qui agit comme un filtre passe-bas. Par conséquent, ce filtre étouffe légèrement le signal de parole ainsi filtré.

Les valeurs idéales de λ et ν sont fonction du débit utilisé. Pour éliminer l'effet de la pente spectrale on peut utiliser un filtre de la forme :

$$F(z) = (1 - \mu z^{-1}) \frac{A(z/\lambda)}{A(z/\nu)}, \tag{3.83}$$

où μ est un coefficient fixe ou adaptatif. Si μ est adaptatif, il s'adapte aux caractéristiques du signal.

On peut également utiliser un post-filtrage basé sur la périodicité du signal, de la forme :

$$\frac{1}{P'(z)} = \frac{1}{1 - \eta G' z^{-T}}, \tag{3.84}$$

où G' est le gain quantifié du dictionnaire adaptatif et T le délai tonal.

En général, le post-filtrage cause une amplification du signal de parole et il est nécessaire d'employer une technique de contrôle du gain pour éviter toute amplification sur le signal de sortie.

3.9 Délai de traitement d'un codeur [3-27]

Le délai de traitement d'un codeur est important puisqu'il indique si le codeur peut être utilisé pour des applications en temps réel. Un délai de traitement trop long peut devenir désagréable.

Le délai d'un codeur de parole correspond à l'intervalle de temps mesuré entre l'application d'un échantillon de signal à l'entrée de l'encodeur, et l'observation de cet échantillon produit à la sortie du décodeur. Ainsi, le délai de codage est donnée par la somme des délais suivants :

- **Le délai de mémorisation algorithmique au niveau de l'encodeur** : ici l'encodeur travaille sur la base de trames de signal, et doit donc attendre la

mise en mémoire des échantillons d'une trame entière de signal, avant de commencer l'encodage des échantillons de celle-ci.

- **Le délai de traitement au niveau de l'encodeur** : les codeurs de parole utilisent typiquement la longueur d'une trame, en temps, pour traiter les échantillons mis en mémoire.
- **Le délai de transmission** : le délai de transmission est fonction du type de canal utilisé et peut donc fortement varier. Il comprend le délai dû au codage de canal pour la protection d'erreurs et le délai réel de transmission.
- **Le délai de mémorisation algorithmique au niveau du décodeur** : ici l'encodeur travaille sur la base des trames du signal, et doit le reconstruire trame par trame.
- **Le délai de traitement au niveau du décodeur** : celui-ci est généralement très inférieur à celui de l'encodage.

Si le délai de transmission entre l'encodeur et le décodeur n'est pas considéré on estime le délai de traitement d'un codeur, à environ 3.5 fois la durée d'une trame de signal. Cette estimation a été faite pour un codeur en bande étroite. Ce délai peut être réduit en diminuant le temps de traitement : il faut pour cela veiller à réduire la complexité de l'algorithme d'encodage.

Certains codeurs contiennent un algorithme de recouvrement de trames, qui fonctionne si une partie ou la totalité de l'information relative à une trame de signal, est perdue en cours de transmission. Naturellement, un tel algorithme ajoute un délai de traitement supplémentaire au codeur.

La limite du délai de traitement d'un codeur, pour une conversation en temps réel sur un système multimédia, est comprise entre 400 et 600 ms en l'absence d'échos [3-28]. Pour les systèmes où un écho risque d'apparaître, alors le délai de traitement doit être inférieur à 100 ms.

3.10 Résumé du chapitre et conclusions

Ce chapitre a présenté les concepts théoriques d'un codeur CS-ACELP ainsi que leur description algorithmique et mathématique.

Les principes des algorithmes d'encodage et de décodage d'un tel codeur ont été introduits à la Section 3.2. La Section 3.3 a décrit les algorithmes relatifs à l'extraction et à la quantification des paramètres LPC, qui permettent de modéliser l'enveloppe spectrale du signal de parole. Une méthode permettant de mesurer les performance de ces algorithmes a été donnée à la Sous-section 3.3.3. Les Sections 3.4 et 3.5 ont décrit la modélisation du vecteur d'excitation du filtre LPC par extraction des mots de code des

dictionnaires adaptatif et innovateur, et par extraction de leur gain respectif. La Section 3.6 a traité de la quantification vectorielle de ces gains.

La Section 3.7 a décrit comment les mémoires des filtres de synthèse LPC et de pondération perceptuelle sont mis à jour pour le passage d'une sous-trame de signal à une autre. Le post-filtrage, permettant un rehaussement de la qualité du signal reconstruit, a été décrit à la Section 3.8. Finalement, le problème du délai de traitement d'un encodeur de parole a été introduit à la Section 3.9.

Les concepts théoriques et leurs descriptions algorithmiques présentés au cours de ce chapitre vont permettre de décrire le codeur propriétaire, développé dans le cadre de ce travail de thèse. Celui-ci est présenté aux Chapitres 5 et 6.

3.11 Références

[3-1] K. Paliwal et W. Kleijn, "Quantization of LPC parameters", Chapter 12, dans *Speech coding and synthesis*, pp. 433-466, édité par W. Kleijn et K. Paliwal, Elsevier, Amsterdam, 1995.

[3-2] G. Guibé, H. How et L. Hanzo, "Comparative study of wideband speech spectral quantization schemes", dans *Proc. 3rd ITG Conference on source and channel coding*, pp. 181-186, Munich, Allemagne, Jan. 2000.

[3-3] M. Ferhaoui et S. Van Gerven, "LSP quantization in wideband speech coders", dans *Proc. IEEE Workshop on speech coding proceedings 1999,* pp. 25-27, Porvoo, Finlande, Juin 1999.

[3-4] S. Grassi, "Line spectrum pairs and the CELP FS1016 speech coder", Chapter 5, dans *Optimized implementation of speech processing algorithms*, pp. 73-112, Thèse publiée par la Faculté des Sciences de l'Université de Neuchâtel, Neuchâtel, 1998.

[3-5] Y. Bistritz et S. Peller, "Immittance spectral pairs (ISP) for speech encoding", dans *Proc. IEEE International conference on acoustics, speech and signal processing 1993, ICASSP'93,* Vol. 2, pp. 9-12, Minneapolis, USA, Avril 1993.

[3-6] L. R. Rabiner et R. W. Schafer, "Digital models for the speech signal", Chapter 3, dans *Digital processing of speech signals*, pp. 38-115, Prentice-Hall Signal Processing Series, Alan V. Oppenheim Series Editor, 1978.

[3-7] B. Atal, R. Cox, et P. Kroon, "Spectral quantization and interpolation for CELP coders", dans *Proc. IEEE International conference on acoustics, speech and signal processing 1989*, *ICASSP'89,* Vol.1, pp. 69-72, Mai 1989.

[3-8] K. Paliwal, "Interpolation properties of linear prediction parametric representations", dans *Proc. 4th European conference on speech communication and technology, EUROSPEECH'95*, Vol. 2, pp. 1029-1032, Madrid, Espagne, Sep.1995.

[3-9] F. Itakura, "Line spectrum representation of linear predictive coefficients of speech signals", dans *J. of the acoustical society of America*, Vol. 57, pp. S35, 1975.

[3-10] A. Kondoz, "LPC parameter quantization using LSFs", Chapter 4, dans *Digital speech. Coding for low bit rate communication systems,* pp. 79-115, édité par John Wiley & Sons, Chichester, Grande Bretagne, 1994.

[3-11] F. Soong et B. Juang, "Line spectrum pair (LSP) and speech data compression", dans *Proc. IEEE Int. Conf. on acoustics, speech, and signal processing*, *ICASSP'84*, pp. 1.10.1-1.10.4, 1984.

[3-12] C.-F. Chan, "Efficient quantization of LPC parameters using a mixed LSP/PARCOR representation", dans *Signal processing VII : Theories and applications*, pp. 939-942, M. Holt & Co. Eds., 1994.

[3-13] K. Paliwal et B. Atal, "Efficient vector quantization of LPC parameters at 24 Bits/Frame", dans *IEEE Trans. on speech and audio processing*, Vol. 1, No 1, pp. 3-14, Jan. 1993.

[3-14] F. Soong et B. Juang, "Optimal quantization of LSP parameters", dans *IEEE Trans. on speech and audio processing*, Vol. 1, No. 1, pp. 15-24, 1993.

[3-15] S. Lloyd, "Least squares quantization in PCM", dans *IEEE Transactions on information theory*, Vol. IT-28, No 2, pp. 129-137, Mars 1982.

[3-16] K. Paliwal et B. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame", *IEEE Trans. on speech and audio processing*, Vol. 1, No. 1, pp. 3-14, 1993.

[3-17] W. LeBlanc B. Bhattacharya et S. Mahmoud, "Efficient search and design procedures for robust multi-stage VQ of LPC parameters for 4 kbps speech coding", dans *IEEE Trans. on speech and audio processing*, Vol. 1, No. 4, pp. 373-385, Oct. 1993.

[3-18] N. Phamdo et N. Farvardin, "Coding of speech LSP parameters using TSVQ with interblock noiseless coding", dans *Proc. IEEE Int. conf. on acoustics, speech, and signal processing*, *ICASSP'90* , Vol. 1, pp.193-196, Avr. 1990.

[3-19] J. Skoglund et J. Lindén, "Predictive VQ for noisy channel spectrum coding : AR or MA", dans *Proc. IEEE Int. conf. on acoustics, speech, and signal processing, ICASSP'97*, Vol. 2, pp. 1351-1354, 1997.

[3-20] R. Ramachandran, M. Sondhi, N Seshadri et B. Atal, "A two codebook format for robust quantization of line spectral frequencies", dans *IEEE Trans. on speech and audio processing*, Vol. 3, No. 3, pp. 157-168, Mai 1995.

[3-21] S. Saito et K. Nakata, *Fundamentals of speech signal processing,* Chapter 9*,* Academic Press, New York, 1985.

[3-22] P. Kabal et P. Ramachandran, "The computation of line spectral frequencies using chebyshev polynomials", dans *IEEE Trans. on acoustics, speech and signal processing*, Vol. 34, No. 6, pp. 1419-1426, 1986.

[3-23] R. Salami, L. Hanzo, R. Steele, K. Wong et I. Wassell, "Speech coding", Chapter 3, dans *Mobile radio communications*, pp. 186-346, Raymond Steele Ed. Pentech Press Publishers, Londres, Grande Bretagne, 1992.

[3-24] C. Laflamme, J.-P. Adoul, R. Salami, S. Morissette et P. Mabileau, "16 kbps wideband speech coding technique based on algebraic CELP" dans *Proc. IEEE Int. conf. on acoustics, speech, and signal processing, ICASSP'91*, Vol. 1, pp.13-16, Mai 1991.

[3-25] J.Ashley, E. Cruz-Zeno, U. Mittal et W. Peng, "Wideband coding of speech using a scalable pulse codebook", dans *Proc. IEEE Workshop on speech coding*, pp. 148-150, Delavan, Wisconsin, USA, Sept. 2000.

[3-26] L. Hanzo, F. Somerville et J. Woodard, "Backward-adaptive code excited linear prediction", Chapter 8, dans *Voice compression and communications*, pp. 279-338, IEEE Series on Digital & Mobile Communication, John Wiley & Sons, Inc., Publication, NY, USA, 2001.

[3-27] A. Kondoz, "Low delay speech coding", Chapter 7, dans *Digital speech. Coding for low bit rate communication systems,* pp. 215-237, édité par John Wiley & Sons, Chichester, Grande Bretagne, 1994.

[3-28] A. Hardman et S. Hailes, "Mobile multimedia access for the Internet ", Chapter 8, dans *Insights into Mobile multimedia communications,* pp. 111-132, édité par D. Bull, N. Canagarajah et A. Nix, Academic Press, Londres, Grande Bretagne, 1999.

Chapitre 4
Etat de l'art

4.1 Introduction

Ce chapitre présente une étude de l'état de l'art pour le codage de la parole en bande élargie. Cette étude montre les principales tendances de ce type de codage et est orientée par le projet de recherche présenté ici. Elle ne traite que les articles de la littérature scientifique, qui décrivent des codeurs respectant dans les grandes lignes les contraintes fixées pour le codeur développé : le P-MRWB-ACELP (Proprietary Multi-Rate Wide-Band ACELP). Ces contraintes concernent le débit de transmission, la qualité du signal reconstruit, le délai de traitement et la complexité algorithmique. Elles sont inspirées par celles imposées par l'ETSI pour le processus de standardisation lancé en 1999, visant à obtenir un codeur à débit variable pour le signal de parole en bande élargie (cf. Section 1.1). Les contraintes imposées par l'ETSI sont décrites à la Section 4.2

La Section 4.3 présente et discute l'étude de l'état de l'art jusqu'en 1999. Cette étude a été réalisée au commencement de ce travail de recherche, afin de comprendre quel type de codeur pouvait respecter les contraintes sus-mentionnées. La Section 4.4 présente et discute l'évolution de l'état de l'art de l'année 2000 à l'année 2002. Elle permet une comparaison ultérieure avec le codeur développé (cf. Chapitre 8).

Bien que l'état de l'art pour le codage de la parole en bande élargie ne soit pas aussi riche en publications que celui pour la parole en bande étroite, il existe de nombreux articles décrivant un tel codage. Pour chacun des articles présentés, lorsque l'on parle de "qualité", il s'agit de la qualité du signal reconstruit que les auteurs disent obtenir avec le codeur décrit. Cette qualité est généralement donnée en comparaison avec la qualité d'un standard. Les auteurs des articles se réfèrent principalement au standard G.722 [4-1], mais également aux standards G.722.1 [4-2] de l'ITU, et MPEG4 (Moving Picture

Experts Groupe - 4) version 1 : le CELP/RPE [4-3]. Ces trois standards conviennent aussi bien au codage de la parole en bande élargie qu'au codage de la musique.

Le standard G.722 est le plus ancien standard pour le codage de la parole en bande élargie. Le codeur qu'il définit traite le signal en deux sous-bandes de fréquences (0-4 et 4-8 kHz) et encode chaque sous-bande en utilisant un algorithme ADPCM (Adaptive-Differential Pulse-Code Modulation) [4-4]. Ce codeur fonctionne selon trois modes, correspondant à des débits de 64, 56 et 48 kbits/s. On qualifie ici les modes de fonctionnement du G.722 par les sigles G.722 A, G.722 B et respectivement G.722 C. Ce standard sert de référence à tous les articles parus jusqu'à 1999.

Le standard G.722.1 a été sélectionné au cours de l'année 1999 par l'ITU. Il décrit un algorithme de codage de parole fonctionnant à des débits compris entre 24 et 32 kbits/s. Il est recommandé pour des applications mains libres, ne comprenant qu'une faible probabilité de perte de l'information des trames du signal lors de la transmission. Il permet de changer le débit de transmission à chaque nouvelle trame de 20 ms. Il effectue un codage par transformée à chevauchement et modulation (MLT, Modulated Lapped Transform).

L'algorithme de codage pour la parole en bande élargie du MPEG-4, est de type CELP/RPE (Regular Pulse Excitation). Il fonctionne à des débits compris entre 12 et 24 kbits/s.

4.2 Contraintes et performances requises pour le codeur WB–AMR de l'ETSI

Le document [4-5] décrit les contraintes et les performances requises initialement par l'ETSI pour le nouveau standard WB-AMR. Celui-ci devait avoir au moins deux modes (ici appelés mode 1 et mode 2) permettant une implantation sur le GSM-FR, qui fonctionne à un débit total de 22.8 kbits/s (codage de source et de canal) [4-6]. Le mode 1 devait en plus satisfaire la contrainte de permettre le transfert des bits du codage de source à travers l'interface A-ter du GSM. Celle-ci ne supporte qu'un débit total de 16 kbits/s. Ainsi le mode 1 du WB-AMR a été défini comme ne devant pas excéder 14.25 kbits/s [4-7]. Le mode 2 pouvait dépasser un débit de 16 kbits/s. En l'absence d'erreurs de transmission du canal, les modes 1 et 2 doivent permettre d'obtenir une qualité du signal reconstruit similaire à celle du G.722 C et respectivement à celle du G.722 B. Toutefois, l'objectif était d'atteindre la qualité du G.722 B et respectivement celle du G.722 A. Pour tous les autres modes, la qualité requise était celle du G.722 A. De plus, la complexité de calcul, la taille de la mémoire RAM, la taille de la mémoire ROM et le nombre d'opérations de base définies par l'ETSI, ne devaient pas

dépasser 40 wMOPS (weighted Million Operations Per Second), 15000 mots, 18000 mots et respectivement 5821 opérations de base ETSI [4-5].

4.3 Etat de l'art jusqu'en 1999

Cette section discute l'étude de l'état de l'art jusqu'en 1999, pour le codage de la parole en bande élargie. Elle traite diverses publications scientifiques, qui décrivent des codeurs respectant dans les grandes lignes les contraintes et performances sus-mentionnées. Ces publications sont présentées à l'Annexe C.1.

L'état de l'art fait apparaître trois catégories de codeurs : les codeurs de type CELP, les codeurs par transformée et les codeurs mixtes. Les codeurs de type CELP se divisent en deux sous-catégories : les codeurs encodant le signal en une seule bande de fréquences, et les codeurs séparant le signal en plusieurs sous-bandes de fréquences avec encodage séparé de chaque sous-bande. Ces derniers sont dénotés ici par l'acronyme SB-CELP (Subband-CELP ou Splitband-CELP) alors que les premiers sont qualifiés de CELP. Les codeurs mixtes sont basés sur un codeur CELP, mais ils utilisent un codage partiel par transformée.

L'état de l'art jusqu'à l'année 1999 est présenté à l'Annexe C.1 par catégories et sous-catégories. A l'intérieur d'une même catégorie les articles sont classés selon leur ordre chronologique de parution. Le Tableau 4-1 permet de comparer différentes caractéristiques des codeurs présentés à l'Annexe C.1. La complexité algorithmique des codeurs présentés n'est malheureusement donnée par les auteurs qu'en de rares exceptions et ne peut donc pas être comparée.

4.3.1 Résumé

Le Tableau 4-1 regroupe les différentes caractéristiques des codeurs présentés à l'Annexe C.1. Ces caractéristiques sont le débit, l'ordre de prédiction linéaire à court-terme (LPC), la durée d'une trame et d'une sous-trame de signal, le nombre de bits nécessaires à la quantification des LPC d'une trame, le type de quantification utilisée pour les paramètres LPC, la gamme de valeurs sur laquelle le délai tonal est recherché (cette gamme est calculée à 16 kHz) et finalement la qualité du signal reconstruit par comparaison à la qualité des différents modes du codeur G.722. Le numéro des codeurs (N°) est indiqué dans l'ordre chronologique de leur présentation à l'Annexe C.1. Certaines rubriques du tableau sont vides (–) car les informations correspondantes ne sont pas décrites dans l'article de référence. Les abréviations suivantes sont utilisées :

- pTR : par transformée;
- * : à 16 kHz;
- * *: selon référence [C-4];
- (a) : indique la durée d'une sous-trame utilisée pour la recherche de l'excitation adaptative;
- (i) : indique la durée d'une sous-trame utilisée pour la recherche de l'excitation innovatrice;
- bw : "backward";
- BI : bande de fréquences inférieure;
- BS : bande de fréquences supérieure;
- MA-n : prédiction par moyenne glissante d'ordre n;
- VQ : quantification vectorielle;
- MSVQ : quantification vectorielle sur plusieurs étages;
- NU : quantification non-uniforme;
- SVQ : quantification vectorielle séparée en sous-vecteurs;
- nT_s : nombre de périodes d'échantillonnage.

Le Tableau 4-1 montre une tendance pour l'implantation de codeurs CELP, SB-CELP et mixtes, mais il présente également un codeur par transformée et un codeur de type MP-Dec.

On constate que s'ils fonctionnent à un débit inférieur ou égal à 16 kbits/s, sept des huit codeurs de type CELP en une seule bande de fréquences ont une "qualité" comprise entre celle du G.722 A et celle du G.722 C. La "qualité" du codeur N° 1, n'est malheureusement pas comparée à celle du G.722. De plus, s'ils fonctionnent à un débit inférieur ou égal à 16 kbits/s, cinq des sept codeurs de type SB-CELP ont la "qualité" du G.722 C. La "qualité" du codeur N° 9 n'est pas comparée à celle du G.722, alors que celle du codeur N° 10 ne l'est que partiellement. On conclut de ce qui précède que les codeurs CELP et les SB-CELP peuvent remplir les contraintes en qualité et en débit décrites à la Section 4.2.

La littérature présente cinq codeurs mixtes, ayant au moins un mode de fonctionnement à un débit inférieur ou égal à 16 kbits/s. Malheureusement, la "qualité" de ces codeurs, n'est comparée que dans un cas à celle du G.722. Il est ainsi difficile de juger quelle qualité de codage peut être obtenue avec un type de codeur donné.

Le codeur par transformée, présenté par Moriya et *al.* en [C-17] (cf. Annexe C) fonctionne à 16 kbits/s. Sa "qualité" est celle du G.722 C. Ce codeur présente le désavantage d'introduire un délai de traitement élevé pour une application bi-directionnelle en temps réel (communication en "full duplex").

Type de codeur	N° du codeur et premier auteur	Réfé-rence	Débit [kbits/s]	Ordre de prédiction LPC	Durée d'une trame [ms]	Durée d'une sous-trame [ms]	Nb. de bits pour les QLPC	Type de quantification pour les coefficients LPC	Délai tonal [nT_s]	Qualité de reconstruction auditive
CELP	**1. Laflamme**	[C-1]	13	16	15	5	–	–	40-295	Haute
	2. Salami	[C-2]	9.6 & 14	16	30	6	54	–	40-295	G.722 B (14 kbits/s)
	3. Harborg	[C-3]	16	16 à 20	20	2	60 à 80	Scalaire**	32-287**	G.722 A
	4. McElroy	[C-5]	16 & 24	20	25	3.125 & 5	70	–	–	G.722 B-C (16 kbits/s)
	5. Black	[C-6]	14.1	16	20	2.5	60	NU, scalaire	–	G.722 C
	6. Sasaki	[C-7]	16	20	10	5(a)/2.5(i)	36	MA-4, MSVQ, SVQ	33-288	G.722 B
	7. Serizawa	[C-8]	16	16	10	5(a)/2.5(i)	36	–	–	G.722 B
	8. Koishida	[C-9]	16	20	10	2.5	21	MA-5, MSVQ, SVQ	33-224	G.722 A
SB-CELP	**9. Roy**	[C-10]	16	16	20	2.5	48	NU, différentielle, scalaire	–	–
	10. McElroy	[C-11]	7.2 -14.4	10 (BI) / 2 (BS)	25-11	–	?/6(BS)	BS : NU/scalaire	–	G.722 B ou inférieure
	11. Paulus	[C-12]	16	14	10	5(a)/2.5(i)	44	MA, SVQ	40-257*	G.722 C
	12. Ubale	[C-13]	16	16	10	2.5	28	MA-2, MSVQ	41-296	G.722 C
	13. Schnitzler	[C-14]	13	14	20	5(a)/2.5(i)	32	MA-4, MSVQ, SVQ	–	G.722 C
	14. Combescure	[C-15]	16 et 24	12/8 (BS)	20	5(a)/2.5(i)	33 / 10 (BS)	MA, MSVQ, SVQ / VQ (BS)	–	G.722 C/B (16/24 kbits/s)
	15. Black	[C-16]	16	10 bw, (BI) / 6 (BS)	1.75 (BI) / 7(BS)	1.75 (BI) / 7 (BS)	16	BI : pas de quantification; BS : SVQ	–	G.722 C
pTR	**16. Moriya**	[C-17]	16	16	8-64	–	19	MA, MSVQ, SVQ	–	G.722 C
Mixtes	**17. Lefebvre**	[C-18]	16 & 24	16	24	6	48	SVQ	40-295	Très bonne
	18. Xie	[C-19]	16	12	24	6	48	SVQ	40-295	Plus que très bonne
	19. Salami	[C-20]	16 & 24	16	10	5	34	MA, SVQ	29-281	G.722 C/B (16/24 kbits/s)
	20. Chen	[C-21]	16 - 32	16	20	4	49	SVQ	–	Haute
	21. Schnitzler	[C-22]	15.5 & 20	16/44 bw	20	4	43	MA, MSVQ, SVQ	–	–
Autre	**22. Abreu**	[C-23]	16-33	14	16	4 ou 2	–	–	40-294	G.722 A (33 kbits/s)

Tableau 4-1 : Etat de l'art jusqu'en 1999 : Type de codeur, N° du codeur et premier auteur et référence (selon l'Annexe C.1), débit, ordre de prédiction LPC, durée d'une trame, durée d'une sous-trame, nombre (nb.) de bits attribués à la quantification des LPC (QLPC) pour une trame, type de quantification pour les LPC, gamme de recherche du délai tonal et "qualité" telle qu'estimée dans la référence.

Finalement, la "qualité" du signal reconstruit obtenue avec le codeur MP-Dec, présenté par Abreu et Docampo, est inférieure à celle du G.722 C pour son mode à 16 kbits/s. Ce codeur ne semble pas pouvoir remplir les contraintes en qualité et en débit décrites à la Section 4.2.

4.3.2 Discussion

En conséquence de ce qui précède et en considérant le rapport débit - qualité, les codeurs de type CELP en une seule bande de fréquences et les codeurs SB-CELP semblent les plus adaptés pour réaliser un codage du signal de parole en bande élargie. Il serait intéressant de confronter non seulement leur rapport débit – qualité, mais également leur complexité algorithmique et leur délai de traitement. Cependant, la complexité algorithmique n'est donnée par les auteurs qu'en de rares exceptions. Toutefois, il est dit en [4-8] qu'un codeur de type CELP repris de la bande étroite et adapté à la bande élargie, souffre d'une complexité très élevée. Un codeur SB-CELP permet de réduire cette complexité de calcul mais augmente le délai de traitement, ce qui peut limiter le traitement en temps réel. En effet, comme l'on n'a pas à faire à un codeur d'onde, on doit absolument éviter les effets de recouvrement dus au filtrage en sous-bandes. Ainsi, il faut utiliser un filtre sélectif de degré suffisant pour bien atténuer les bandes de fréquences coupées. Un tel filtre introduit un délai de traitement.

Par rapport aux codeurs CELP en une seule bande de fréquences, les codeurs SB-CELP sont moins complexes, à débit égal ils donnent une "qualité" inférieure et ils ont un délai de traitement plus long (à moins de réaliser un filtrage "off-line" comme le font Ubale et Gersho avec le codeur N° 12 de l'Annexe C.1). Or, si la complexité est le principal obstacle pour un codeur CELP en une seule bande de fréquences, l'utilisation d'un codeur de type ACELP permet de réduire considérablement la complexité de calcul par rapport à un codeur CELP standard.

Un codeur CELP standard possède un dictionnaire innovateur composé de vecteurs d'excitation stochastique fixes, dont chaque élément est généré par un nombre gaussien aléatoire. Par rapport à un tel codeur, le dictionnaire algébrique d'un codeur ACELP permet de réduire considérablement la complexité de calcul relative à l'extraction de l'excitation innovatrice (cf. Sous-section 2.3.5). En effet, un tel dictionnaire est composé en grande majorité de valeurs nulles. Bien que pour des codeurs en bande élargie, la recherche dans un dictionnaire algébrique reste très complexe, des approches sous-optimales telles qu'une recherche focalisée ou un ré-ordonnement d'impulsions [4-9], permettent une diminution de la complexité

algorithmique. 8 des 22 codeurs, présentés dans la revue de l'état de l'art de l'Annexe C.1 et discutés ci-dessus, utilisent un algorithme ACELP.

4.4 Evolution de l'état de l'art de 2000 à 2002

Cette section discute l'étude de l'état de l'art des années 2000 à 2002, pour le codage de la parole en bande élargie. Elle traite diverses publications scientifiques, qui décrivent des codeurs respectant dans les grandes lignes les contraintes et performances décrites à la Section 4.2. Ces publications sont présentées à l'Annexe C.2 par catégories et sous-catégories. A l'intérieur d'une même catégorie, les articles sont classés selon leur ordre chronologique de parution. La Sous-section 4.4.1 décrit brièvement le nouveau standard WB-AMR de l'ETSI.

L'état de l'art décrit à l'Annexe C.2 fait apparaître une catégorie principale de codeurs : les codeurs de type SB-ACELP. De plus, on a un codeur de type MELP, deux codeurs par transformée et un codeur mixte combinant le codage CELP et le codage par transformée. Certains codeurs se prêtent aussi bien à l'encodage de la parole en bande élargie qu'à celui de la parole en bande étroite. Seule leur description pour la bande élargie est donnée à l'Annexe C.2.

4.4.1 Nouveau standard ETSI : le WB-AMR [4-10], [4-11], [4-12] et [4-13]

Le nouveau standard ETSI pour le codage à débit variable de la parole en bande élargie a été développé par les sociétés Nokia et VoiceAge. Ce standard décrit un codeur de type SB-ACELP, qui fonctionne à neufs débits, ou modes, différents. Ces neufs débits valent : 23.85, 23.05, 19.85, 18.25, 15.85, 14.25, 12.65, 8.85 et 6.6 kbits/s. Un mode supplémentaire, fonctionnant à 1.75 kbits/s, est destiné à encoder le bruit de fond contenu dans le signal, en cas de transmission discontinue (DTX : discontinuous transmission).

Les modes correspondant à un débit supérieur à 8.85 kbits/s offrent une haute qualité de signal reconstruit. Les débits les plus bas, soit 8.85 et 6.6 kbits/s ne sont utilisés que temporairement, lorsque les conditions du canal radio sont très mauvaises, ou pendant une congestion du réseau.

Le codage se fait en deux bandes de fréquences. La bande inférieure (50-6400 Hz) est encodée en utilisant un algorithme ACELP. Diverses particularités y sont ajoutées. Elles permettent d'obtenir une haute qualité subjective du signal reconstruit, même à des bas débits d'encodage. La bande

de fréquences supérieure (6400-7000 Hz) est reconstruite sur la base de l'information de la bande inférieure.

4.4.1.1 Bande de fréquences inférieure

Pour le codage de la bande de fréquences inférieure, le signal d'entrée est sous-échantillonné à 12.8 kHz, décimé et traité par trames de 20 ms. Le signal décimé est filtré passe-haut et pré-accentué. Le filtrage passe-haut sert de précaution contre les composantes indésirables en très basse fréquence. Le filtre de pré-accentuation est de la forme $H_{pré-acc}(z) = 1 - \mu_1 \cdot z^{-1}$. Il rehausse les hautes fréquences du signal. L'analyse par prédiction linéaire LPC est réalisée sur le signal pré-accentué. Les coefficients LPC sont transformés en ISP et quantifiés vectoriellement (cf. Sous-Section 5.6.2). Pour la recherche dans les dictionnaires adaptatif et fixe (algébrique), les trames sont séparées en 4 sous-trames de 5 ms.

Le filtre de pondération perceptuelle, donné généralement par l'équation (2.20), est de la forme :

$$W(z) = A(z/\gamma_1) \cdot H_{dé-acc.}(z) = \frac{A(z/\gamma_1)}{1 - \beta_1 z^{-1}}. \qquad (4.1)$$

$H_{dé-acc.}(z)$ compense la pré-accentuation du signal d'entrée utilisée pour extraire les coefficients LPC.

Le délai tonal est recherché en boucle ouverte toutes les 10 ms, et en boucle fermée toutes les 5 ms. La recherche en boucle ouverte se base sur la parole pondérée perceptuellement. Elle favorise les délais les plus bas et tient compte du caractère voisé, ou non voisé, des sous-trames traitées dans le passé. Exceptionnellement, à 6.6 kbits/s, elle ne s'effectue que toutes les 20 ms. Le délai tonal de la première et de la troisième sous-trames d'une trame est encodé avec 9 bits. La précision du délai tonal pour ces sous-trames est de ¼ de la période d'échantillonnage T_s, pour les valeurs de délai tonal comprises entre 34 et 127, de ½ T_s pour les valeurs comprises entre 128 et 159, et de 1 T_s pour les valeurs supérieures. Le délai tonal de la seconde et de la quatrième sous-trame est encodé de façon différentielle, en utilisant 6 bits, avec une précision de ¼ T_s.

Une fois le délai tonal en boucle fermée déterminé, le gain G_1 de l'excitation correspondante est calculé. De plus, un second gain G_2 est calculé sur la base de cette même excitation filtrée passe-bas (0-2.8 kHz). Dans le cas où $G_2 > G_1$, alors l'excitation filtrée est retenue pour la suite du codage. Un bit est nécessaire pour transmettre cette information au décodeur. Cette technique constitue une des particularités du codeur. Si le codeur fonctionne à un débit de 6.6 kbits/s, l'excitation adaptative est toujours filtrée

passe-bas. Une seconde particularité du codeur est que la borne supérieure du gain pour l'excitation adaptative, est fonction du caractère périodique ou non-périodique des sous-trames passées.

Le dictionnaire innovateur (fixe) est un dictionnaire algébrique de structure ISPP (cf. Sous-Section 3.5.1). Les 64 positions des vecteurs de code (5 ms) sont divisées en 4 pistes de positions entrelacées. Les différents débits du codeur sont construits en plaçant un nombre différent d'impulsions sur chaque piste. Ce nombre varie entre 1 et 6. A 6.6 kbits/s, on n'a que deux pistes et une impulsion par piste.

Une autre particularité du codeur est que pour la recherche de l'excitation innovatrice, le dictionnaire algébrique est suivi d'un filtre adaptatif $F(z)$, combiné au filtre de pondération perceptuelle $W(z)$. $F(z)$ se décompose en deux parties : $F_1(z)$ et $F_2(z)$. $F_1(z)$ permet d'améliorer la périodicité de l'excitation. Il est de la forme $1/(1-0.85\cdot z^{-T})$, où T est la partie entière du délai tonal. $F_2(z)$ accentue les hautes fréquences et est de la forme $(1-\mu_2\cdot z^{-1})$.

Lors de la recherche de l'excitation innovatrice, deux impulsions provenant de pistes différentes sont sélectionnées à chaque itération. L'une des impulsions ne peut prendre qu'un certain nombre de positions. Ces positions sont déterminées à l'aide d'un vecteur $\mathbf{b}$, dont les composantes $b(n)$ correspondent à une estimation de la probabilité qu'une impulsion occupe la position n.

Le gain du dictionnaire adaptatif et le facteur de correction γ (rapport entre le gain du dictionnaire algébrique et sa valeur prédite : cf. Sous-Section 3.5.2) sont quantifiés vectoriellement.

4.4.1.2 Bande de fréquences supérieure

L'excitation de la bande de fréquences supérieure est obtenue en générant un bruit blanc, dont la puissance est égalisée à celle de l'excitation de la bande de fréquences inférieure, puis en multipliant ce bruit par un facteur de gain g_{HB}. g_{HB} est basé sur la caractéristique voisée ou non-voisée du signal, ainsi que sur la pente spectrale du signal reconstruit de la bande inférieure. L'excitation est passée dans un filtre de synthèse LPC, dont les coefficients sont obtenus sur la base des ISF quantifiés de la bande inférieure. Finalement, la sortie du filtre de synthèse LPC est filtrée passe-bande (6.4-7 kHz), et donne le signal de la bande de fréquences supérieure.

Dans le cas du débit le plus élevé, le facteur de gain g_{HB} est calculé au niveau de l'encodeur. Il correspond au rapport entre l'énergie du signal en haute fréquence (6.4-7 kHz) et celle du signal de sortie du filtre de synthèse

LPC cité ci-dessus. Pour tous les autres débits, ce facteur de gain n'est calculé qu'au niveau du décodeur.

Les coefficients du filtre de synthèse LPC pour la bande de fréquences supérieure sont obtenus au niveau du décodeur comme suit. Pour le débit le plus bas, les 16 ISF quantifiés de la bande inférieure sont extrapolés en un vecteur de 20 ISF. Pour tous les autres débits, les 16 ISF quantifiés sont pondérés pour être adaptés à la conversion de la fréquence d'échantillonnage du signal, qui passe de 12.8 à 16 kHz. Ainsi, l'enveloppe de la bande spectrale comprise entre 5.1 et 5.6 kHz, avec un échantillonnage à 12.8 kHz, est appliquée à la bande de fréquences comprise entre 6.4 et 7 kHz, avec un échantillonnage à 16 kHz. La bande supérieure de fréquences est donc reconstruite au niveau du décodeur sur la seule base de la bande inférieure, à l'exception du cas où le débit vaut 23.85 kbits/s.

4.4.1.3 Post-traitement au niveau du décodeur

Diverses techniques de post-traitement sont appliquées au niveau du décodeur afin d'améliorer la qualité subjective du signal reconstruit. Il s'agit d'une part, d'un lissage du gain de l'excitation algébrique permettant d'améliorer cette excitation si le signal est bruité. D'autre part, une procédure de rehaussement du délai tonal modifie l'excitation totale, en filtrant l'excitation algébrique à travers un filtre d'innovation, dont la réponse en fréquence accentue les hautes fréquences et dont les coefficients sont liés à la périodicité du signal.

4.4.2 Résumé

Le numéro des codeurs (N°) indiqués dans cette section se réfèrent à la numérotation adoptée dans l'Annexe C.2.

L'étude de l'état de l'art de 2000 à 2002 fait apparaître une tendance pour l'utilisation de codeurs de type SB-ACELP. Le nouveau standard WB-AMR en fait partie. Pour une grande partie des codeurs SB-ACELP présentés à l'Annexe C.2, la bande inférieure (0-4 kHz) est encodée à l'aide d'un algorithme de codage de la parole en bande étroite, tel que le G.729 E, le GSM EFR, ou le NB-AMR. Certains codeurs reconstruisent aussi bien de la parole en bande élargie que de la parole en bande étroite.

En plus des codeurs de type SB-ACELP, l'état de l'art comprend un codeur de type ACELP en une seule bande de fréquences, un codeur MELP, deux codeurs par transformée et un codeur mixte combinant le codage CELP et le codage par transformée.

Le codeur ACELP N° 23 en une seule bande de fréquences permet d'obtenir à 16 kbits/s, soit la "qualité" du G.722 B, soit celle du G.722 C. Parmi les codeurs SB-ACELP, où la bande de fréquences inférieure est encodée à l'aide d'un algorithme pour la bande étroite et où la bande supérieure n'est pas encodée par transformée, les codeurs N° 25 et 28 permettent d'obtenir avec un débit de 16 kbits/s, la "qualité" du G.722 B et celle du G.722 A respectivement, alors que le codeur N° 27 permet d'obtenir la "qualité" du G.722 B pour un débit de 14 kbits/s. Pour les autres codeurs de ce type, la "qualité" n'est pas comparée au G.722. Cependant, le codeur N° 24 a été candidat au processus de standardisation WB-AMR de l'ETSI. On peut donc supposer qu'il respecte les contraintes et les performances décrites à la Section 4.2. Parmi les codeurs SB-ACELP, où la bande de fréquences inférieure est encodée à l'aide d'un algorithme pour la bande étroite et où la bande supérieure est encodée par transformée, le codeur N° 29 permet d'obtenir la qualité du G.722 B.

La qualité du codeur MELP à 8.4 kbits/s est celle du G.722 C, alors qu'aucune comparaison qualitative avec un standard n'est donnée pour les codeurs par transformée.

4.4.3 Discussion

L'évolution de l'état de l'art de 2000 à 2002 fait transparaître une forte tendance pour l'utilisation d'un algorithme ACELP, que ce soit pour le codage en une seule bande de fréquences, en deux bandes de largeurs égales (0-4 et 4-8 kHz), ou encore en deux bandes de largeurs inégales (nouveau standard WB-AMR et codeur N° 24 de l'Annexe C.2).

Les codeurs de type SB-ACELP, où la bande de fréquences inférieure est encodée à l'aide d'un algorithme pour la bande étroite sont majoritaires. L'utilisation de tels codeurs s'explique par la volonté des auteurs de réaliser des codeurs applicables aussi bien pour le codage de la parole en bande étroite que pour le codage de la parole en bande élargie. Cette volonté a certainement été influencée par le processus de standardisation de l'ETSI.

Le nouveau standard WB-AMR réduit la complexité de calcul relative à l'utilisation d'un codeur ACELP en une seule bande de fréquences, en filtrant le signal original à 6.4 kHz, et en traitant un signal échantillonné à 12.8 kHz. Ainsi, le nombre d'échantillons compris dans une sous-trame de signal de 5 ms peut être réduit de 80 pour une fréquence de 16 kHz, à 64. En utilisant un dictionnaire algébrique avec des pistes de 16 positions, seules 4 pistes sont nécessaires et non plus 5. Ceci permet de réduire la complexité de calcul et le débit, tout en ne négligeant que la qualité de la bande de fréquences comprise

entre 6.4 et 7 kHz, puisque pour le codage de la parole en bande élargie le signal original est filtré dans la bande [50-7'000 Hz]. La bande comprise entre 6.4 et 7 kHz ne contient que peu d'information et est reconstruite séparément.

4.5 Références

[4-1] ITU-T Recommendation G.722, "7 kHz audio – coding within 64 kbit/s", dans *Blue Book, fascicule III.4*, Melbourne, Australie, 1988.

[4-2] Recommandation ITU-T G722.1, "Codage aux débits de 24 et 32 kbit/s pour utilisation en mains libres sur les systèmes à faible perte de trame", Sept. 1999.

[4-3] F. Wuppermann, R. Taori et A. Gerrits, "A low complexity wideband-CELP for MPEG-4", dans *Proc. ITG-Fachbericht*, pp. 23-27, Aachen, Allemagne, Mars 1998.

[4-4] D. O'Shaughnessy, "Coding of speech signals", Chapter 7, dans *Speech communication; Human and machine*, Addison-Wesley publishing company, 1987.

[4-5] Document S4/SMG11 Tdoc 90/00, "AMR wideband performance requirements (WB-3) version 1.2", publié par le 3GPP, en : http://www.3gpp.org/ftp/tsg_sa/TSG_SA/TSGS_07/Docs/PDF/SP-000134.pdf (Sept. 2002).

[4-6] L. Hanzo, F. Somerville et J. Woodard, "Standard forward-adaptive CELP codecs", Chapter 7, dans *Voice compression and communications*, pp. 207-278, IEEE Series on Digital & Mobile Communication, John Wiley & Sons, Inc., Publication, NY, USA, 2001.

[4-7] Communication Email privée avec le Dr. Redwan Salami, VoiceAge (Sept. 2002).

[4-8] A. Ubale et A. Gersho, "A multi-band CELP wideband speech coder", dans *Proc. IEEE Int. conf. acoustic, speech and signal processing 1997, ICASSP'97*, Vol. 2, pp. 1367-1370, Munich, Allemagne, Avr. 1997.

[4-9] S. Pujalte et A. Moreno, "Wideband ACELP at 16 kbit/s with multi-band excitation", dans *Proc. European conference on speech communication and technology, EUROSPEECH 2001*, pp. 2001-2004, Aalborg, Danemark, 2001.

[4-10] http://www.mobilein.com/3GVoiceWhitepaper.pdf (29 Août 2002).

[4-11] J. Rotola-Pukkila, J. Vainio, H. Mikkola, K. Järvinen, B. Bessette, R. Lefebvre, R. Salami et M. Jelinek, "AMR wideband codec – leap in mobile communication voice quality", dans *Proc. European conference on speech communication and technology, EUROSPEECH 2001*, pp. 2303-2306, Aalborg, Danemark, Sept. 2001.

[4-12] B. Bessette, R. Lefebvre, R. Salami et M. Jelinek, J. Vainio, J. Rotola-Pukkila, H. Mikkola et K. Järvinen, "Techniques for high-quality ACELP coding of wideband speech", dans *Proc. European conference on speech communication and technology, EUROSPEECH 2001*, pp. 1997-2000, Aalborg, Danemark, Sept. 2001.

[4-13] Document 3GPP TS 26.190 V5.0.0 (2001–03), publié par le 3GPP en : ftp://ftp.3gpp.org/Specs/2001-09/Rel-5/26_series/ (14 Nov. 2001).

Chapitre 5
Choix du codeur développé et principales contributions

5.1 Introduction

Ce chapitre décrit la conception du codeur P-MRWB-ACELP (Proprietary Multi-Rate Wide-Band ACELP) et les contributions et innovations principales de ce travail de thèse.

Ce chapitre est articulé comme suit. La Section 5.2 présente les différentes étapes de développement qui ont mené à l'implantation du codeur de parole en bande élargie P-MRWB-ACELP, fonctionnant à trois débits différents. La Section 5.3 définit les contraintes proposées dans le cadre de ce travail de thèse pour la conception du codeur. Ces contraintes ont été définies comme un objectif à atteindre. La Section 5.4 motive les choix initiaux qui ont orienté l'implantation du codeur. Ces choix ont été basés sur les contraintes précitées et sur l'étude de l'état de l'art. La Section 5.5 décrit la méthodologie utilisée pour qualifier le codeur développé, ainsi que les bases de données créées pour entraîner et tester le codeur développé.

La Section 5.6 présente le développement et l'implantation d'un quantificateur vectoriel pour l'encodage des paramètres LPC. La Section 5.7 décrit trois innovations qui ont permis de réduire le bruit de reconstruction du signal, lié à un algorithme de type CELP. Ces innovations sont :

- L'introduction d'un pré-filtrage du dictionnaire adaptatif;
- Le contrôle du gain de l'excitation innovatrice;
- L'introduction de deux filtres de pondération formantique liés.

La Section 5.8 présente deux méthodes d'extraction de l'excitation adaptative alors que la Section 5.9 décrit brièvement les différents modes implantés pour l'extraction de l'excitation algébrique. Finalement, la Section 5.10 résume et discute les différentes innovations et contributions.

5.2 Etapes de développement

Cette section décrit les étapes de développement qui ont mené à l'implantation (en langage C) du codeur P-MRWB-ACELP. Ce codeur de parole en bande élargie fonctionne à trois débits différents.

La conception du codeur a débuté par l'établissement de contraintes à respecter, telles que le débit, la complexité, la qualité du signal reconstruit et le délai de traitement algorithmique. Ces contraintes ont été fixées en fonction des besoins du marché et dans l'optique d'une participation possible au processus de standardisation lancé par l'ETSI. La décision d'implanter un codeur de type ACELP en une seule bande de fréquences est basée sur ces contraintes, ainsi que sur l'état de l'art jusqu'en 1999 présenté et discuté au Chapitre 4 et à l'Annexe C.1.

L'implantation du codeur a débuté par la programmation en langage C des routines élémentaires d'un codeur ACELP. Ces routines sont utilisées par divers standards pour le codage de parole en bande étroite. Ainsi, les algorithmes du standard G.729 de l'ITU [1-2] (codeur de type ACELP fonctionnant à un débit de 8 kbits/s) ont été adaptés au codage de la parole en bande élargie. Ensuite, une méthodologie de test pour qualifier le signal de parole reconstruit a été développée.

Un effort conséquent a été déployé pour implanter un quantificateur des paramètres LSP. La méthodologie développée correspond à une contribution importante de cette thèse. Une fois la quantification des paramètres LSP définie, un algorithme de base pour l'extraction de l'excitation adaptative a été développé. Puis, l'algorithme d'extraction de l'excitation innovatrice de type algébrique a été implanté. Cet algorithme fonctionne selon trois modes correspondant à trois débits, complexités algorithmiques et qualités du signal reconstruit.

A ce stade du développement, il fallait ajouter au codeur une unité de pré-traitement ainsi qu'une unité de post-traitement. Cependant, quel que soit le mode de fonctionnement utilisé, le signal reconstruit était fortement bruité et de qualité insuffisante. Un réglage fin des paramètres de pondération perceptuelle et l'insertion des blocs de pré-traitement et post-traitement n'ont permis qu'une légère diminution des bruits contenus dans le signal reconstruit. La réduction de ces bruits a fait l'objet d'une grande partie de la recherche présentée ici.

L'origine des bruits de reconstruction est liée à l'extraction de l'excitation. Celle-ci est réalisée en boucle fermée et il est difficile d'isoler un problème particulier dans un algorithme bouclé. Pour réduire ces bruits, trois nouveaux concepts ont été introduits et différents blocs de base du codeur ACELP ont été modifiés. Toutefois, pour des voix de femmes, le signal reconstruit contenait encore une certaine rugosité. Ainsi, la prédiction à long-terme, ou extraction de l'excitation adaptative, a été traitée avec une grande attention. Diverses méthodes de recherche de l'excitation adaptative en boucle ouverte et en boucle fermée ont été étudiées. Finalement, deux modes d'extraction ont été retenus.

5.3 Contraintes

La première étape pour le développement du codeur a consisté en l'établissement d'un certain nombre de contraintes, permettant de définir le codeur à développer.

Les principaux attributs d'un codeur de parole, visant une utilisation bi-directionnelle en temps réel (communication en "full duplex"), sont :

- le débit de transmission;
- la qualité du signal reconstruit;
- le délai de traitement;
- la complexité algorithmique.

Pour la bande élargie, la qualité d'un codeur est couramment définie en comparant le signal reconstruit avec celui des trois modes du standard G.722 de l'IUT [5-3], qualifiés de G.722 A, G.722 B et respectivement G.722 C (cf. Section 4.1).

Les contraintes proposées dans le cadre de ce travail de thèse ont été définies comme un objectif à atteindre. Elles ont été inspirées par les besoins du marché et en particulier dans le processus de standardisation lancé par l'ETSI en 1999, visant à obtenir un codeur à débit variable pour le signal de parole en bande élargie (cf. Section 1.1). Les performances requises par l'ETSI sont décrites dans le document [5-2] et résumées à la Section 4.2. Cependant, comme l'étendue du travail nécessaire à les satisfaire et à les tester est trop conséquente, seules certaines contraintes ont été retenues. Ici, les performances du codeur pour un signal bruité n'ont pas été considérées. De même, les problèmes liés à la perte de l'information par le canal de transmission, et donc à la correction d'erreurs, n'ont pas été traités. De plus, aucun algorithme de transmission discontinue n'a été développé. Un tel algorithme permet une réduction considérable du débit de transmission

pendant les périodes où le signal ne contient pas de parole, mais uniquement un bruit de fond.

Ainsi, pour le développement du codeur présenté ici, les contraintes fixées initialement étaient les suivantes :

- **Débit de transmission** : le but étant d'obtenir un codeur à débit variable, une seule contrainte de débit a été fixée initialement : concevoir au moins un mode de fonctionnement, à un débit inférieur ou égal à 14.25 kbits/s. Ce mode aurait pu être intégré au GSM-FR.
- **Qualité du signal reconstruit** : le but étant d'obtenir un codeur à débit variable, nous nous proposions de concevoir au moins trois modes de fonctionnement, chacun d'entre eux permettant d'obtenir pour un signal de parole non-bruité, un signal reconstruit de qualité similaire à celle du G.722 A, B et respectivement C,
- **Délai de traitement** : l'objectif était une implantation future du codeur développé pour une application bi-directionnelle, en temps réel. Selon le document [5-2], pour les deux modes permettant une implantation sur le GSM-FR, un délai maximum de 125 ms au total est autorisé. Ce délai de traitement est celui du codeur NB-AMR fonctionnant à 12.2 kbits/s [5-4]. Il comprend les délais de mémorisation et de traitement algorithmique du codeur et du décodeur, ainsi que les délais d'entrelacement et de dé-entrelacement des bits pour le codage de canal. Le codeur NB-AMR référencé, traite des trames de signal de 20 ms, sans "look-ahead"[6]. Si le délai de transmission entre l'encodeur et le décodeur n'est pas considéré, on estime le délai de traitement d'un codeur à environ 3.5 fois la durée d'une trame de signal (cf. Section 3.5). Par rapport à un codeur en bande étroite, nous pouvons supposer que la complexité de calcul augmente considérablement pour un codeur en bande élargie. Or, le délai de traitement algorithmique d'un codeur est fonction de sa complexité. Cependant, l'évolution rapide des puces DSP doit permettre de ne pas augmenter ce délai de traitement même si la complexité augmente. Ainsi, une durée maximale de trame de 20 ms sans "look-ahead", doit convenir à une implantation en temps réel également pour un codeur en bande élargie.
- **Complexité de calcul et exigences en taille de mémoire** : la complexité de calcul définie pour le nouveau standard WB-AMR se réfère aux opérations de base définies par l'ETSI. Cependant, le codeur développé ici n'utilise pas ces opérations, et sa complexité ne peut leur être comparée. Toutefois, la complexité de calcul et les exigences en mémoire d'un

[6] Un codeur avec / sans "look-ahead" consiste à inclure / renoncer à une analyse anticipée des trames futures.

algorithme déterminent le coût et la consommation en puissance de la plate-forme sur laquelle il est implanté. Pour maintenir un faible coût et une basse consommation, les algorithmes de codage de la parole sont sensés fonctionner sur un circuit intégré. En prévision de l'implantation du codeur développé sur un DSP[7] du marché, il est important de veiller à restreindre la complexité de l'algorithme développé. De plus, la réduction de la complexité de calcul permet de réduire le délai de traitement du codeur. Dans le cadre de ce travail de recherche, l'objectif a été de réduire au maximum la complexité, dans la mesure où une telle réduction respecte les contraintes décrites précédemment.

5.4 Choix du codeur à développer

Cette section présente les motivations qui ont mené, d'une part au choix du type d'algorithme à développer et implanter, et d'autre part au choix de la structure à donner au codeur. Ces choix sont basés sur les contraintes décrites à la Section 5.3 et sur l'étude de l'état de l'art présentée à la Section 4.3 et à l'Annexe C.1.

Le projet de recherche présenté ici a débuté au cours de l'année 1999. A ce moment là, l'état de l'art pour le codage de la parole en bande élargie faisait clairement apparaître une tendance pour l'implantation de codeurs axés sur des algorithmes de type CELP en une seule bande de fréquences et de type SB-CELP (cf. Sous-section 4.3.1).

5.4.1 Rapport débit - qualité

Selon l'état de l'art présenté à la Section 4.3, les codeurs de type CELP et SB-CELP semblent les plus adaptés pour réaliser un codage du signal de parole en bande élargie, en termes de rapport débit – qualité. Les Tableaux 5–1 et 5–2 situent la "qualité" des codeurs de type CELP et respectivement SB-CELP de l'état de l'art jusqu'en 1999, fonctionnant à un débit inférieur ou égal à 16 kbits/s. Le numéro des codeurs (N°) est indiqué dans l'ordre chronologique de leur présentation à l'Annexe C.1.

En considérant le rapport débit – qualité, à ce stade de la réflexion notre choix s'est porté sur un codeur de type CELP en une seule bande de fréquences. En effet, la littérature montre qu'à débit égal, un tel codeur donne de meilleurs résultats en termes qualitatifs, qu'un codeur de type SB-CELP. De plus, un tel codeur permet d'obtenir non seulement une qualité de signal

[7] http://www.entegra.co.uk/3g.htm (27 sept. 2002).

reconstruit similaire à celle du G.722 C pour un débit inférieur à 14.25 kbits/s, mais également une qualité de signal reconstruit similaire à celle du G.722 B (voir G.722 A) à un débit de 16 kbits/s.

N° du codeur et premier auteur	Réfé-rence	Débit [kbits/s]	Qualité de reconstruction auditive
2. Salami	[C-2]	14	G.722 B
3. Harborg	[C-3]	16	G.722 A
4. McElroy	[C-5]	16	G.722 B-C
5. Black	[C-6]	14.1	G.722 C
6. Sasaki	[C-7]	16	G.722 B
7. Serizawa	[C-8]	16	G.722 B
8. Koishida	[C-9]	16	G.722 A

Tableau 5-1 : Etat de l'art jusqu'en 1999, tel que présentée à l'Annexe C.1 : qualité auditive du signal reconstruit des codeurs de type CELP en une seule bande de fréquences, fonctionnant à un débit inférieur ou égal à 16 kbits/s.

N° du codeur et premier auteur	Réfé-rence	Débit [kbits/s]	Qualité de reconstruction auditive
11. Paulus	[C-12]	16	G.722 C
12. Ubale	[C-13]	16	G.722 C
13. Schnitzler	[C-14]	13	G.722 C
14. Combescure	[C-15]	16	G.722 C
15. Black	[C-16]	16	G.722 C

Tableau 5-2 : Etat de l'art jusqu'en 1999, tel que présentée à l'Annexe C.1 : qualité auditive du signal reconstruit des codeurs de type SB-CELP, fonctionnant à un débit inférieur ou égal à 16 kbits/s.

5.4.2 Complexité de calcul et délai de traitement

Un algorithme de type CELP, repris de la bande étroite et adapté à la bande élargie, souffre d'une complexité très élevée. Un algorithme de type SB-CELP permet de réduire cette complexité de calcul.

En considérant la complexité de calcul des algorithmes de type CELP et de type SB-CELP, notre choix aurait pu se porter sur un codeur de type SB-CELP. Cependant, les codeurs de type SB-CELP présentent divers inconvénients. En effet, le délai de traitement d'un codeur SB-CELP est plus élevé que celui d'un codeur de type CELP. De plus, pour un même débit, la qualité du signal reconstruit que permet d'obtenir un codeur de type SB-CELP est inférieure à celle offerte par un codeur CELP.

5.4.3 Choix du type de codeur

Notre choix s'est finalement porté sur un codeur CELP en une seule bande de fréquences et en particulier sur un codeur ACELP, où le dictionnaire innovateur est de type algébrique. En effet, les sous-sections précédentes montrent que nos objectifs de qualité, de débit et de délai de traitement (cf. Section 5.3) devraient être atteints avec un tel codeur. Le principal obstacle à une telle implantation semble être la complexité de calcul.

En 1999, le codage de la parole en bande étroite et à débit moyen était dominé par les codeurs ACELP. Le nouveau standard AMR pour la parole en bande étroite, le NB-AMR, venait de sortir. Ce standard est basé sur le GSM-EFR, qui définit un codeur ACELP proche du G.729.

Un dictionnaire algébrique présente l'avantage de ne pas nécessiter de mémoire. De plus, la complexité de calcul d'un codeur CELP contenant un dictionnaire algébrique est considérablement réduite par rapport à celle d'un codeur CELP contenant un dictionnaire innovateur composé de vecteurs d'excitation stochastique fixes, où chaque élément est généré par un nombre gaussien aléatoire (cf. Sous-section 2.3.5). Toutefois, pour des codeurs en bande élargie, la recherche dans un dictionnaire algébrique reste complexe. Cependant des approches sous-optimales, telles qu'une recherche focalisée ou un ré-ordonnement d'impulsions, permettent de réduire cette complexité [5-5]. Ainsi, le choix d'un codeur de type ACELP allège l'obstacle lié à la complexité.

Finalement, le choix d'un codeur ACELP a été consolidé par le fait que les plates-formes de traitement du signal digital (DSP) évoluent rapidement et permettent d'implanter des algorithmes toujours plus complexes. De plus, l'utilisation d'un dictionnaire algébrique permet de varier le débit d'encodage très facilement.

5.4.4 Choix de la structure de base

L'implantation du codeur P-MRWB-ACELP est initialement basée sur la structure du codeur G.729. Le G.729 permet d'obtenir un signal de parole reconstruit de qualité suffisante pour la téléphonie en bande étroite. Cette qualité est communément définie par le terme anglais "toll". Nous avons supposé qu'en adaptant ce codeur à la bande élargie, nous devions obtenir une bonne qualité de signal reconstruit.

Nous étions conscients des difficultés qui apparaissent lorsque le modèle CELP est appliqué à des signaux de parole en bande élargie. Ces difficultés sont liées à la complexité et à la qualité. Il est clairement dit en [5-6] qu'un tel

codeur, fonctionnant à un débit limité, souffre d'un signal reconstruit bruité. Plusieurs particularités doivent lui être ajoutées pour qu'il puisse produire une haute qualité de signal reconstruit. Ces particularités sont principalement liées à la pondération perceptuelle, mais également à la prédiction du délai tonal. Si un filtre de pondération perceptuelle inadéquat est utilisé pour modéliser les hautes fréquences, un codeur de type CELP pour la bande élargie produit un signal reconstruit bruité. En effet, le spectre de fréquences de la parole en bande élargie a une pente spectrale importante : la chute en énergie entre les basses et les hautes fréquences est d'environ 35 dB.

5.4.5 Choix de la durée des trames et des sous-trames

Initialement, la durée d'une trame et d'une sous-trame de signal ont été fixées à 20 et respectivement 5 ms. Toutefois, le programme C a été développé de sorte qu'une adaptation de la durée de ces trames et sous-trames puisse se faire à tout moment. En effet, ces valeurs ont été fixées comme paramètres pouvant être changés.

5.4.5.1 Durée d'une trame de signal

L'enveloppe spectrale du signal de parole change relativement lentement dans le temps (cf. Section 2.3) et peut être considérée comme stationnaire sur des intervalles de temps compris entre 10 et 30 ms. Le Tableau 4-1, montre une tendance à utiliser des trames de 10 ou 20 ms pour l'extraction des coefficients LPC. Le Tableau 5-3 illustre la quantification des paramètres LPC, pour les codeurs réalisant une prédiction linéaire ordinaire ("forward") sur une seule bande de fréquences. Avec de tels codeurs, la durée moyenne d'une trame est de 20 ms. Notre choix s'est donc porté sur une trame de 20 ms. Naturellement, plus la trame de laquelle les LPC sont extraits et quantifiés est courte, moins l'excitation est sollicitée pour compenser les inexactitudes de l'enveloppe spectrale. De même, plus cette trame est courte et plus le délai de traitement du codeur est court.

5.4.5.2 Durées des sous-trames

En consultant le Tableau 4-1, on remarque une forte tendance à utiliser des sous-trames de 2.5 ms pour la mise à jour de l'excitation innovatrice, et de 5 ms pour l'excitation adaptative.

N° du codeur et premier auteur	Référence	Ordre de prédiction LPC	Trame [ms]	Nb. bits pour QLPC	Type de quantification pour coefficient LPC
2. Salami	[C-2]	16	30	54	
3. Harborg	[C-3]	16-20	20	60-80	Scalaire (selon réf. [C-4])
4. McElroy	[C-5]	20	25	70	
5. Black	[C-6]	16	20	60	NU, scalaire
9. Roy	[C-10]	16	20	48	NU, différentielle, scalaire
12. Ubale	[C-13]	16	10	28	MA-2, MSVQ
17. Lefebvre	[C-18]	16	24	48	SVQ
18. Xie	[C-19]	12	24	48	SVQ
20. Chen	[C-21]	16	20	49	SVQ

Tableau 5-3 : Quantification de l'enveloppe spectrale pour les codeurs de l'état de l'art jusqu'en 1999, tels que présentés à l'Annexe C.1, réalisant une prédiction linéaire LPC "forward" sur une seule bande de fréquences, avec : ordre de prédiction LPC, durée de la trame en ms, nombre (Nb.) de bits nécessaires à la quantification des LPC (QLPC) et type de quantification utilisée.

Le G.729, traite des sous-trames de 5 ms pour les deux types d'excitations. Il encode l'excitation innovatrice à l'aide d'un dictionnaire algébrique ISPP (cf. Sous-section 3.5.1) et utilise 4 impulsions par vecteur de code. Un vecteur de code a la durée d'une sous-trame et est divisé en 4 pistes : trois pistes de 8 positions et une piste de 16 positions, pour un total de 40 positions. Une impulsion est attribuée à chaque piste. Les 4 impulsions sont encodées avec 17 bits : 13 bits de positions et 4 bits de signe.

A 16 kHz, en utilisant trois pistes de 16 positions et une piste de 32, 21 bits sont nécessaires pour encoder 4 impulsions. Comme la richesse du signal reconstruit dépend principalement de l'excitation innovatrice, pour le codage de la parole en bande élargie le nombre d'impulsions par sous-trame de 5 ms doit être au minimum de 8. Ainsi, pour une sous-trame de 5 ms divisée en 4 pistes, comme expliqué ci-dessus, 4 bits de signe (cf. Sous-Section 3.5.1) et 34 bits de positions sont nécessaires, pour un total de 38 bits par sous-trame.

En traitant l'excitation innovatrice par sous-trames de 2.5 ms et en reprenant la structure du G.729, nous aurions besoin de 34 bits par sous-trame de 5 ms. Le débit passerait de 7.6 kbits/s à 6.8 kbits/s pour l'encodage de cette excitation. La complexité serait réduite mais également la qualité. En effet, nous aurions un dictionnaire de 2^{34} vecteurs et non plus de 2^{42} vecteurs. De plus, bien que le Tableau 4-1 illustre une tendance à utiliser des sous-trames de 2.5 ms pour la mise à jour de l'excitation innovatrice, sur les 8 codeurs de type CELP en une seule bande de fréquences, seulement trois sont des ACELP, dont deux extraient l'excitation algébrique toutes les 5 ms.

Notre choix s'est ainsi porté sur une sous-trame de 5 ms, aussi bien pour l'excitation innovatrice que pour l'excitation adaptative.

5.5 Mesures des performances et bases de données

Cette section décrit les mesures de performances utilisées pour qualifier le codeur, ainsi que les bases de données créées pour entraîner et tester le codeur développé.

5.5.1 Mesures des performances

La mesure des performances d'un codeur de parole est une mesure critique. La Section 2.6 et la Sous-section 3.3.3 introduisent les mesures objective et subjective couramment utilisées pour le codage de la parole.

La mesure objective standard de la qualité d'un codeur est la moyenne du rapport signal sur bruit calculé sur des segments de parole (SNR par segments). Comme les codeurs CELP ne sont pas des codeurs d'ondes, cette mesure ne donne pas toujours une information correcte de la qualité du codage. En effet, elle ne tient pas compte des propriétés perceptuelles de l'oreille utilisées par les codeurs CELP. Ainsi, le SNR par segments ne donne qu'une mesure informative. Pour qualifier les codeurs CELP, une mesure perceptuelle est plus adéquate.

La mesure perceptuelle standard utilisée ici, est la mesure "double aveugle avec référence cachée" selon l'échelle des DMOS (cf. Section 2.6). Pour être significative, elle doit être réalisée par un certain nombre d'auditeurs et avec une base de données de test variée. Pour simplifier la procédure, en cours de développement les tests auditifs n'ont été réalisés que par une à deux personnes, et les signaux de sortie obtenus à l'aide de deux implantations différentes du codeur ont simplement été comparés.

Les mesures présentées ci-dessus s'appliquent aux performances du signal reconstruit. Les performances de la quantification des LPC sont usuellement obtenues en se basant sur la racine carrée de la distorsion spectrale (cf. Sous-section 3.3.3).

5.5.2 Bases de données : entraînements et tests

Les bases de données TIMIT [5-7], BD-SONS [5-8] et ITU [5-9] ont été utilisées au cours de ce travail de recherche. Ce sont respectivement des bases de données en langue anglaise, française et multi-linguistique. BD-SONS est composée de 7 CD-ROM, numérotés de 1 à 7. Une base de données en français et en italien, appelée BD-IMT, a été créée en enregistrant les voix de quatre enfants et celle d'une collègue ayant une fréquence fondamentale

moyenne très élevée. Les enfants sont deux fillettes de 5 et 8 ans et deux garçons de 3 et 11 ans. BD-IMT a été enregistrée à 44.1 kHz, en utilisant un DAT (Digital Audio Tape), puis a été convertie à 16 kHz par un algorithme d'interpolation.

Une base de données pour l'entraînement du codeur (BD-TRAIN) et une base de données indépendante pour le tester (BD-TEST) ont été créées. BD-TRAIN contient les 1183 premiers fichiers de TIMIT (60'), les 42 fichiers de ITU (59') et 77 fichiers de BD-SONS n°6 (60'). BD-TEST contient les 30 fichiers du répertoire "LABISE" de BD-SONS n°3. BD-IMT a également été utilisée pour différents tests auditifs.

Tous les fichiers utilisés sont préalablement filtrés à l'aide du filtre P341, défini par l'ITU en [5-10]. Ce filtre caractérise la bande passante de la parole en bande élargie pour la téléphonie. Les parties des signaux de BD-TRAIN correspondant à du silence ont été retirées de la base de données d'entraînement.

5.6 Conception d'un dictionnaire de quantification pour les coefficients de prédiction linéaire

La quantification des paramètres LPC pour les codeurs de parole en bande étroite a largement été étudiée et il existe une quantité considérable de littérature scientifique sur le sujet. Tel n'est pas le cas pour le codage de la parole en bande élargie.

Cette section décrit le développement et l'implantation d'un quantificateur pour l'encodage des paramètres LPC du codeur P-MRWB-ACELP. Pour quantifier les coefficients LPC de façon efficace et stable, ils sont transformés en paires de lignes spectrales (LSP). Les LSP sont une représentation alternative des LPC, couramment utilisés pour le codage de la parole. La Sous-section 3.3.6 définit les LSP et en décrit les propriétés.

La Sous-section 5.6.1 rappelle quelques points importants relatifs aux méthodes de quantification adaptées aux LSP, et à leur mesure de performances. La Sous-section 5.6.2 décrit sommairement les quantificateurs spectraux proposés dans la littérature pour la parole en bande élargie. La Sous-section 5.6.3 présente divers schémas de quantification vectorielle (VQ) expérimentés pour encoder les LSP d'ordre 16 du codeur P-MRWB-ACELP. La Sous-section 5.6.4 décrit le schéma de quantification sélectionné. Finalement la Sous-section 5.6.5 discute les tests et les résultats obtenus.

5.6.1 Méthodes de quantification des LSP et méthodes de tests

5.6.1.1 Méthodes de quantification adaptées aux LSP

La Sous-section 3.3.7 décrit différentes méthodes de quantification adaptées aux LSP ainsi que leurs propriétés. Quelques points importants sont rappelés ici:

- La quantification vectorielle exploite la corrélation intra-trame des paramètres LSP. Par rapport à la quantification scalaire, à débit égal, elle permet de réduire la distorsion spectrale. Un quantificateur vectoriel unique pour 16 LSP est irréalisable. En effet, son implantation est prohibitive du point de vue de la taille de la mémoire nécessaire à son stockage, et de la complexité algorithmique relative à l'extraction du vecteur de LSP quantifié. Diverses méthodes de quantification vectorielle sous-optimales, telles que la quantification vectorielle séparée en sous-vecteurs (SVQ) ou la quantification vectorielle sur plusieurs étages de quantification (MSVQ) sont utilisables. Toutefois, à débit égal ces méthodes abaissent les performances du quantificateur. La quantification SVQ traite le vecteur de LSP en sous-vecteurs. A chacun des sous-vecteurs est associé un "sous-dictionnaire" de quantification dans lequel une recherche complète est réalisée. La quantification MSVQ consiste en une séquence de quantificateurs vectoriels, où chaque quantificateur traite l'erreur de quantification de l'étage précédent. Dans la pratique, la quantification SVQ est préférée à la MSVQ. En effet, les LSP ont une sensibilité spectrale localisée. En utilisant la quantification SVQ, qui sépare le vecteur de LSP en régions fréquentielles, les distorsions dues à la quantification sont limitées à la région fréquentielle de chaque sous-vecteur. De plus, la quantification SVQ est robuste aux erreurs de transmission.
- La quantification vectorielle prédictive exploite les propriétés de corrélation temporelle (inter-trames) des LSP. La prédiction à moyenne glissante (MA) est souvent utilisée car elle est peu sensible aux erreurs de transmission du canal.
- Une mesure de distance pondérée, utilisée pour extraire les vecteurs de LSP quantifiés, assigne plus de poids aux LSP les plus sensibles d'un point de vue perceptif. Elle exploite la relation entre la distance séparant les LSP consécutifs et les formants de l'enveloppe spectrale. Elle peut permettre de réduire légèrement le débit.

5.6.1.2 Méthodes de tests

La Sous-section 3.3.3 décrit la mesure objective des performances d'un quantificateur de LPC. Cette mesure est basée sur la moyenne de la distorsion spectrale, SD, calculée pour un grand nombre de trames de signal. Pour le codage de la parole en bande étroite, il existe un critère de transparence couramment utilisé et accepté. Ce n'est pas le cas pour le codage de la parole en bande élargie. Nous avons décidé d'utiliser les mesures sur lesquelles est basé le critère précité pour la bande étroite, avec les bornes fréquentielles de mesures fixées à 0 et 7000 Hz, et de coupler ces mesures à des tests auditifs. Cette démarche a permis de déterminer si le critère de transparence de la bande étroite, avec les nouvelles bornes d'intégration, est également valable pour la bande élargie (critère A); si celui fixé par Ferhaoui et Gerven (critère B) est plus adéquat (cf. Sous-section 3.3.3); ou s'il faut utiliser un autre critère.

Les tests auditifs ont été réalisés à l'aide du système illustré à la Figure 5.1. Une analyse spectrale est effectuée sur le signal original $s(n)$. Les LPC d'ordre 16 $(a_{16}(1),\ldots,a_{16}(16))$ sont calculés et transformés en LSP. Les LSP sont quantifiés et transformés en LPC quantifiés $(\hat{a}_{16}(1),\ldots,\hat{a}_{16}(16))$. Le signal original est passé dans le filtre d'analyse $A(z)$, cascadé au filtre de synthèse $1/\hat{A}(z)$, où :

$$A(z) = 1 + \sum_{i=1}^{16} a_{16}(i) \cdot z^{-i};$$

$$\frac{1}{\hat{A}(z)} = \frac{1}{1 + \sum_{i=1}^{16} \hat{a}_{16}(i) \cdot z^{-i}}.$$

La sortie du filtre de synthèse est appelée signal quantifié $\hat{s}(n)$. Ce signal est comparé au signal original $s(n)$ par des tests auditifs.

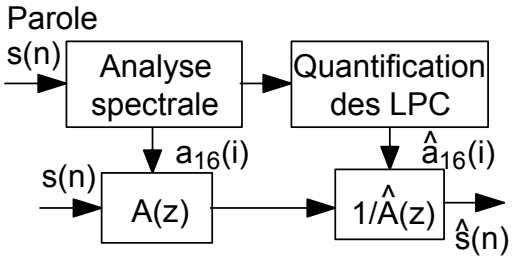

Figure 5.1 : Système utilisé pour réaliser les tests auditifs subjectifs.

5.6.2 Exemples de quantificateurs spectraux

Divers quantificateurs spectraux proposés dans la littérature, pour le codage de la parole en bande élargie, sont reportés ci-dessous :

- L'article de Guibé et *al.* [5-11] décrit les techniques de quantification spectrale des codeurs en bande élargie N° 3, 11, 12, 14, 17 et 20, présentés à l'Annexe C.1. Le Tableau 5-4 montre les caractéristiques de ces techniques. Cet article présente également une étude comparative de différents types de quantificateurs vectoriels : les quantificateurs SVQ, MSVQ, S-MSVQ, prédictifs, non-prédictifs et passant en fonction des besoins d'une quantification prédictive à une quantification non-prédictive ou vice-versa (Safety-net : SNVQ). Un quantificateur S-MSVQ combine les techniques SVQ et MSVQ. Le Tableau 5-5 contient les résultats présentés. La notation utilisée pour décrire les quantificateurs S-MSVQ est illustrée ici par un exemple où sur le premier étage les 16 LSP sont séparés en 3 sous-vecteurs de dimension 6, 7 et 3. Chacun de ces sous-vecteurs est quantifié en utilisant 7 bits. Sur le deuxième étage le vecteur résiduel est séparé en 4 sous-vecteurs de dimensions 4, 4, 4 et 4. Chacun de ces sous-vecteurs est quantifié en utilisant 5 bits. Un total de 41 bits est nécessaire. La notation 41-$[(6,7,3)_{7,7,7};(4,4,4,4)_{5,5,5,5}]$ décrit ce schéma.
  Certaines rubriques du Tableau 5-5 sont vides, les informations correspondantes n'étant pas décrites dans l'article référencé. Guibé et *al.* utilisent la quantification prédictive de type AR et d'ordre 1 (AR-1). Pour les quantificateurs SNVQ, un bit (+1) sert à notifier quelle est la quantification utilisée. Le schéma SNVQ combine un quantificateur non prédictif à 41 bits et un quantificateur prédictif à 36 bits. Il semble donner de bons résultats en termes de distorsion spectrale, mais implique une grande taille de mémoire et une complexité de calcul élevée. Les auteurs ne spécifient pas les bornes fréquentielles sur lesquelles la mesure la distorsion spectrale (BFSD) est réalisée.
- L'article de Gibbs et Hoskin [5-12] considère un quantificateur vectoriel, adapté à la parole en bande élargie, en présence de parole traitée par un codeur en bande étroite. Ce quantificateur encode les LSP d'ordre 18, extraits toutes les 20 ms et séparés en 6 sous-vecteurs de 3 LSP (SVQ). La quantification se fait en utilisant une mesure de distance pondérée (WQ : weighted quantization). Un total de 40 bits est alloué de façon optimale aux 6 sous-dictionnaires. Une prédiction MA d'ordre 0, 1 (MA-1) et 2 (MA-2) est testée. Le Tableau 5-5 décrit les meilleurs schémas de quantification considérés, pour des prédictions d'ordre 0, 1 et 2.
- L'article de Ferhaoui et Gerven [5-13] décrit un quantificateur MSVQ et propose un nouveau critère de transparence pour le codage en bande

élargie (critère B). Ce critère est décrit à la Sous-section 3.3.3. Un quantificateur MSVQ est testé pour des LSP d'ordre 18, extraits toutes les 15 ms en utilisant une fenêtre de 20 ms. Le Tableau 5-5 décrit diverses configurations testées.

- L'article de Ragot et *al.* [5-14] décrit une méthode pour quantifier les LSP à l'aide d'une classe spécifique des codes en treillis pratiquement ellipsoïdaux. Ces codes sont appelés "codes de Voronoi généralisés". Avec 44 bits, les auteurs atteignent la transparence selon le critère A. Ils quantifient des LSP d'ordre 16, extraits toutes les 20 ms, en utilisant une fenêtre de 40 ms. Le Tableau 5-5 décrit l'un des schémas de quantification proposés. Il utilise les treillis appelés RE_{16}.
- Le document [5-15] décrit le codeur WB-AMR de l'ETSI. Le signal de parole est décimé à 12.8 kHz, et une analyse LPC d'ordre 16 (pour la bande de fréquences [0, 6400] Hz) est effectuée toutes les 20 ms en utilisant une fenêtre de 30 ms (5 ms de la trame précédente, la trame courante, 5 ms de "look-ahead"). Les LPC sont convertis en ISP puis quantifiés. Le codeur fonctionne selon 9 modes correspondant à des débits variant de 6.6 à 23.85 kbits/s. La quantification des ISP est réalisée en utilisant les schémas S-MSVQ à deux étages suivants : 36-$[(9,7)_{8,8;}(5,4,7)_{7,7,6}]$, MA-1 pour le mode à 6.6 kbits/s, et pour tous les autres modes 46-$[(9,7)_{8,8;}(3,3,3,3,4)_{6,7,7,5,5}]$, MA-1. La mesure de distance utilisée est Euclidienne non pondérée.

Pour réaliser une comparaison des quantificateurs spectraux présentés ci-dessus, ceux-ci devraient tous traiter des LSP de même ordre, extraits à l'aide d'une fenêtre de même durée et avec le même taux de remise à jour. De plus, les tests devraient être réalisés avec la même base de données, et la même mesure de distorsion spectrale.

Nous avons expérimenté les schémas de quantification SVQ et S-MSVQ. Nous avons testé les quantificateurs prédictifs et non-prédictifs. Finalement, nous avons considéré des quantificateurs utilisant une mesure de distance Euclidienne pondérée et non-pondérée pour la sélection des vecteurs de code dans les différents dictionnaires. L'expérimentation des différents schémas de quantification spectrale est présentée à la Sous-section 5.6.3.

5.6.3 Expérimentation de différents schémas de quantification spectrale

5.6.3.1 Extraction des LPC

Les trames de signal du codeur P-MRWB-ACELP ont une durée de 20 ms (cf. Sous-section 5.4.5). Les coefficients de prédiction linéaire LPC sont extraits une fois par trame de signal, en minimisant l'énergie ε_p de l'erreur de prédiction selon la méthode des moindres carrés. ε_p est donnée par l'équation (2.10) :

$$\varepsilon_p = \sum_{n=-\infty}^{\infty} x^2(n) = \sum_{n=-\infty}^{\infty} \left[s(n) + \sum_{k=1}^{p} a_p(k) \cdot s(n-k) \right]^2 .$$

Cette somme devient limitée en fenêtrant le signal original sur 30 ms (20 ms de la trame traitée et 10 ms de la trame précédente). La fenêtre $w(n)$ utilisée est asymétrique et discontinue. Son poids est concentré sur la quatrième sous-trame de la trame traitée. Elle est composée de deux parties : la première est une demi-fenêtre de Hamming; la seconde est un quart de cycle de la fonction cosinus. Elle est illustrée à la Figure 5.2 et est donnée par l'équation suivante :

$$w(n) = \begin{cases} 0, & n < 0 \text{ et } n \geq W, \\ 0.54 - 0.46 \cdot \cos\left(\dfrac{2\pi n}{2K-1} \right), & n = 0, \ldots, K-1, \\ \cos\left(\dfrac{2\pi (n-K)}{4(W-K)-1} \right), & n = K, \ldots, W-1, \end{cases} \qquad (5.1)$$

où $K = 440$ et $W = 480$. Les coefficients LPC d'ordre 16 sont calculés par la méthode d'auto-corrélation et en utilisant la récursion de Levinson-Durbin introduite à la Sous-section 2.3.3

N° du codeur et premier auteur	Réfé-rence	Ordre de prédiction LPC	Trame [ms]	Nb. de bits pour les QLPC	Type de quantification pour les coefficients LPC
3. Harborg	[C-3]]	16 à 20	20	60 à 80	Scalaire (selon référence [4-11])
11. Paulus	[C-12]	14	10	44	MA, SVQ
12. Ubale	[C-13]	16	10	28	MA-2, MSVQ
14. Combescure	[C-15]	12/8 (BS)	20	33 / 10 (BS)	MA, MSVQ, SVQ / VQ (BS)
17. Lefebvre	[C-18]	16	24	48	SVQ
20. Chen	[C-21]	16	20	49	SVQ

Tableau 5-4 : Techniques de quantification spectrale de différents codeurs en bande élargie présentés par Guibé (les références citées sont fournies dans l'Annexe C).

Premier auteur	Réfé-rence	Ordre de prédiction LPC	Type de quantification	Nb. de bits	Configuration	BFSD	SD moyenne [dB]	Trames avec 2<SD≤4 dB en [%]	Trames avec SD>4 dB en [%]
Guibé	[5-11]	16	S-MSVQ	41	41-$[(6,7,3)_{7,7,7;}(4,4,4,4)_{5,5,5,5}]$		1.46		
			AR-1, SVQ	36	36-$[(4,4,4,4)_{9,9,9,9}]$, AR-1		1.45		
			AR-1, SVQ	40	40-$[(4,4,4,4)_{10,10,10,10}]$, AR-1		1.09	4.24	0
			SNVQ	41/36 +1			1.09	0.38	0
Gibbs	[5-12]	18	SVQ, WQ	40	40-$[(3,3,3,3,3,3)_{6,8,8,7,6,5}]$	0-7000	1.570		
			MA-1, SVQ, WQ	40	40-$[(3,3,3,3,3,3)_{7,8,7,7,6,5}]$, MA-1	0-7000	1.324		
			MA-2,SVQ, WQ	40	40-$[(3,3,3,3,3,3)_{7,8,8,7,6,4}]$, MA-2	0-7000	1.264		
Ferhaoui	[5-13]	16	MSVQ	36	36-$[(16)_6;(16)_6;(16)_6;(16)_6;(16)_6;(16)_6]$	0-7000	1.62	21.74	3.13
			MSVQ	56	56-$[(16)_8;(16)_8;(16)_8;(16)_8;(16)_8;(16)_8;(16)_8]$	0-7000	1.05	7.42	0.48
Ragot	[5-14]	16	Treillis RE_{16}	44		50-7000	0.87	1.41	0.10

Tableau 5-5 : Résultats obtenus par Guibé et *al.*, Gibbs et Hoskin, Ferhaoui et Gerven, et Ragot et *al.*, pour divers schémas de quantification.

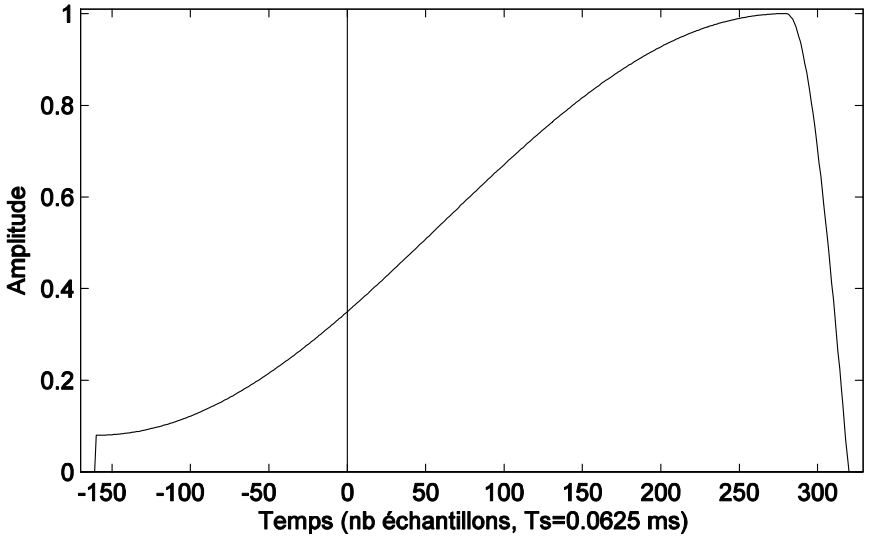

Figure 5.2 : Fenêtre d'analyse pour l'extraction des coefficients LPC

5.6.3.2 Quantification SVQ et allocation optimale des bits

Pour toute l'expérimentation présentée ici, les dictionnaires ont été entraînés en utilisant l'algorithme de Linde-Buzo-Gray LBG [5-16]. Nous avons commencé par tester la quantification SVQ, sans prédiction et avec une mesure de distance Euclidienne non-pondérée (DE).

Le nombre minimal de bits permettant d'atteindre la transparence subjective a été testé. Le premier problème a consisté en la recherche de la séparation optimale d'un vecteur composé de 16 LSP. Dans un premier temps, la complexité algorithmique et la taille de la mémoire de stockage n'ont pas été considérées. Naturellement, ces paramètres diminuent avec le nombre de séparations effectuées mais il en est de même pour les performances du quantificateur. En [5-11], Guibé et *al.* affirment que les 3 derniers LSP d'un vecteur de 16, correspondant aux plus hautes fréquences, ont un comportement statistique différent des 13 premiers LSP. Ainsi, une première séparation en 13 et 3 LSP est envisageable. En [5-12] Gibbs et Hoskin constatent qu'un choix commun pour le codage de la parole en bande étroite est d'utiliser des sous-vecteurs de 3 LSP. Nous avons donc effectué nos premières expériences avec un vecteur de 16 LSP séparé en 5 sous-vecteurs de dimensions 4-3-3-3-3.

Nous avons déterminé l'allocation optimale des bits en utilisant une méthode itérative. Initialement, 3 bits sont attribués à chacun des 5 sous-dictionnaires, pour un budget total de 15 bits. Puis, à chacune des itérations le budget a été incrémenté de 1 bit. Ce bit supplémentaire est alloué provisoirement à chacun des sous-dictionnaires. Les sous-dictionnaires sont ainsi entraînés puis testés 5 fois par itération. La distorsion spectrale est

mesurée et la configuration qui donne la meilleure amélioration en termes de distorsion spectrale est choisie. Cette procédure itérative a été réalisée pour un nombre total de bits compris entre 16 et 45. Le Tableau 5-6 regroupe les résultats obtenus. Pour un budget de bits donné, il décrit le meilleur schéma de quantification obtenu, ainsi que les caractéristiques de la distorsion spectrale permettant de définir la transparence objective (critère A). Ce tableau montre par exemple, qu'à la première et à la seconde itération, la meilleure amélioration marginale, en termes de distorsion spectrale moyenne, a été obtenue en attribuant le bit supplémentaire au premier sous-dictionnaire.

Nombre total de bits	**Nombre de bits alloués aux sous-dictionnaires 1 à 5**	**Distorsion spectrale (SD) moyenne [dB]**	**Trames dont la SD vaut 2<SD≤4 dB [%]**	**Trames dont la SD vaut SD >4 dB [%]**
16	4,3,3,3,3	3.2116	85.89	11.85
17	5,3,3,3,3	3.0491	86.45	9.28
18	5,4,3,3,3	2.9166	87.76	6.16
19	5,4,4,3,3	2.8064	89.16	3.65
20	6,4,4,3,3	2.7100	87.14	2.84
21	7,4,4,3,3	2.5901	82.42	2.12
22	7,4,5,3,3	2.4929	78.21	1.51
23	7,5,5,3,3	2.3907	74.56	0.94
24	7,6,5,3,3	2.2827	67.37	0.58
25	7,6,5,4,3	2.1904	62.10	0.18
26	7,6,5,5,3	2.1129	56.19	0.18
27	7,6,6,5,3	2.0360	49.8	0.09
28	7,7,6,5,3	1.9543	42.39	0.03
29	8,7,6,5,3	1.8741	34.69	0.05
30	8,7,7,5,3	1.8104	29.58	0.04
31	8,7,7,5,4	1.7439	23.38	0.01
32	8,7,7,6,4	1.6807	18.53	0.02
33	9,7,7,6,4	1.6046	14.34	0.01
34	9,8,7,6,4	1.5430	10.78	0.00
35	10,8,7,6,4	1.4815	8.40	0.02
36	10,9,7,6,4	1.4268	6.27	0.00
37	10,9,8,6,4	1.3700	4.83	0.00
38	10,9,8,7,4	1.3178	3.73	0.00
39	11,9,8,7,4	1.2638	2.83	0.00
40	11,9,8,8,4	1.2186	2.50	0.00
41	11,9,8,8,5	1.1748	1.32	0.00
42	11,9,9,8,5	1.1251	0.91	0.00
43	11,9,9,8,6	1.0801	0.73	0.00
44	11,10,9,8,6	1.0354	0.50	0.00
45	12,10,9,8,6	0.9889	0.35	0.00

Tableau 5-6 : Allocation optimale des bits à chacun des sous-dictionnaires de dimensions 4-3-3-3-3, en fonction du nombre total de bits; distorsion spectrale moyenne; trames (en %) dont la distorsion spectrale SD est comprise entre 2 et 4 dB, ou supérieure à 4 dB.

Les critères de transparence A et B, décrits au paragraphe 5.6.1.2, sont respectivement atteints pour un budget de 45 et 34 bits. Sur la base de ces résultats, les expérimentations décrites aux points 5.6.3.3 et 5.6.3.4 ont été menées pour un nombre total de bits correspondant à 34 et 45 bits.

Des tests auditifs ont montré que la configuration 45-$[(4,3,3,3,3)_{12,10,9,8,6}]$ est transparente, alors que la configuration 34-$[(4,3,3,3,3)_{9,8,7,6,4}]$ entraîne une distorsion audible significative. La complexité liée à la configuration 45-$[(4,3,3,3,3)_{12,10,9,8,6}]$ est de 21952 multiplications, 21952 soustractions et 16000 additions par trame de signal traitée, soit 21952 MAC (multiplications et accumulations) et ADD (additions ou soustractions). De plus, pour une telle configuration, une mémoire de 21952 valeurs doubles (MEM) est nécessaire.

5.6.3.3 Ordre de prédiction et extraction des coefficients

La quantification utilisée au point 5.6.3.2 est une quantification "sans mémoire" puisque aucune prédiction temporelle n'est réalisée. Bien que la quantification sans mémoire soit robuste contre les erreurs du canal de transmission, la quantification prédictive exploite la corrélation temporelle (inter-trames) des vecteurs de LSP consécutifs, et permet une réduction du débit. La prédiction est soit auto régressive (AR) soit à moyenne glissante (MA). Généralement, pour obtenir des performances égales, une prédiction MA requiert un ordre de prédiction supérieur à celui d'une prédiction AR. L'avantage principal d'un système MA est que la réponse impulsionnelle finie, du filtre de prédiction correspondant, limite la propagation des erreurs de transmission en cas de canal bruité. Nous avons donc décidé d'utiliser une prédiction MA.

Pour un système MA, la i^{ème} valeur linéairement prédite, $\hat{f}_i(n)$ est donnée par [5-17] :

$$\hat{f}_i(n) = \sum_{k=1}^{q} b_i(k) \cdot \tilde{e}_i(n-k), \qquad (5.2)$$

où les $\tilde{e}_i(n-k)$ sont les erreurs de prédiction précédemment encodées, les $b_i(k)$ sont les coefficients de prédiction et q est l'ordre de prédiction.

Les coefficients de prédiction MA sont déterminés en utilisant soit un algorithme LMS (Least-Mean-Square) [5-18], soit une méthode basée sur une approximation AR d'un ordre élevé du processus MA [5-19]. Selon le théorème de décomposition de Wold (1938), n'importe quel processus MA peut être représenté par un modèle AR d'ordre aussi grand que possible [5-20]. Si un processus d'ordre q, dénoté MA(q), peut être modélisé par un modèle AR d'ordre p, dénoté AR(p), où $p \gg q$, alors le filtre $B(z)$ de type

MA et d'ordre q, est exprimé comme une fonction du filtre $A(z)$ de type AR et d'ordre p :

$$B(z) = \frac{1}{A(z)}, \tag{5.3}$$

où

$$A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k}, \quad B(z) = \sum_{k=0}^{q} b_k z^{-k}. \tag{5.4}$$

Les ensembles de paramètres $\{b_k\}$ et $\{a_k\}$ sont liés par la relation :

$$a_n + \sum_{k=1}^{q} b_k a_{n-k} = \begin{cases} 1, & n = 0, \\ 0, & n \neq 0. \end{cases} \tag{5.5}$$

L'ensemble de paramètres $\{a_k\}$ s'obtient en faisant correspondre le signal au modèle AR(p). Pour calculer les paramètres $\{b_k\}$, une meilleure correspondance est obtenue en minimisant l'erreur carrée ε donnée par l'équation (5.6) :

$$\varepsilon = \sum_{n=0}^{p} \left[a_n + \sum_{k=1}^{q} b_k a_{n-k} \right]^2 - 1; \tag{5.6}$$

$$\text{avec} \quad a_0 = 1 \quad \text{et} \quad a_k = 0, \quad \text{si} \quad k < 0.$$

Ceci correspond à résoudre une équation de Levinson-Durbin d'ordre q en utilisant les paramètres $\{a_k\}$ comme entrée.

Nous avons testé la prédiction d'ordre 0 (sans prédiction), 1, 2 et 3. Pour réduire la complexité de l'expérimentation et bien que ce soit légèrement sous-optimal, les prédicteurs MA ont été entraînés en boucle ouverte et maintenus constants pour tous les tests réalisés. De plus, nous avons utilisé la valeur moyenne des coefficients MA extraits d'un même sous-vecteur de LSP, puisque nous avons observé que ces coefficients sont très proches. Pour simplifier la procédure, une fois les coefficients MA déterminés, les dictionnaires de quantification vectorielle sont entraînés en boucle ouverte.

La distorsion spectrale a été mesurée avec les schémas de quantification 45-$[(4,3,3,3,3)_{12,10,9,8,6}]$ et 34-$[(4,3,3,3,3)_{9,8,7,6,4}]$, pour une prédiction d'ordre 0, 1, 2 et 3. La Figure 5.3 illustre les résultats obtenus. Elle montre que par rapport à une prédiction d'ordre nul, une prédiction d'ordre 1 permet de réduire considérablement la distorsion spectrale moyenne. Par contre, si l'ordre de prédiction augmente davantage, la distorsion spectrale moyenne n'est que légèrement réduite.

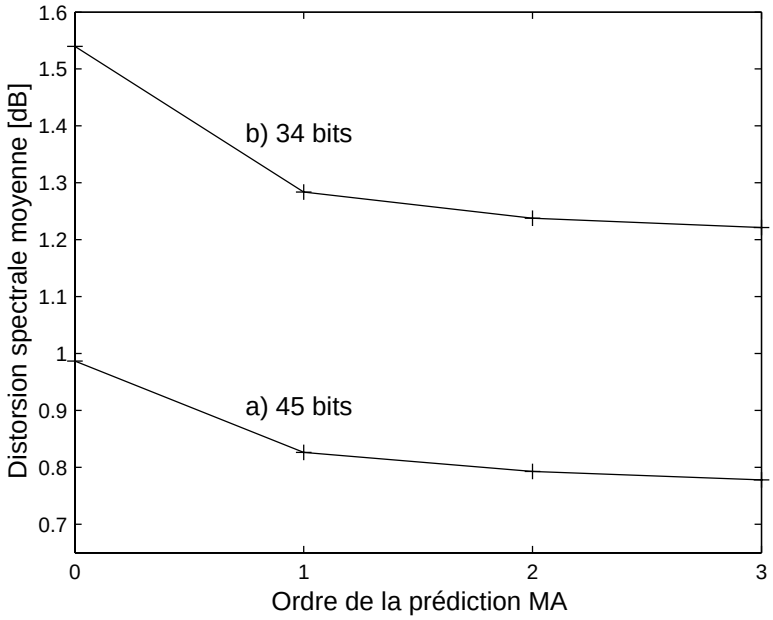

Figure 5.3 : Distorsion spectrale moyenne pour différents ordres de prédiction MA, et les schémas de quantifications à 45 et 34 bits sus-mentionnés.

5.6.3.4 Pondération perceptuelle

La mesure de distance Euclidienne (DE), utilisée pour les expérimentations des points 5.6.3.2 et 5.6.3.3 est non-pondérée. La mesure de distance Euclidienne pondérée (DEP), $d(\mathbf{f},\hat{\mathbf{f}})$, mesurée entre le vecteur $\mathbf{f}$ et sa valeur quantifiée $\hat{\mathbf{f}}$, tient compte des propriétés de perception de l'oreille humaine et assigne un poids w_i différent à chacun des LSP, f_i, en fonction de son importance perceptuelle :

$$d(\mathbf{f},\hat{\mathbf{f}}) = \sum_{i=1}^{m} w_i \left(f_i - \hat{f}_i \right)^2 \quad [\mathrm{Hz}^2]. \tag{5.7}$$

Un poids supérieur est attribué aux LSP correspondant aux fréquences spectrales des pics formantiques du signal. De même, plus de poids est alloué aux LSP des formants de grande amplitude, qu'à ceux des formants de basse amplitude. Les LSP sont regroupés autours des formants. Un groupe de 2 à 3 LSP caractérise une fréquence formantique et la largeur de bande du formant dépend de la proximité des LSP du groupe. Les poids w_i du schéma de pondération du G.729 sont définis par l'équation :

$$1 \le i \le 10; \quad d_i = \omega_{i+1} - \omega_{i-1}; \quad \omega_i = 2 \cdot \pi \cdot f_i;$$
$$\omega_0 = 0.04 \cdot \pi; \quad \omega_{11} = 0.92 \cdot \pi; \quad \alpha = 1.0; \quad \beta = 1.0;$$
$$w_i = \begin{cases} \beta, & \text{si } d_i > \alpha, \\ 10 \cdot \dfrac{(d_i - \alpha)^2}{\alpha^2} + \beta, & \text{ailleurs,} \end{cases} \tag{5.8}$$

où les poids w_5 et w_6 (donnés par l'équation (5.8)) sont multipliés par 1.2.

Nous avons modifié ce schéma pour tenir compte de la densité supérieure des LSP, lors du codage de la parole en bande élargie. Ainsi, nous avons utilisé pour le schéma de pondération suivant :

$$1 \le i \le 16; \quad d_i = \omega_{i+1} - \omega_{i-1}; \quad \omega_i = 2 \cdot \pi \cdot f_i;$$
$$\omega_0 = 0; \quad \omega_{17} = \pi; \quad \alpha = 0.6; \quad \beta = 0.1;$$
$$w_i = \begin{cases} \beta & \text{si } d_i > (1-\beta) \cdot \alpha, \\ \dfrac{1}{(1+\beta)} \left[\dfrac{(d_i - \alpha)^2}{\alpha^2} + \beta \right] & \text{ailleurs.} \end{cases} \tag{5.9}$$

Les paramètres α et β ont été déterminés expérimentalement et correspondent aux meilleurs résultats obtenus. Si $\beta = 0.1$ et $\alpha = 0.9$ notre schéma (5.9) est pratiquement identique à celui du G.729 (5.8).

Nous avons testé la distorsion spectrale moyenne avec les schémas de quantification 45-$[(4,3,3,3,3)_{12,10,9,8,6}]$ et 34-$[(4,3,3,3,3)_{9,8,7,6,4}]$ et une prédiction d'ordre 0, 1, 2 et 3. Nous avons utilisé la mesure DE et l'algorithme LBG pour entraîner les sous-dictionnaires. Le quantificateur a été testé en utilisant la mesure DEP avec les poids donnés par l'équation (5.9). La Figure 5.4 illustre les résultats. Par rapport à la mesure DE, la mesure DEP réduit la distorsion spectrale moyenne d'environ 0.04 et respectivement 0.03 dB. La Figure 5.5 montre l'effet de la mesure DEP sur l'histogramme de la distorsion spectrale du schéma de quantification 34-$[(4,3,3,3,3)_{9,8,7,6,4}]$, MA-1.

L'utilisation de la mesure DEP utilisée lors de l'entraînement des sous-dictionnaires a été testée. Elle n'apporte aucune amélioration.

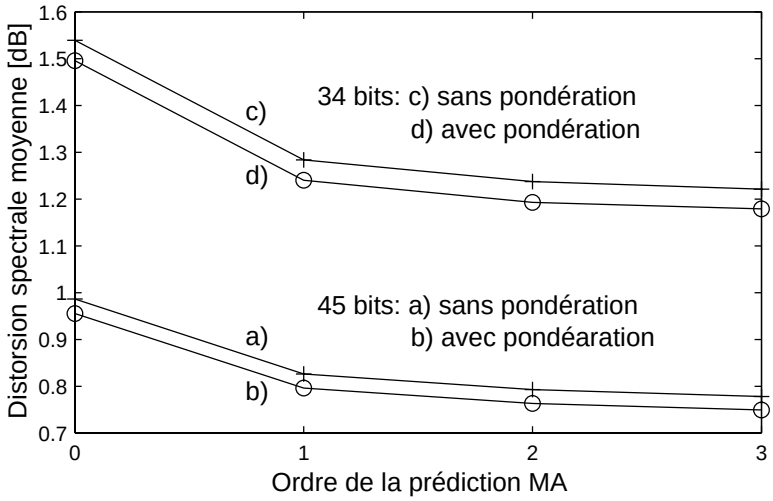

Figure 5.4 : Distorsion spectrale moyenne des schémas à 45 et 34 bits du Tableau 5-6, avec les mesures DE et DEP, et une prédiction MA d'ordre 0, 1, 2 et 3.

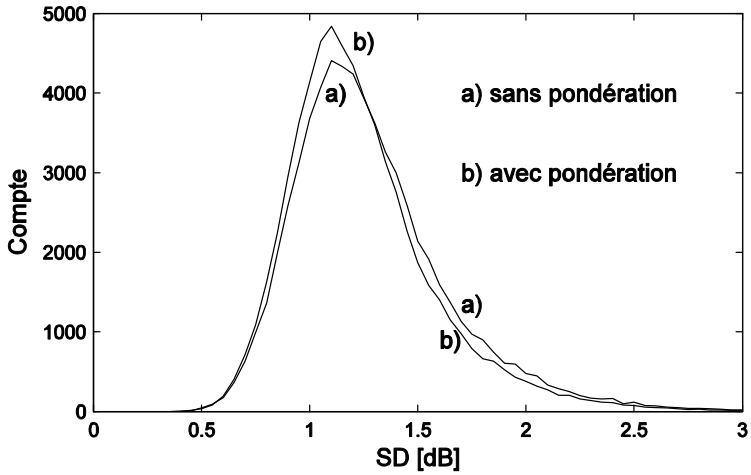

Figure 5.5 : Histogramme de la distorsion spectrale (SD) avec les mesures DE et DEP, pour le schéma à 34 bits du Tableau 5-6 avec prédiction MA-1.

5.6.3.5 Seconde itération d'allocation des bits

Pour déterminer la meilleure allocation de bits en utilisant la prédiction MA-1 et la mesure DEP, nous avons répété la procédure décrite au point 5.6.3.2. Le Tableau 5-7 regroupe les résultats obtenus. Selon le critère A, la transparence est atteinte avec le schéma 40-$[(4,3,3,3,3)_{11,9,8,7,5}]$, MA-1, DEP. Ce schéma permet une réduction de 5 bits par rapport au schéma 45-$[(4,3,3,3,3)_{12,10,9,8,6}]$. La complexité liée à cette configuration est de 21952 multiplications, 10976 soustractions et 8000 additions par trame de signal traitée et de 10976 valeurs doubles en mémoire.

Nombre total de bits	Nombre de bits alloués aux sous-dictionnaires 1 à 5	Distorsion spectrale (SD) moyenne [dB]	Trames dont la SD vaut 2<SD≤4 dB [%]	Trames dont la SD vaut SD >4 dB [%]
16	4,3,3,3,3	2.7904	85.66	4.65
17	4,4,3,3,3	2.6437	82.93	3.01
18	5,4,3,3,3	2.5097	78.01	2.02
19	5,4,4,3,3	2.3749	72.06	1.26
20	6,4,4,3,3	2.2696	64.33	0.90
21	6,5,4,3,3	2.1629	57.25	0.57
22	7,5,4,3,3	2.0739	49.05	0.45
23	7,5,4,4,3	1.9832	41.93	0.24
24	7,5,5,4,3	1.8936	34.78	0.13
25	7,6,5,4,3	1.8083	28.42	0.10
26	8,6,5,4,3	1.7342	23.26	0.05
27	8,6,6,4,3	1.6623	19.35	0.04
28	8,6,6,5,3	1.5905	14.55	0.03
29	8,7,6,5,3	1.5241	11.37	0.02
30	9,7,6,5,3	1.4643	9.43	0.02
31	9,7,6,6,3	1.4084	7.60	0.00
32	9,7,7,6,3	1.3515	6.06	0.00
33	9,7,7,6,4	1.2956	4.59	0.00
34	9,8,7,6,4	1.2431	3.60	0.00
35	10,8,7,6,4	1.1956	2.69	0.00
36	10,8,7,7,4	1.1484	2.19	0.00
37	10,8,8,7,4	1.1043	1.75	0.00
38	10,9,8,7,4	1.0626	1.46	0.00
39	11,9,8,7,4	1.0224	1.14	0.00
40	11,9,8,7,5	0.9815	0.73	0.00
41	11,9,8,8,5	0.9451	0.63	0.00
42	11,9,9,8,5	0.9075	0.46	0.00
43	11,10,9,8,5	0.8588	0.25	0.00
44	11,10,9,8,6	0.8353	0.28	0.00
45	12,10,9,8,6	0.7976	0.21	0.00

Tableau 5-7 : Allocation optimale des bits à chacun des sous-dictionnaires de dimensions 4-3-3-3-3 en fonction du nombre total de bits, avec prédiction MA-1 et mesure DEP; distorsion spectrale moyenne et trames (en %) dont la distorsion spectrale SD est comprise entre 2 et 4 dB, ou supérieure à 4 dB.

Avec le schéma 40-[$(4,3,3,3,3)_{11,9,8,7,5}$], MA-1, DEP, 2 % des vecteurs de LSP quantifiés sont instables et doivent être réordonnés. De plus, à l'intérieur d'un vecteur quantifié, 5 % de LSP consécutifs (pris 2 à 2) sont trop proches et doivent être déplacés. Par conséquent, nous avons ajouté à l'algorithme d'encodage des LSP, un bloc qui teste les LSP quantifiés et les réordonne ou les éloigne les uns des autres de 50 Hz, si nécessaire. Les tests auditifs ont montré que la transparence est ainsi atteinte pour tous les fichiers de BD-TEST. Toutefois dans quelques cas rares, la qualité du signal reconstruit $\hat{s}(n)$ est légèrement dégradée.

5.6.3.6 Réduction de la complexité

Pour la suite de l'expérimentation nous avons optimisé le nombre et la dimension des sous-vecteurs pour la quantification SVQ, et testé la quantification S-MSVQ, ceci dans l'optique de réduire la complexité algorithmique.

Nous avons fixé un budget avoisinant 40 bits/trame. Nous avons utilisé la mesure DE, puisque la mesure DEP double le nombre de multiplications nécessaires à la quantification des LSP, alors qu'elle ne réduit la distorsion spectrale moyenne que de 0.03 à 0.04 dB. Le Tableau 5-8 regroupe les résultats obtenus. Il décrit les différents schémas de quantification testés et caractérise la distorsion spectrale obtenue en utilisant la base de données BD-TEST. Il expose également la complexité algorithmique en termes de multiplications (mult.), additions (add.) et soustractions (sous.), ainsi que le nombre de valeurs doubles nécessaire en mémoire[8] (mém.). La complexité de calcul due à la prédiction d'ordre 1, soit 16 soustractions et multiplications, n'est pas incluse. De plus, les coefficients de prédictions ne sont plus moyennés par sous-vecteur, mais considérés individuellement. Pour la suite de la discussion, l'expression "meilleure configuration" désigne la meilleure configuration en termes de distorsion spectrale et le terme #n désigne le numéro (n) du schéma donné dans le Tableau 5-8. Les schémas #1 et #2 correspondent aux configurations transparentes décrites aux points 5.6.3.2 et 5.6.3.5.

Selon Paliwal et Atal [5-21], pour un budget de bits donné, les solutions optimales en termes de complexité sont celles qui assignent le même nombre de bits à chacun des sous-dictionnaires d'un quantificateur SVQ. Nous avons constaté que cette assertion n'est pas absolument vraie. En effet, dans le cas où les sous-vecteurs sont de dimensions fortement différentes, parfois la complexité est réduite si l'attribution de bits est inégale. Cependant, comme cette assertion est exacte dans la gamme des dimensions des sous-vecteurs considérés dans la pratique, nous l'avons retenue.

.

[8] Ce codeur a été implanté en code ANSI C, avec une arithmétique en virgule flottante et double précision (64 bits).

#	Schémas de quantification	$\overline{SD}$ [dB]	2<SD≤4 dB [%]	SD>4 dB [%]	Sous.	Mult.	Add.	Mém.
1	45-$[(4,3,3,3,3)_{12,10,9,8,6}]$	0.9889	0.35	0	21952	21952	16000	21952
2	40-$[(4,3,3,3,3)_{11,9,8,7,5}]$ MA-1, distance pondérée	0.9815	0.73	0	10976	21952	8000	10976
3	45-$[(3,3,3,3,4)_{9,9,9,9,9}]$	1.0017	0.47	0	8192	8192	5632	8192
4	44-$[(4,3,4,5)_{11,11,11,11}]$	0.9915	0.31	0	32768	32768	22528	32768
5	45-$[(3,2,3,2,2,4)_{8,8,8,7,7,7}]$	1.0789	0.93	0	3072	3072	1920	3072
6	40-$[(3,2,3,3,5)_{8,8,8,8,8}]$ MA-1	1.0297	1.47	0	4096	4096	2816	4096
7	41-$[(3,3,3,3,4)_{8,9,8,8,8}]$ MA-1	0.9818	1.03	0	4864	4864	3328	4864
8	43-$[(3,2,2,2,3,4)_{8,7,7,7,7,7}]$ MA-1	0.9594	0.76	0	2432	2432	1536	2432
9	30-$[(3,2,3,3,5)_{6,6,6,6,6}]$ MA-1	1.5164	11.89	0.02	1024	1024	704	1024
10	42-$[(3,2,3,3,5)_{6,6,6,6,6}; (4,6,6)_{4,4,4}]$ MA-1	1.0439	1.72	0	1280	1280	912	1280
11	42-$[(3,2,3,3,5)_{6,6,6,6,6}; (3,5,4,4)_{3,3,3,3}]$ MA-1	1.0429	1.64	0.01	1152	1152	800	1152
12	44-$[(3,2,3,3,5)_{6,6,6,6,6}; (4,6,6)_{4,5,5}]$ MA-1	0.9683	1.19	0	1472	1472	1072	1472
13	44-$[(3,2,3,3,5)_{6,6,6,6,6}; (4,4,3,5)_{4,3,3,4}]$ MA-1	0.9727	1.30	0	1224	1224	856	1224
14	43-$[(3,2,3,3,5)_{6,6,7,6,6}; (4,7,5)_{4,4,4}]$ MA-1	1.0073	1.63	0	1472	1472	1040	1472
15	28-$[(3,3,4,6)_{7,7,7,7}]$ MA-1	1.5720	15.09	0.02	2048	2048	1536	2048
16	42-$[(3,3,4,6)_{7,7,7,7}; (8,8)_{7,7}]$ MA-1	0.9920	1.31	0	4096	4096	3328	4096
17	43-$[(3,3,4,6)_{7,7,7,7}; (6,5,5)_{5,5,5}]$ MA-1	0.9748	1.23	0	2560	2560	1952	2560
18	43-$[(3,3,4,6)_{7,7,7,7}; (5,4,3,4)_{4,4,4,3}]$ MA-1	0.9784	1.26	0	2272	2272	1704	2272
19	21-$[(4,4,8)_{7,7,7}]$ MA-1	1.9946	44.50	0.31	2048	2048	1664	2048
20	42-$[(4,4,8)_{7,7,7}; (5,5,6)_{7,7,7}]$ MA-1	0.9773	1.34	0	4096	4096	3328	4096
21	41-$[(4,4,8)_{7,7,7}; (4,4,3,5)_{5,5,5,5}]$ MA-1	1.0344	1.53	0	2560	2560	2048	2560
22	41-$[(4,4,8)_{7,7,7}; (2,4,3,3,4)_{4,4,4,4,4}]$ MA-1	1.0372	1.63	0.01	2304	2304	1840	2304
23	42-$[(4,4,8)_{7,7,7}; (3,5,3,5)_{5,5,5,6}]$ MA-1	0.9901	1.21	0	2720	2720	2176	2720
24	42-$[(4,4,8)_{7,7,7}; (3,3,3,3,4)_{4,4,4,5,4}]$ MA-1	1.0016	1.30	0	2352	2352	1872	2352
25	43-$[(4,4,8)_{7,7,7}; (2,3,3,3,5)_{4,4,4,5,5}]$ MA-1	0.9654	1.10	0	2432	2432	1936	2432
26	14-$[(5,11)_{7,7}]$ MA-1	2.5283	77.30	2.35	2048	2048	1792	2048
27	42-$[(5,11)_{7,7;}\ (2,4,3,3,4)_{5,6,6,6,5}]$ MA-1	0.9656	0.96	0	2880	2880	2368	2880
28	42-$[(5,11)_{7,7;}\ (2,3,2,2,3,4)_{4,5,4,5,5,5}]$ MA-1	0.9763	0.98	0	2496	2496	2080	2496
29	41 $[(5,11)_{7,7;}(4,4,4,4)_{7,7,7,6}]$ MA-1	0.9869	1.09	0	3840	3840	3136	3840
30	42-$[(16)_6; (16)_6; (2,3,3,3,5)_{6,6,6,6,6}]$ MA-1	0.9473	0.98	0	3072	3072	2624	3072
31	41-$[(16)_6; (16)_6; (2,3,3,3,5)_{5,6,6,6,6}]$ MA-1	0.9756	1.16	0	3008	3008	2592	3008

Tableau 5-8 : Schémas de quantification (LSP) testés, avec les caractéristiques de la distorsion spectrale et de la complexité algorithmique.

Nous avons commencé par réduire la complexité du schéma #1. Pour un budget de 45 bits, nous avons testé tous les schémas possibles avec 5 sous-vecteurs (séparation du vecteur total en 5) d'au moins 2 LSP, et une attribution de 9 bits à chacun des sous-dictionnaires. Pour chaque configuration, nous avons entraîné les sous-dictionnaires et testé la quantification. La meilleure configuration est la #3. Elle est nettement moins complexe que la #1, pour une distorsion spectrale moyenne à peine supérieure. Nous avons également testé toutes les configurations possibles avec 4 séparations et une attribution de 11 bits par sous-dictionnaire, pour un budget total de 44 bits/trame. Dans ce cas, la meilleure configuration est la #4.

Ensuite, nous avons testé toutes les configurations avec 6 séparations et une attribution de 3 fois 6 bits et 3 fois 7 bits, pour un budget de 45 bits/trame. La meilleure configuration est la #5. Elle est moins complexe que les configurations #1, #3 et #4 mais ne permet pas d'atteindre la transparence selon le critère A. Nous observons ici un compromis entre la complexité et le nombre de sous-vecteurs utilisés. Ce compromis est typique de la quantification SVQ.

Nous avons essayé de réduire la complexité du schéma #2, en utilisant la prédiction MA-1 et une mesure DE. Nous avons testé toutes les configurations avec 5 séparations et une attribution de 8 bits par sous-dictionnaire. La meilleure configuration est la #6 mais elle n'atteint pas la transparence. Nous avons donc testé toutes les séparations possibles en attribuant une fois 9 bits et quatre fois 8 bits aux différents sous-dictionnaires. La meilleure configuration est la #7. Elle correspond à un budget de 41 bits/trames. Elle est transparente et moins complexe que la #2. Nous avons également étudié toutes les configurations avec 6 séparations. Pour atteindre la transparence, la meilleure configuration est la #8. Elle est moins complexe que la #7 mais correspond à un budget de 43 bits/trames.

Nous avons ensuite essayé de réduire la complexité de la configuration #7, en utilisant une quantification S-MSVQ (split-multistage VQ) à deux étages, telle que représentée à la Figure 5.6. Parmi les schémas décrits jusqu'ici, la configuration #7 nous semble la meilleure en termes de transparence, complexité et débit. En assignant le même nombre de bits à chacun des sous-vecteurs, la complexité (en termes de multiplications, soustractions et mémoire), ne dépend pas du nombre de séparations mais uniquement du nombre de bits attribués à chaque sous-dictionnaire. Ainsi, pour réduire la complexité de la configuration #7, nous ne devons pas attribuer plus de 7 bits à chaque sous-dictionnaire du premier et du second étage de quantification. Nous visions un budget total compris entre 40 et 42 bits. Pour le premier étage, nous avons étudié les configurations à 5, 4, 3 et 2 séparations. En effet, avec un tel budget et plus de 5 séparations, la transparence semble inaccessible.

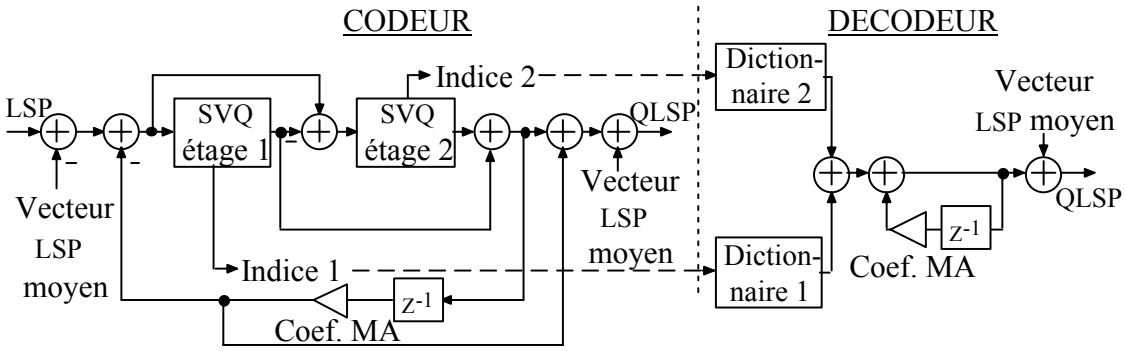

Figure 5.6 : Schéma de quantification S-MSVQ à 2 étages, avec prédiction MA-1.

Nous avons testé 5 séparations avec 6 bits par sous-dictionnaire sur le premier étage. La meilleure configuration est la #9. Elle correspond à un budget de 30 bits. Nous avons donc testé 3 et 4 séparations sur le deuxième étage en attribuant 4 et respectivement 3 bits à chaque sous-dictionnaire. Les meilleures configurations obtenues sont les #10 et #11. Comme elles ne sont pas transparentes, nous avons ajouté un bit puis deux sur le deuxième étage, et testé toutes les configurations possibles avec 3 et 4 séparations. La transparence n'est atteinte que pour un débit total de 44 bits/trame. Les meilleures configurations sont les #12 et #13. Pour réduire le débit, nous avons ajouté un bit sur le premier étage. Nous avons testé toutes les configurations possibles avec une fois 7 bits et quatre fois 6 bits (31 bits au total), puis utilisé 12 bits sur le second étage en testant toutes les configurations avec 3 séparations et 4 bits. La meilleure configuration est la #14, elle ne permet pas d'atteindre le critère A, mais en est très proche et est considérée comme transparente.

Nous avons considéré 4 séparations sur le premier étage avec 7 bits par sous-vecteur. La meilleure configuration est la #15. Nous avons testé 2, 3 et 4 séparations sur le deuxième étage. Avec 2 séparations et 7 bits par sous-dictionnaire, la meilleure configuration est la #16. Elle est transparente pour un débit total de 42 bits/trame. Par rapport à la configuration #7 elle requiert un bit/trame de plus et a une complexité légèrement plus faible. Par rapport aux configurations #12, #13 et #14, elle utilise un à deux bits de moins par trame, mais est nettement plus complexe. Dans le cas de 3 séparations sur le deuxième étage et 5 bits par sous-dictionnaire, la meilleure configuration est la #17. Elle est transparente et correspond à un débit de 43 bits/trame. En posant une fois 4 bits et 2 fois 5 bits sur le deuxième étage pour un débit total de 42 bits/trames, la transparence n'est jamais atteinte. Finalement nous avons testé 4 séparations sur le deuxième étage avec 3 fois 4 bits et une fois 3 bits pour un débit total de 43 bits. La meilleure configuration est la #18.

Nous avons testé 3 séparations avec 7 bits par sous-dictionnaire sur le premier étage. La meilleure configuration est la #19. Nous avons ensuite testé 3, 4 et 5 séparations sur le deuxième étage, avec respectivement 7, 5 et 4 bits par sous-dictionnaire et obtenu les schémas #20, #21 et #22. Alors que la configuration #20 est transparente pour un débit de 42 bits/trame, les schémas #21 et #22 ne le sont pas. Ils correspondent à un débit de 41 bits/trame. Pour le cas #21, avec 4 séparations sur le deuxième étage, un bit supplémentaire par trame permet d'atteindre la transparence (#23). Le schéma #24 avec 42 bits/trame n'est pas tout à fait transparent. Ainsi pour le cas #22 avec 5 séparations sur le deuxième étage, deux bits supplémentaires sont nécessaires, soit 43 bits/trames (#25).

Avec 2 séparations sur le premier étage et 7 bits par sous-dictionnaire, la meilleure configuration est la #26. En utilisant les mêmes procédures que précédemment et en testant 5 et 6 séparations sur le deuxième étage, nous avons obtenu les schémas #27 et #28 qui correspondent à un débit de 42 bits/trame. De même, avec 4 séparations sur le deuxième étage la meilleure configuration est la #29. Elle correspond à un débit de 41 bits/trame et sa complexité est relativement diminuée par rapport celle du schéma #7.

Finalement, nous avons testé la solution à trois étages et un budget de 42 bits/trames, donnée par la configuration #30. Comme cette configuration est plus que transparente, nous avons testé une solution à 41 bits, en supprimant un bit sur le troisième étage, et en testant donc toutes les configurations à une fois 5 bits et quatre fois 6 bits sur cet étage. La meilleure configuration est la #31.

5.6.4 Choix d'un schéma de quantification pour le codeur propriétaire

Les schémas de quantification testés, utilisant la prédiction MA-1 et permettant d'obtenir la transparente selon le critère A, sont regroupé dans la première partie du Tableau 5-9. Ils sont d'abord ordonnés selon leur débit, puis selon leur complexité. Le Tableau 5-9 montre un compromis entre le débit et la complexité. Aucun des schémas les moins complexes n'a un débit aussi bas que le #7.

#	Schémas de quantification	$\overline{SD}$ [dB]	2<SD≤4 dB [%]	SD>4 dB [%]	Sous.	Mult.	Add.	Mém.
31	41-[$(16)_6$; $(16)_6$; $(2,3,3,3,5)_{5,6,6,6,6}$] MA-1	0.9756	1.16	0	3008	3008	2592	3008
29	41 [$(5,11)_{7,7}$; $(4,4,4,4)_{7,7,7,6}$] MA-1	0.9869	1.09	0	3840	3840	3136	3840
7	41-[$(3,3,3,3,4)_{8,9,8,8,8}$] MA-1	0.9818	1.03	0	4864	4864	3328	4864
24	42-[$(4,4,8)_{7,7,7}$;$(3,3,3,3,4)_{4,4,4,5,4}$] MA-1	1.0016	1.30	0	2352	2352	1872	2352
28	42-[$(5,11)_{7,7}$; $(2,3,2,2,3,4)_{4,5,4,5,5,5}$] MA-1	0.9763	0.98	0	2496	2496	2080	2496
23	42-[$(4,4,8)_{7,7,7}$; $(3,5,3,5)_{5,5,5,6}$] MA-1	0.9901	1.21	0	2720	2720	2176	2720
27	42-[$(5,11)_{7,7;}$ $(2,4,3,3,4)_{5,6,6,6,5}$] MA-1	0.9656	0.96	0	2880	2880	2368	2880
30	42-[$(16)_6$; $(16)_6$; $(2,3,3,3,5)_{6,6,6,6,6}$] MA-1	0.9473	0.98	0	3072	3072	2624	3072
20	42-[$(4,4,8)_{7,7,7}$; $(5,5,6)_{7,7,7}$] MA-1	0.9773	1.34	0	4096	4096	3328	4096
16	42-[$(3,3,4,6)_{7,7,7,7}$; $(8,8)_{7,7}$] MA-1	0.9920	1.31	0	4096	4096	3328	4096
14	43-[$(3,2,3,3,5)_{6,6,7,6,6}$; $(4,7,5)_{4,4,4}$] MA-1	1.0073	1.63	0	1472	1472	1040	1472
18	43-[$(3,3,4,6)_{7,7,7,7}$; $(5,4,3,4)_{4,4,4,3}$] MA-1	0.9784	1.26	0	2272	2272	1704	2272
8	43-[$(3,2,2,2,3,4)_{8,7,7,7,7,7}$] MA-1	0.9594	0.76	0	2432	2432	1536	2432
25	43-[$(4,4,8)_{7,7,7}$; $(2,3,3,3,5)_{4,4,4,5,5}$] MA-1	0.9654	1.10	0	2432	2432	1936	2432
17	43-[$(3,3,4,6)_{7,7,7,7}$; $(6,5,5)_{5,5,5}$] MA-1	0.9748	1.23	0	2560	2560	1952	2560
13	44-[$(3,2,3,3,5)_{6,6,6,6,6}$; $(4,4,3,5)_{4,3,3,4}$] MA-1	0.9727	1.30	0	1224	1224	856	1224
12	44-[$(3,2,3,3,5)_{6,6,6,6,6}$; $(4,6,6)_{4,5,5}$] MA-1	0.9683	1.19	0	1472	1472	1072	1472
32	41-[$(6,10)_{8,8}$; $(2,4,3,3,4)_{5,5,5,5,5}$] MA-1	0.9749	0.97	0	4608	4608	3936	4608
33	42-[$(4,4,8)_{8,8,8}$; $(6,5,5)_{6,6,6}$] MA-1	0.9477	1.01	0	5120	5120	4160	5120
34	46- [$(9,7)_{8,8}$; $(3,3,3,3,4)_{6,7,7,5,5}$] MA-1	0.8761	0.46	0	5280	5280	4384	5280
35	36- [$(9,7)_{8,8}$; $(5,4,7)_{7,7,6}$] MA-1	1.1979	2.64	0	5696	5696	4864	5696

Tableau 5-9 : Schémas de quantification (LSP) testés, qui sont absolument transparents selon le critère A et qui utilisent une prédiction MA-1, avec les caractéristiques de la distorsion spectrale et de la complexité algorithmique.

Nous avons réalisé des tests auditifs à l'aveugle, pour toutes les configurations regroupées au Tableau 5-9 et selon la Figure 5.1. Ces tests montrent que le schéma #27 donne la meilleure qualité de signal reconstruit. Le schéma #27 est non seulement le meilleur subjectivement, mais il représente un bon compromis en termes de qualité, complexité et débit. Il a été retenu pour être implanté dans le codeur P-MRWB-ACELP.

Bien que le schéma #27 soit le meilleur en qualité subjective, les schémas #8 et #30 donnent de meilleurs résultats objectifs en termes de distorsion spectrale. Toutefois, la qualité subjective obtenue avec les schémas #8 et #30 est très proche de celle obtenue avec le schéma #27.

Les tests auditifs avec la base de données BD-TEST ont montré que le critère A n'est pas suffisant pour atteindre la transparence. En effet, si plus de 1% des trames donnent une distorsion spectrale supérieure à 2 dB, alors le signal quantifié présente des distorsions audibles. Ainsi, nous définissons un nouveau critère de transparence (critère C). Ce critère est défini comme suit : considérant la mesure de distorsion spectrale SD_n donnée par l'équation (3.8) où les bornes d'intégration f_1 et f_2 valent 0 et 7000 Hz, alors :

- La distorsion spectrale moyenne doit être inférieure à 1.0 dB;
- Aucune trame ne doit entraîner une distorsion spectrale supérieure à 4 dB;
- Le nombre de trames pouvant entraîner une distorsion spectrale comprise entre 2 et 4 dB doit être inférieur à 1 %.

Finalement, avec un dernier test nous avons vérifié qu'en utilisant 2 ou 3 séparations sur le premier étage et 8 bits par sous-dictionnaire, puis respectivement 3 et 5 séparations sur le deuxième étage pour un débit total inférieur à 43 bits, le schéma #27 est toujours le meilleur. Naturellement, dans un tel cas la complexité augmente. Nous avons obtenu les schémas #32 et #33 regroupés dans la deuxième partie du Tableau 5-9. Des tests auditifs ont montré que bien que ces schémas soient transparents selon le critère C, la meilleure configuration reste la #27.

Le fait que le schéma #27 soit le meilleur parmi les quantificateurs à deux étages s'explique ainsi : le schéma #27 n'utilise que 14 bits sur le premier étage et ne réalise que deux séparations sur celui-ci. Ces 14 bits sont suffisants à modéliser pratiquement toute la corrélation intra-trame des LSP. Le résidu de quantification du premier étage n'est pratiquement plus corrélé et les 26 bits restant, attribués au 5 sous-dictionnaires du deuxième étage, permettent de quantifier avec suffisamment de précision la diversité (les détails) de ce vecteur résiduel non-corrélé. Le fait que le schéma #28 soit moins bon est dû à la séparation supplémentaire sur le deuxième étage (6 séparations), qui entraîne une perte en qualité. Les schémas à 3, 4 et 5 séparations sur le premier étage perdent la caractéristique précitée du schéma

#27. Cette théorie est consolidée par le fait que les schémas #20 et #16, bien que plus complexes que le #27, mais de débit égal, résultent en une qualité auditive inférieure. Le fait que le schéma #30 soit moins bon que le #27, et ceci bien que le résidu de prédiction des deux premiers étages de quantification ne contienne plus de corrélation intra-trame, ne semble lié qu'à un aspect purement subjectif. Toutefois, le choix du schéma #27 par rapport au schéma #30 semble judicieux dans le cas où le canal de transmission serait bruité, puisqu'une quantification SVQ est alors plus robuste qu'une quantification MSVQ.

Une fois le schéma #27 implanté dans le codeur P-MRWB-ACELP, les coefficients de prédiction MA-1 et les différents sous-dictionnaires ont été ré-entraînés itérativement.

La troisième partie du Tableau 5-9 regroupe les résultats obtenus avec les paramètres et les schémas du WB-AMR. Ces résultats sont présentés à la Sous-section 5.6.5.

5.6.5 Conclusions de la section

L'expérimentation décrite à la Sous-section 5.6.3 nous a permis d'implanter un quantificateur spectral transparent, peu complexe et ne nécessitant que 42 bits par trame de signal traité, soit 2.1 kbits/s. Aucun standard basé sur la prédiction linéaire et encodant le signal en une seule bande de fréquences n'existe pour permettre une comparaison. De plus, parmi les schémas décrits à la Sous-section 5.6.2, seul le schéma de type SNVQ présenté par Guibé est transparent selon le critère C et aucun des différents auteurs n'utilise la même base de donnée de test que nous.

Pour information, nous avons entraîné et testé les schémas de quantification du WB-AMR de l'ETSI, avec les bases de données BD-TRAIN et BD-TEST. La troisième partie du Tableau 5-9 regroupe les résultats obtenus. Une comparaison entre le schéma #27 et les schémas #34 et #35 ne peut être rigoureuse. En effet, les schémas #34 et #35 sont utilisées pour quantifier des ISP correspondant à une largeur de bande comprise entre 50 à 6400 Hz, et non pas des LSP correspondant à une largeur de bande comprise entre 50-7000 Hz comme c'est le cas pour le schéma #27. Toutefois, bien que les schémas #34 et #35 encodent une bande de fréquences plus étroite, ils sont plus complexes que tous les autres schémas du Tableau 5-9. Quant au schéma #34, il présente une qualité meilleure que celle donnée par le critère C. Une telle qualité se justifie par le fait que le WB-AMR est un codeur en virgule fixe alors que nos tests sont réalisés en virgule flottante. Un test en virgule

fixe entraînerait une distorsion supérieure. De plus, la transmission via un canal bruité entraînerait d'ultérieures dégradations.

Il faudrait tester nos configurations avec des erreurs de canal. Dans ce cas, il est fort possible qu'une autre configuration que la #27 soit choisie. En effet, cette configuration n'exploite que peu la sensibilité spectrale localisée des LSP sur le premier étage de quantification, permettant de contrer les problèmes liés aux erreurs de transmission.

5.7 Contributions à la réduction du bruit de quantification, d'un codeur de type ACELP pour la parole en bande élargie

En général, la parole encodée, décodée et reconstruite par un algorithme de type CELP souffre de plusieurs distorsions : une dégradation qualifiée de "rugosité" (roughness), plus marquée pour les voix de femmes que pour les voix d'hommes; une légère réverbération contenue dans certains segments de parole voisée; un bruit de fond comparable à un bruissement ou à un grésillement pendant les transitions rapides entre segments voisés et non-voisés. Ces distorsions sont liées aux limites d'un codeur CELP dont la taille des dictionnaires est restreinte et ne suffit pas à modéliser les hautes fréquences du signal, ainsi que ses transitions rapides [5-22].

La qualité d'un algorithme ACELP dépend de la richesse de son dictionnaire d'excitation algébrique. Cet algorithme est efficace pour le codage de la parole en bande étroite. Par contre, pour le traitement de la parole en bande élargie, des difficultés apparaissent. Celles-ci sont non seulement liées à une complexité algorithmique élevée, mais également à la qualité du signal reconstruit. En effet, un codeur de parole de type ACELP, fonctionnant à un débit limité, produit un signal reconstruit bruité en haute fréquence [5-23]. Or, le spectre de fréquences de la parole en bande élargie a une pente importante : l'énergie du signal chute d'environ 35 dB entre les basses et les hautes fréquences (cf. Figure 2.2). Ainsi, pour le codage de la parole en bande élargie, diverses particularités doivent être ajoutées au codeur ACELP pour qu'il produise un signal reconstruit adéquat. Celles-ci sont liées à l'extraction des excitations adaptative et algébrique qui doivent modéliser au mieux les hautes fréquences du signal.

Nous avons été confrontés aux difficultés précitées, puisque notre codeur ACELP de base, implanté en langage C, produisait un signal reconstruit fortement bruité et de qualité absolument insuffisante. La Sous-section 5.7.1 décrit la nature des problèmes de reconstruction et présente les innovations développées pour résoudre ces problèmes.

5.7.1 Problèmes de reconstruction du signal, recherche de solutions et brevets

Même en utilisant un dictionnaire d'excitation algébrique très riche, notre codeur ACELP de base produisait un signal reconstruit de qualité globale fort insuffisante. Ce signal manquait de présence, son niveau d'énergie était variable et le timbre de la voix du locuteur était méconnaissable. D'autre part, le signal reconstruit était corrompu par trois types de bruits :

- Un bruit harmonique en haute fréquence, appelé "comb-like noise";
- Un fort bruit en haute fréquence, tel un bruit de quantification;
- Un bruit en basse fréquence, appelé "rumbling noise", tel un balai de paille frappé à intervalles réguliers sur le sol.

Les problèmes précités ont d'abord été considérés séparément. Ils sont discutés aux points 5.7.1.1 et 5.7.1.2. En leur cherchant une solution, nous avons constaté qu'ils sont liés et proviennent de l'extraction de l'excitation totale. Or, celle-ci se fait par un algorithme en boucle fermée et il est difficile d'isoler un problème particulier dans un tel algorithme. Afin d'obtenir un signal reconstruit de bonne qualité, nous avons modifié différents blocs de base du codeur ACELP et développé des solutions innovatrices. Celles-ci font l'objet de trois brevets déposés en juillet 2002.

5.7.1.1 Qualité insuffisante du signal reconstruit

Le problème lié à la qualité insuffisante du signal reconstruit semblait complexe, puisque même avec un dictionnaire algébrique contenant 30 impulsions par excitation (6 impulsions par piste), la qualité du signal reconstruit était pauvre. Nous avons tenté plusieurs stratégies :

- Réglage fin des paramètres du filtre de pondération $W(z)$, utilisé pour extraire les excitations en minimisant l'erreur de quantification perçue par l'oreille (cf. Sous-section 2.4.2, équation (2.20)).
- Correction de l'effet clairsemé de l'excitation algébrique, dû à ses nombreuses composantes nulles qui introduisent des artefacts dans le signal reconstruit. Cet effet est plus conséquent pour un signal non-voisé que pour un signal voisé. En effet, avec un signal non voisé, l'excitation algébrique participe davantage que l'excitation adaptative à la reconstruction du signal. Pour corriger cet effet, nous avons essayé d'ajouter une composante aléatoire à la phase de l'excitation algébrique. Cet ajout est fonction du gain du dictionnaire adaptatif et est opéré à

l'encodeur et au décodeur. Nous avons également essayé de lisser la puissance de l'excitation algébrique. Nous avons donc filtré cette excitation au niveau du décodeur, avec un filtre à phase aléatoire en haute fréquence. Ces méthodes sont décrites par Hagen et *al.* en [5-23].

- Introduction d'une excitation algébrique adaptative. Si le gain du dictionnaire adaptatif est inférieur à un seuil donné, nous avons essayé de remplacer l'excitation adaptative par une excitation innovatrice plus riche. Cette méthode est décrite par Gerson et Jasiuk en [5-24].

Aucune des stratégies précitées n'a été concluante. Nous avons alors essayé d'éliminer les différents bruits du signal. Les solutions trouvées nous ont permis d'obtenir une qualité du signal reconstruit satisfaisante.

5.7.1.2 Les bruits

Le signal reconstruit était corrompu par les trois types de bruits que nous avons traités séparément.

Le bruit harmonique

Le bruit harmonique est lié à l'excitation adaptative et en particulier à la répétition des échantillons qu'elle introduit lorsque le délai tonal T est inférieur à la durée d'une sous-trame de parole (80 échantillons ou 5 ms). Pour de tels délais, une partie des échantillons de l'excitation adaptative est reconstruite artificiellement et ne correspond pas à des échantillons contenus dans l'excitation passée. De plus, le bruit harmonique est lié au gain adaptatif, qui n'est extrait qu'une seule fois pour la totalité de la bande de fréquences. Ceci induit parfois la création de composantes indésirables dans les hautes fréquences du signal reconstruit [5-22]. En outre, le bruit de quantification de l'excitation algébrique se répète et se propage une à plusieurs fois dans le dictionnaire adaptatif si le signal est voisé. Ce problème apparaît sur toute la bande de fréquences.

L'utilisation d'un filtre de correction "harmonique", réalisant un léger filtrage passe-bas de l'excitation adaptative, semble une solution possible pour réduire les harmoniques du signal et le bruit harmonique. Kroon et *al.* proposent cette solution en [5-22]. La Figure 3.1 illustre le schéma classique de quantification des excitations, où la cascade du filtre de pondération $W(z)$ (équation (2.20)) et du filtre de synthèse LPC est dédoublée pour la recherche des excitations. En réalité, dans un codeur ACELP usuel cette cascade n'est pas dédoublée (cf. Sous-section 3.2.1). Ce dédoublement, est cependant utile pour la suite de la discussion et il ne modifie pas le fonctionnement de

l'encodeur. La Figure 5.7 illustre ce même schéma, auquel le filtre de correction harmonique a été ajouté.

Un filtre de correction $C(z)$ possible est donné par :

$$C(z) = 1 + c_1 z^{-1}, \tag{5.10}$$

où c_1 vaut par exemple 0.2. Pour ne pas diminuer la qualité du signal reconstruit non-voisé, il faut éviter de trop filtrer les hautes fréquences.

Le bruit en haute-fréquence

Le bruit en haute fréquence est introduit par d'anciennes excitations algébriques présentes dans le dictionnaire adaptatif. Lors de l'extraction de l'excitation adaptative par maximisation du second terme de l'équation (3.36), les composantes en haute fréquence sont pratiquement ignorées car le critère de sélection favorise les basses fréquences. Ainsi, des hautes fréquences indésirables peuvent être contenues dans l'excitation adaptative sélectionnée et répétées. Elles sont particulièrement perçues durant les trames de parole voisée.

Un filtre de correction "haute fréquence (HF)" peut être utilisé pour éliminer les composantes indésirables en haute fréquence de l'excitation algébrique, avant la mise à jour du dictionnaire adaptatif. En principe, ce filtre peut être de type RIF ou RII de degré suffisant pour couper les hautes fréquences. La Figure 5.8 illustre le schéma de quantification CS-ACELP classique, auquel le filtre de correction HF proposé ici a été ajouté.

L'introduction simultanée du filtre de correction harmonique et du filtre de correction "haute-fréquence" a été testée. Cette solution permet une amélioration de la qualité durant les trames de parole voisée. Afin de réduire la complexité, ces deux filtres ont été combinés et implantés comme un "pré-filtre" du dictionnaire adaptatif. Ce pré-filtre, appelé ici filtre de correction "totale", est illustré à la Figure 5.9. Il est décrit à la Sous-section 6.2.11. Cette solution est innovatrice par sa double fonction et par l'emplacement du filtrage avant le dictionnaire adaptatif. Elle fait l'objet d'un brevet déposé en juillet 2002 [5-25].

En [5-22] Kroon et Atal proposent d'utiliser le filtre de correction illustré à la Figure 5.7, que nous avons appelé filtre de correction harmonique. Ce filtre est soit fixe, soit fonction du gain adaptatif. Il est utilisé comme un post-filtrage du dictionnaire adaptatif et non comme un pré-filtrage.

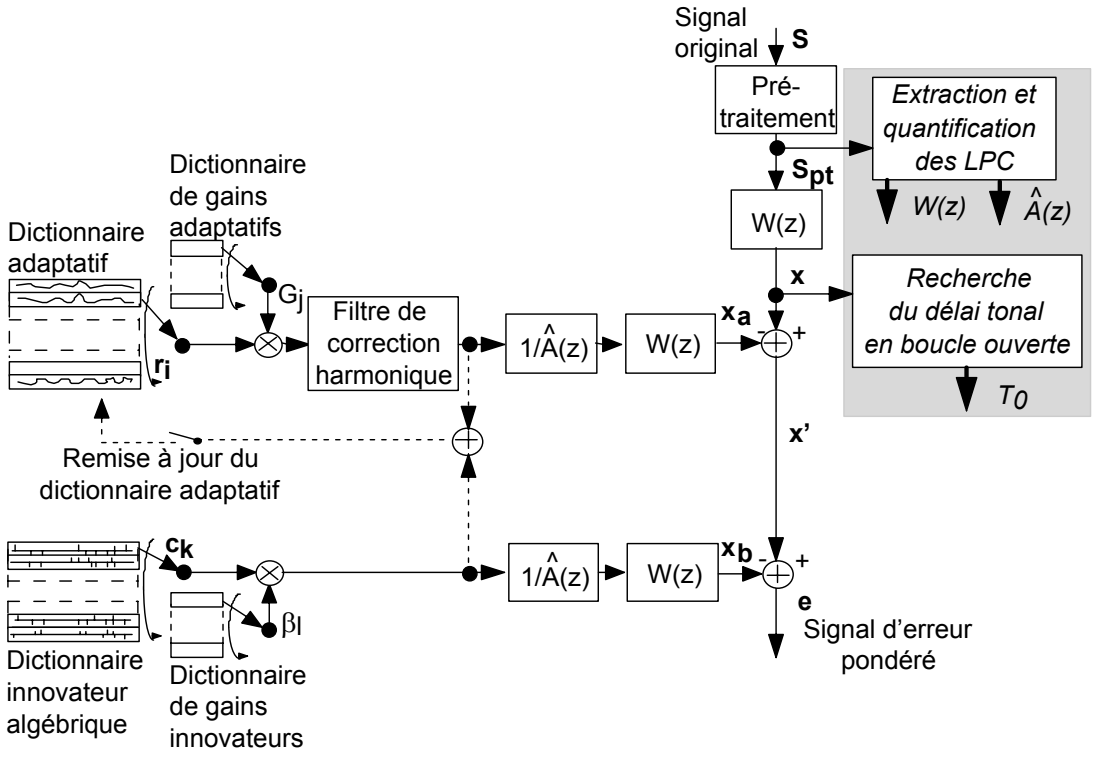

Figure 5.7 : Codeur CS-ACELP avec filtre de correction "harmonique".

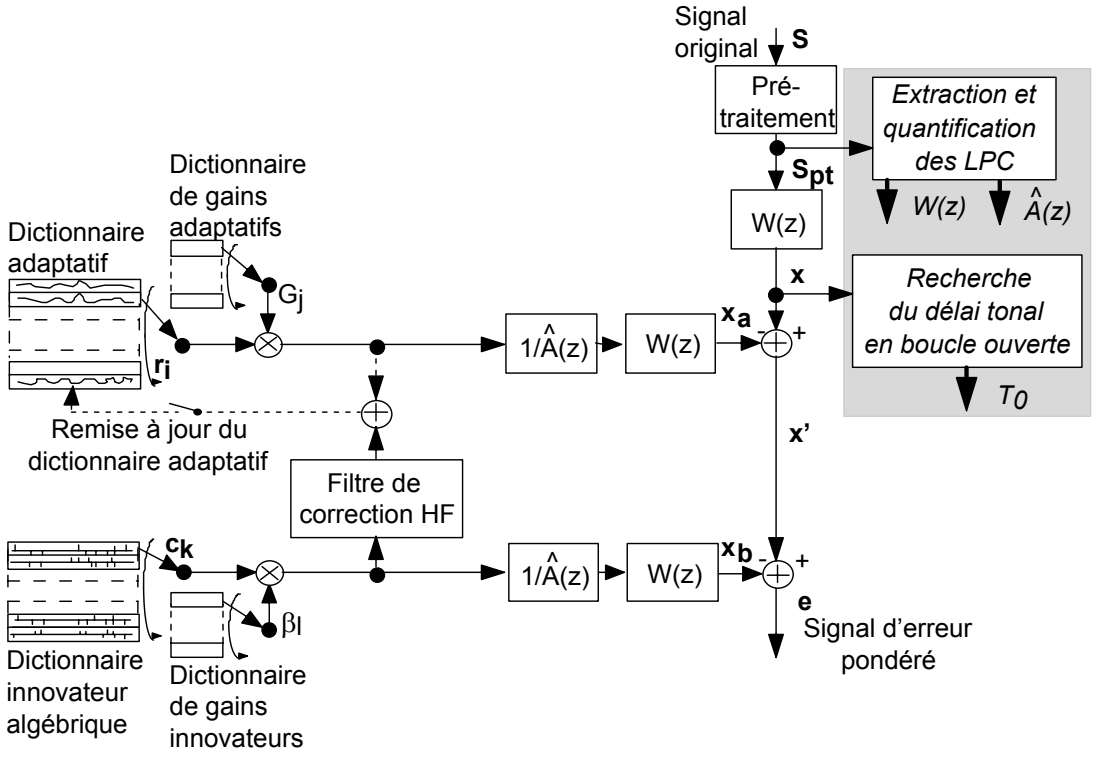

Figure 5.8 : Codeur CS-ACELP avec filtre de correction "haute fréquence" (HF).

Une solution où l'excitation adaptative sélectionnée est filtrée passe-bas lors de l'interpolation, a également été mentionnée par le Dr R. Salami [5-26]. Nous avons testé cette solution et constaté que la nôtre est meilleure en termes de qualité de signal reconstruit. La société VoiceAge est co-auteur du nouveau standard WB-AMR. Dans ce standard, une fois le délai tonal en boucle fermée déterminé, le gain G_1 de l'excitation correspondante est calculé (cf. point 4.4.1.1). Un second gain G_2 est calculé sur la base de cette même excitation filtrée passe-bas (0-2.8 kHz). Dans le cas où $G_2 > G_1$, l'excitation filtrée est retenue pour la suite du codage. De plus, si le codeur fonctionne à un débit de 6.6 kbits/s, l'excitation adaptative est toujours filtrée passe-bas.

Avec notre solution, le filtrage est pris en compte lors de la minimisation de l'erreur, effectuée pour choisir l'excitation adaptative. Ce n'est pas le cas pour la méthode proposée par Kroon et Atal ou utilisée dans le WB-AMR.

Le filtre de correction totale améliore la qualité du signal reconstruit, mais ne résout pas tous les problèmes liés aux bruits de reconstruction. Avec cette seule solution, les trames de parole voisée sont encore corrompues par un bruit comparable à un sifflement. Ce bruit, plus perceptible en haute fréquence, provient de l'excitation algébrique qui introduit des artefacts indésirables. En effet, l'excitation algébrique étant totalement artificielle, elle ne ressemble pas à une source du signal de parole. Ces artefacts ne sont pas présents uniquement en haute fréquence, cependant c'est en ces fréquences que l'oreille les perçoit davantage. De nombreux articles consacrés au codage de la parole en bande étroite et à faible débit discutent le problème lié à ce bruit en haute fréquence. Ici, nous visons un codage à moyen débit mais traitons de la parole en bande élargie, plus difficile à quantifier que la parole en bande étroite.

Le résidu de prédiction de la parole en bande élargie présente une pente spectrale et n'est donc pas similaire à un bruit blanc. Ainsi, comme ici l'excitation algébrique est très sollicitée pour modéliser le résidu, un débit important est nécessaire à l'encoder. Diverses solutions sont proposées dans la littérature :

- En [5-23] Hagen et *al.* proposent d'ajouter une composante aléatoire à la phase de l'excitation innovatrice. Comme décrit au point 5.7.1.1, nous avons testé cette solution.
- En [5-24] Gerson et Jasiuk proposent d'utiliser un filtre de pondération harmonique du bruit. Ce filtre, $D(z)$, est cascadé au filtre de pondération $W(z)$ donné par l'équation (2.20) et est utilisé pour la sélection des excitations. Il est de la forme :

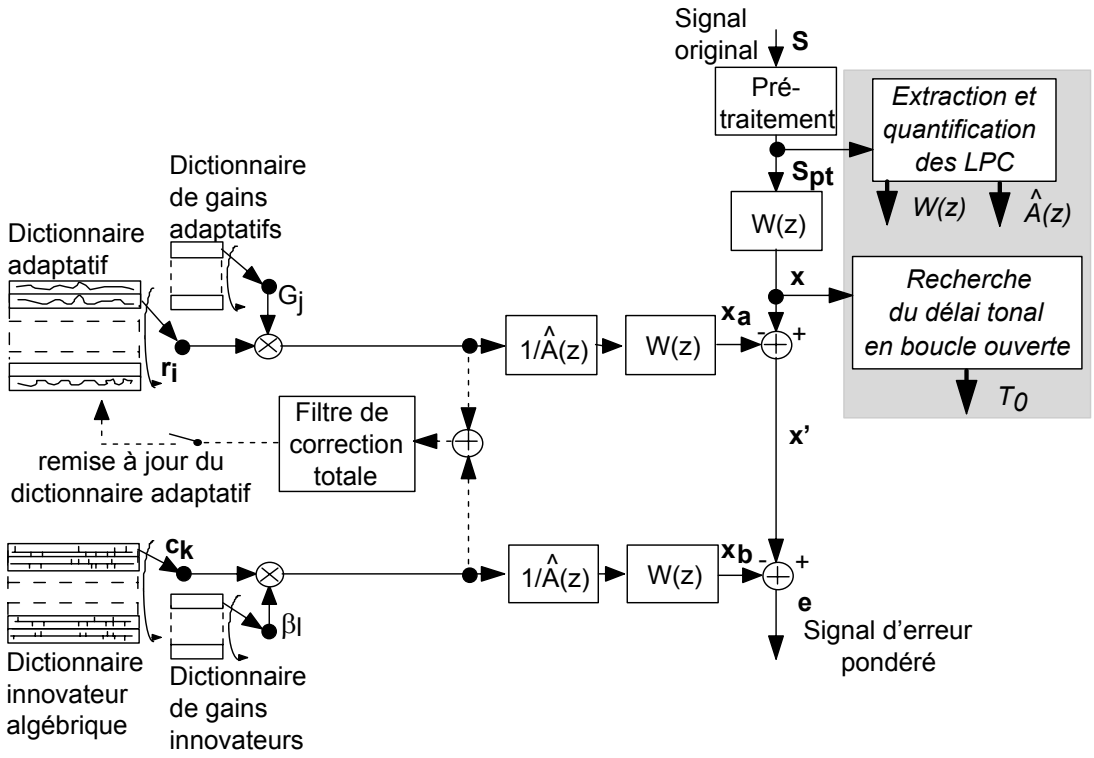

Figure 5.9 : Codeur CS-ACELP avec filtre de correction "totale".

$$D(z) = 1 - \varepsilon \sum_{k=-m1}^{m2} G_k z^{-(T+k)}, \quad \text{avec} \quad 0 \le \varepsilon \le 1.0, \tag{5.11}$$

où ε spécifie le montant de la pondération, T est le délai tonal entier ou l'un de ses multiples et les G_k sont les coefficients de prédiction à long-terme. L'utilisation de $D(z)$ est problématique. En effet, pour la recherche de l'excitation en boucle fermée selon les équations (3.30) et (3.36), la réponse à zéro $\hat{x}_a(n)$ du filtre $D(z)$ cascadé à $W(z)/\hat{A}(z)$ doit être recalculée pour chaque nouvelle valeur de T testée (cf. Sous-section 3.4.1).

- En [5-27] Taniguchi et *al.* proposent une "mise en forme" de l'excitation totale (excitations adaptative et innovatrice multipliées par leur gain respectif) pour mettre à jour le dictionnaire adaptatif. Cette "mise en forme" non-linéaire efface les composantes non-périodiques de l'excitation totale en fonction du délai tonal. Cette implémentation est illustrée à la Figure 5.9, si la "mise en forme" remplace le filtre de correction totale. Les auteurs proposent aussi une solution, où le gain de l'excitation innovatrice est réduit pour mettre à jour le dictionnaire adaptatif, mais est inchangé pour le calcul du signal de sortie.

- En [5-28] Miki et *al.* proposent d'utiliser pour les trames de parole voisée, une excitation innovatrice (de type aléatoire) synchronisée à la périodicité du signal. Cette excitation est répétée à intervalles réguliers correspondant au délai tonal T. Pour implanter cette technique, une partie du dictionnaire adaptatif usuel est remplie d'éléments aléatoires fixes et est utilisée comme partie intégrante du dictionnaire adaptatif. Cette partie du dictionnaire adaptatif est appelée dictionnaire fixe. En fonction des caractéristiques du signal d'entrée, le vecteur de code extrait du dictionnaire adaptatif correspond soit à une excitation passée, soit à un vecteur du dictionnaire fixe. S'il appartient au dictionnaire fixe, l'excitation innovatrice n'est pas synchronisée à T. Les auteurs proposent une variante, où le point de départ de l'excitation innovatrice est déplacé pour correspondre à l'emplacement du pic contenu dans la sortie du filtre de synthèse excité par l'excitation adaptative.
- En [5-29] Shoham propose d'imposer une contrainte à l'excitation innovatrice. Si l'excitation adaptative est suffisamment proche de l'excitation totale à modéliser, l'importance de l'excitation innovatrice est réduite et forcée à n'apporter qu'une très faible contribution. Ceci advient souvent pour les trames de parole voisée.

Il existe donc trois types d'approches : soit une composante aléatoire est ajoutée à la phase de l'excitation innovatrice (Hagen et *al.)*; soit la contribution innovatrice est rendue périodique (Gerson et Jasiuk, Miki et *al.*); soit le gain innovateur est contrôlé de façon adaptative (Taniguchi et *al.*, Shoham). La solution qui rend la contribution innovatrice périodique est complexe du point de vue algorithmique.

Après avoir testé plusieurs des solutions décrites ci-dessus, une solution propre au codeur P-MRWB-ACELP a été implantée. Cette solution est du type "contrôle de gain" et profite du pré-filtrage du dictionnaire adaptatif. Elle est illustrée à la Figure 5.10.

Avec cette solution, la contribution de l'excitation innovatrice est réduite si le gain de l'excitation adaptative est supérieur à un seuil préfixé, par exemple égal à 0.8. La réduction est réalisée en filtrant la contribution de l'excitation innovatrice avec un filtre $F(z)$ dont les coefficients dépendent de la valeur du gain adaptatif. Si $F(z)$ est d'ordre 1, et a une réponse impulsionnelle finie, il est donné par :

$$F(z) = \frac{1 + \lambda \cdot \min(G_j, 1) \cdot z^{-1}}{1 + \lambda \cdot \min(G_j, 1)}, \quad \text{avec} \quad \lambda = 0.5, \tag{5.12}$$

où G_j est le gain adaptatif. Cependant, c'est la contribution de l'excitation innovatrice **non-réduite** qui est utilisée pour la remise à jour du dictionnaire

adaptatif ainsi que pour l'extraction des paramètres qui lui sont relatifs (Figure 5.10 : switch en position a). La contribution réduite est utilisée pour générer le signal de sortie du décodeur et mettre à jour les mémoires des filtres $1/\hat{A}(z)$ et $W(z)$ de l'encodeur. Il est important de préserver la contribution innovatrice dans le dictionnaire adaptatif remis à jour, puisqu'ainsi la richesse de ce dictionnaire est conservée pour les fréquences les plus basses. Cette solution innovatrice élimine le bruit en haute fréquence. Elle fait l'objet d'un brevet déposé en juillet 2002 [5-30].

Notre solution est différente de celle de Taniguchi puisque nous contrôlons l'excitation innovatrice uniquement pour produire le signal de sortie. Taniguchi, contrôle cette excitation pour remettre à jour le dictionnaire adaptatif et produit la sortie normalement. Avec une telle solution la qualité du dictionnaire adaptatif est dégradée. Quant à Shoham, il ne précise pas si le contrôle de l'excitation n'est utilisé que pour mettre à jour le dictionnaire adaptatif, ou si ce contrôle est également utilisé pour calculer le signal de sortie.

Le bruit en basse fréquence

Le filtre de pondération perceptuelle de type formantique $W(z)$ (équation (2.20)) met en forme le bruit de quantification des codeurs LPAS selon les propriétés de perception de l'oreille humaine. La plupart des codeurs CELP et ACELP, décrits dans la littérature et dans les standards, utilisent le même filtre de pondération perceptuelle pour l'extraction de l'excitation adaptative et innovatrice. Cependant, la nature spectrale de l'excitation adaptative est différente de celle de l'excitation innovatrice.

L'excitation adaptative contribue principalement à la modélisation des basses fréquences des excitations voisées, qui ont une structure harmonique (cf. Figure 2.2). Par contre, l'excitation innovatrice contribue à la modélisation des basses fréquences des excitations non-voisées, ainsi qu'à la modélisation des hautes fréquences des excitations voisées et non-voisées. Pendant les périodes de parole non-voisée, l'excitation innovatrice apporte la meilleure contribution possible aux basses fréquences du signal et peut ajouter des artefacts dans la partie supérieure du spectre. L'erreur de codage de l'excitation innovatrice devrait être majoritairement pondérée par un filtre $W_T(z)$, contrôlant la pente spectrale du signal et dans une moindre mesure par le filtre de pondération perceptuelle usuel $W(z)$. Le filtre $W_T(z)$ est donné par :

$$W_T(z) = \frac{1-\mu_1 z^{-1}}{1-\mu_2 z^{-1}}. \qquad (5.13)$$

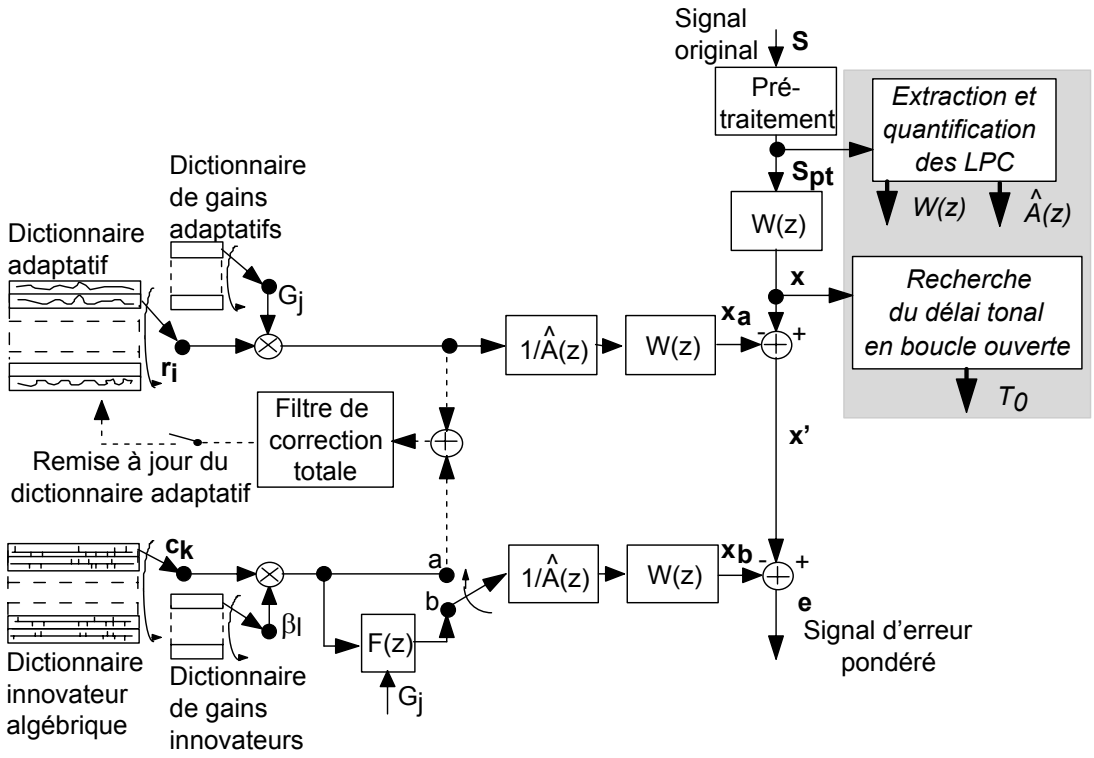

Figure 5.10 : Nouveau codeur CS-ACELP avec filtre de réduction de l'excitation innovatrice $F(z)$ pour le calcul du signal de sortie.

Pour l'extraction de l'excitation adaptative l'erreur de codage devrait être contrôlée par le filtre de pondération perceptuelle $W(z)$, et dans une moindre mesure par un filtre contrôlant la pente spectrale. Rappelons que la procédure d'extraction de l'excitation adaptative assigne peu d'importance aux hautes fréquences du signal.

A la Sous-section 2.5.1, nous proposons d'effectuer en entrée du codeur, une pré-accentuation du spectre du signal dans les hautes fréquences. Celle-ci peut être éliminée lors de l'extraction de l'excitation adaptative, en utilisant un filtre $\tilde{W}_T(z)$ de la forme :

$$\tilde{W}_T(z) = \frac{1}{1-\mu z^{-1}}, \tag{5.14}$$

où le coefficient μ a la même valeur que celui de l'équation (2.21).

La discussion qui précède montre la nécessité de différentier les fonctions de pondération utilisées pour l'extraction des deux types d'excitations, et ceci est d'autant plus important pour le codage de la parole en bande élargie. Il est donc nécessaire de changer le filtre de pondération perceptuelle $W(z)$ et le filtre de pondération de la pente spectrale $W_T(z)$, pour l'extraction de chacune des deux excitations.

Soit deux filtres de pondération perceptuelle indépendants :

$$W_1(z) = \frac{A(z/\gamma_{11})}{A(z/\gamma_{12})}, \text{ avec } 1 \geq \gamma_{11} \geq \gamma_{12} \geq 0, \tag{5.15}$$

utilisé pour la recherche de l'excitation adaptative et :

$$W_2(z) = \frac{A(z/\gamma_{21})}{A(z/\gamma_{22})}, \text{ avec } 1 \geq \gamma_{21} \geq \gamma_{22} \geq 0, \tag{5.16}$$

utilisé pour la recherche de l'excitation innovatrice, où :

$$A(z) = 1 + \sum_{k=1}^{p} a_p(k) z^{-k}. \tag{5.17}$$

L'utilisation de tels filtres avec γ_{11}=γ_{21}=1.0 est proposée par Harborg et *al.* en [5-31] et par Kroon et *al.* en [5-22], alors que de tels filtres avec les coefficients γ_{11}=1.0; γ_{12}=0.4; γ_{21}=0.9; γ_{22}=0.8 sont utilisés en [5-32] par Paulus et Schnitzler. L'article de Kroon et *al*. discute le codage de la parole en bande étroite alors que ceux de Harborg et *al*. et Paulus ou Schnitzler concernent les codeurs de parole en bande élargie décrits à l'Annexe C.1 aux points 3 et 13 (codeur similaire).

Harborg et *al.* utilisent le schéma illustré à la Figure 5.11, où le filtre $H(z) = 1/\hat{A}(z/\gamma)$ est équivalent à $W(z)/\hat{A}(z)$ si les coefficients de $W(z)$ sont quantifiés. L'excitation adaptative est choisie avec $\gamma = 0.7$. Le signal cible $d(n)$ utilisé est le signal de parole filtré par $W(z)$, auquel est soustraite la réponse à une excitation d'entrée nulle du filtre $1/\hat{A}(z/\gamma)$. Pour extraire l'excitation innovatrice, la nouvelle cible est obtenue en soustrayant à $d(n)$, la contribution de l'excitation adaptative et en posant $\gamma = 0.9$. Cette solution nous semble sous-optimale car pour la recherche de l'excitation innovatrice, le seul changement de la valeur de γ est insuffisant. En effet, pour une implantation optimale il faudrait filtrer le signal de parole dans le nouveau filtre $W(z)$ et tenir compte de la nouvelle réponse à une excitation d'entrée nulle du filtre $1/\hat{A}(z/\gamma)$.

Paulus et Schnitzler utilisent vraisemblablement un codeur similaire à celui illustré à la Figure 5.11, où le filtre $H(z) = W(z)/\hat{A}(z)$, et où une première estimation du délai tonal est réalisée en boucle ouverte. Paulus et Schnitzler n'expliquent pas comment le signal cible est calculé et remis à jour pour l'extraction de chacune des excitations.

Kroon et Atal utilisent certainement le schéma illustré à la Figure 5.11 et proposent d'adapter la valeur du dénominateur de $W(z)$ pour l'extraction de chacune des excitations en posant γ_{12}=0.5 et γ_{22}=0.9. Ils ne disent pas comment le signal cible est remis à jour.

A la place de changer les coefficients du filtre de pondération perceptuelle, nous proposons d'introduire deux filtres de pondération perceptuelle $W_1(z)$ et $W_2(z)$ en les implantant comme suit.

Soit $W_{F1}(z)$ et $W_{F2}(z)$, deux filtres de pondération perceptuelle dépendants, utilisés pour l'extraction des excitations adaptative et respectivement innovatrice. Ces filtres sont donnés par :

$$W_{F1}(z) = \tilde{W}_1(z) \quad \text{et} \quad W_{F2}(z) = \tilde{W}_1(z) \cdot \tilde{W}_2(z), \tag{5.18}$$

où

$$\tilde{W}_1(z) = \frac{A(z/\gamma_{11})}{A(z/\gamma_{12})}, \quad \text{avec} \quad 1 \geq \gamma_{11} \geq \gamma_{12} \geq 0, \tag{5.19}$$

$$\tilde{W}_2(z) = \frac{A(z/\gamma_{21})}{A(z/\gamma_{22})}, \quad \text{avec } \gamma_{21} = \gamma_{12} \quad \text{et} \quad 1 \geq \gamma_{11} \geq \gamma_{22} \geq 0 \tag{5.20}$$

($A(z)$ est donné par l'équation (5.17)). Les filtres $\tilde{W}_1(z)$ et $\tilde{W}_2(z)$, utilisés en cascade pour l'extraction de l'excitation innovatrice, sont liés puisque le dénominateur du premier est égal au numérateur du second. Ainsi :

$$W_{F2}(z) = \frac{A(z/\gamma_{11})}{A(z/\gamma_{22})}, \quad \text{avec} \quad 1 \geq \gamma_{11} \geq \gamma_{22} \geq 0. \tag{5.21}$$

Soit deux filtres de pondération de la pente spectrale :

$$W_{T1}(z) = \frac{1-\mu_{11}z^{-1}}{1-\mu_{12}z^{-1}} \quad \text{et} \quad W_{T2}(z) = \frac{1-\mu_{21}z^{-1}}{1-\mu_{22}z^{-1}}, \quad \text{avec} \quad \mu_{12} = \mu_{21}, \tag{5.22}$$

cascadés à $\tilde{W}_1(z)$ et respectivement à $\tilde{W}_2(z)$. Alors, les filtres de pondération perceptuelle $W_1(z)$ et $W_2(z)$ précités sont donnés par :

$$W_1(z) = \tilde{W}_1(z) \cdot W_{T1}(z) \quad \text{et} \quad W_2(z) = \tilde{W}_2(z) \cdot W_{T2}(z). \tag{5.23}$$

L'implantation de ces filtres est illustrée à la Figure 5.12. Le filtre $W_2(z)$ est cascadé au filtre $W_1(z)$ pour l'extraction de l'excitation innovatrice. Par conséquent $W_1(z)$ apparaît dans la boucle de recherche de cette excitation. Notons que la cascade des deux filtres permet une simplification algorithmique puisque $\gamma_{12} = \gamma_{21}$ et $\mu_{12} = \mu_{21}$.

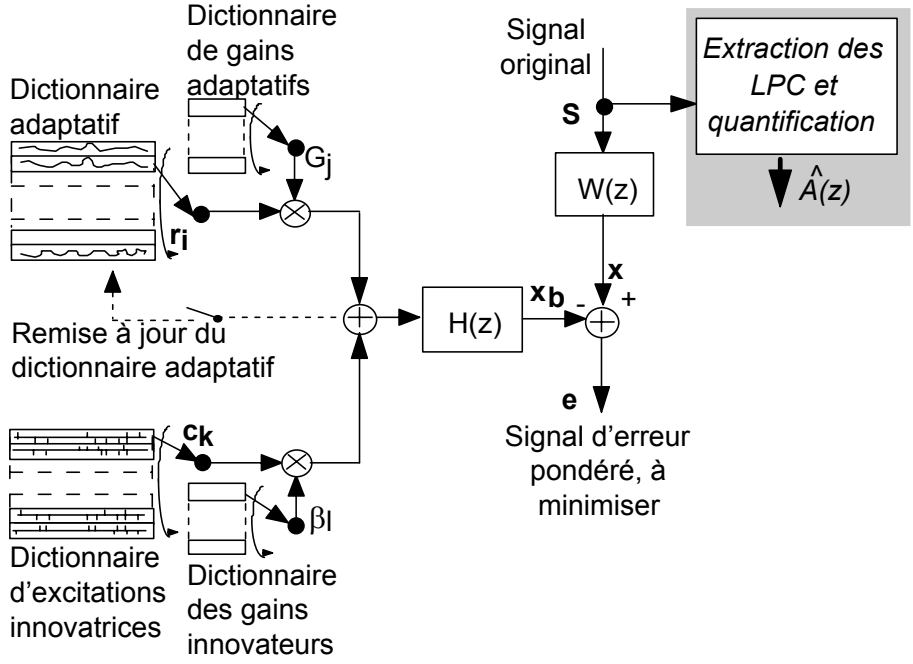

Figure 5.11 : Schéma du codeur CELP vraisemblablement utilisé par Harborg [5-31].

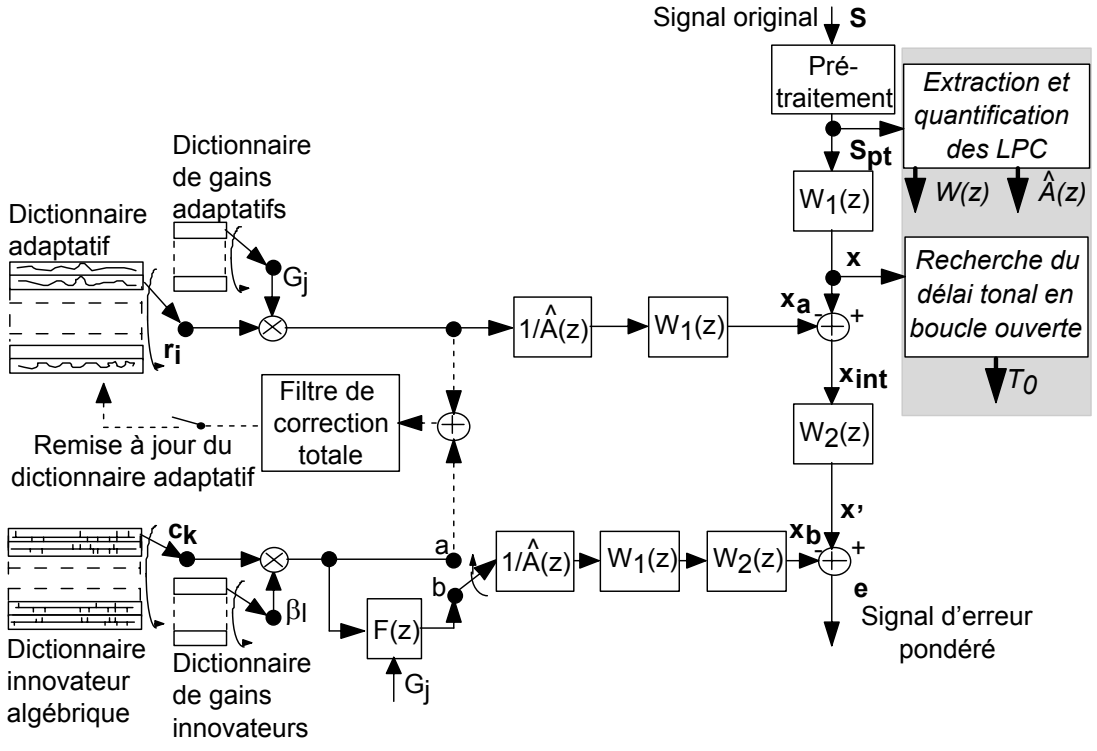

Figure 5.12 : Codeur P-MRWB-ACELP avec pondération du bruit différenciée en fonction de l'excitation extraite.

5.7.2 Conclusions de la section

L'implantation classique du codeur CELP, illustrée à la Figure 5.11, est moins complexe du point de vue algorithmique (calcul et mémoire) que l'implantation proposée pour le codeur P-MRWB-ACELP. Notons que la complexité algorithmique est fortement réduite si $\gamma_{11} = 1.0$ et si les coefficients $a_p(k)$ donnés par l'équation (5.17) et utilisés dans les équations (5.19) à (5.21) sont quantifiés. Toutefois, une telle complexité est nécessaire pour obtenir un signal de parole reconstruit satisfaisant. Notons encore que la pré-accentuation du spectre du signal dans les hautes fréquences, utilisée en entrée du codeur, peut renforcer la pondération de la pente spectrale pour l'extraction de l'excitation innovatrice.

5.8 Méthodes d'extraction de l'excitation adaptative

Les innovations décrites à la Section 5.7 ont permis de réduire, voire éliminer, le bruit de quantification et ont permis d'obtenir une qualité de signal reconstruit acceptable pour les voix d'hommes. Par contre, pour les voix de femmes de fréquence fondamentale moyenne élevée, ces innovations ne suffisent pas pour éliminer une certaine rugosité contenue dans les sons voisés. Cette rugosité est qualifiée de "rugosité tonale".

Jusqu'ici, la méthode d'extraction du délai tonal était basée sur celle du codeur G.729. Avec cette méthode, qualifié ici de "méthode A", le délai tonal en boucle ouverte T_0 est extrait toutes les 10 ms du signal prétraité et pondéré par $W_1(z)$. Ce signal est défini par $\mathbf{x}$ à la Figure 5.12. T_0 est recherché entre 40 et 286 échantillons, soit pour une fréquence fondamentale comprise entre 56 et 400 Hz. La recherche est réalisée en trois étapes : le maximum de la corrélation non normalisée $\tilde{R}_i$ $(i = 0,1,2)$, entre le signal traité et le signal décalé est calculé dans les intervalles de délais T_0 suivants : $g_0 = [40,\ldots,79]$, $g_1 = [80,\ldots,159]$ et $g_2 = [160,\ldots,286]$. La corrélation maximale de chaque intervalle est ensuite normalisée en fonction du délai correspondant. Les 3 maxima normalisés, R_i, sont alors comparés comme suit :

$$\begin{aligned} &W_{bo} = 0.85; \\ &R_{bo} = R_2, \\ &\text{si} \quad R_1 \geq W_{bo} \cdot R_{bo}, \quad \text{alors} \quad R_{bo} = R_1, \\ &\text{si} \quad R_0 \geq W_{bo} \cdot R_{bo}, \quad \text{alors} \quad R_{bo} = R_0. \end{aligned} \tag{5.24}$$

T_0 est le délai qui correspond à la corrélation normalisée R_{bo}. Le poids W_{bo} favorise les bas délais et permet d'éviter d'extraire un multiple du délai tonal. Ce poids est utile si le codeur contient un filtre de prédiction à long-terme. Par

contre, son effet est négligeable si ce filtre est remplacé par un dictionnaire adaptatif. Le délai tonal en boucle fermée est recherché toutes les 5 ms en fonction de T_0. T_1 et T_2 sont les délais tonals calculés en boucle fermée pour la première et respectivement la seconde sous-trame de 5 ms de la trame de 10 ms à partir de laquelle T_0 est extrait. T_1 est recherché entre $T_0 - 6$ et $T_0 + 6$, et T_2 entre $T_1 - 10$ et $T_1 + 9$. T_1 est calculé avec une résolution fractionnelle de 1/3 d'échantillon dans l'intervalle $[39 + 1/3, \ldots, 170 + 2/3]$ et avec une résolution non fractionnelle dans l'intervalle $[171, \ldots, 286]$. T_2 est toujours recherché avec une résolution fractionnelle de 1/3.

En [5-33], Kubin dit que dans un codeur CELP, la seule prédiction du délai tonal contribue à 75 % du rapport signal sur bruit du codeur. De plus, un codeur de parole en bande élargie doit encoder avec une grande précision les composantes en basse fréquence du signal, puisqu'elles sont très importantes du point de vue perceptif (cf. Chapitre 1). Ainsi, la méthode A basée sur l'algorithme du G.729 et simplement adaptée pour la bande élargie n'est pas suffisante, ce qui explique l'origine de la rugosité tonale.

La Sous-section 5.8.1 analyse le problème lié à l'extraction du délai tonal par la méthode A et propose des solutions pour éliminer la rugosité tonale. La Sous-section 5.8.2 discute ces solutions.

5.8.1 Analyse du problème et solutions

Si le délai tonal est inexact, l'excitation adaptative ne peut modéliser correctement les impulsions périodiques du signal ciblé. De plus, par sa nature, l'excitation innovatrice ne peut compenser cette erreur et modéliser ces impulsions sans introduire d'autres distorsions [5-34]. En outre, les algorithmes CELP n'arrivent pas à modéliser correctement les périodes de transition entre un son non voisé et un son voisé, puisque le filtre de prédiction à long-terme (ici l'excitation adaptative) n'est pas remis à jour assez fréquemment [5-35].

Pour vérifier que la "rugosité tonale" est liée à un calcul inexact du délai tonal, nous avons testé le codeur en utilisant la "méthode B". Avec cette méthode, la recherche du délai tonal n'est réalisée qu'en boucle fermée sur tout l'intervalle $[30 + 1/3, \ldots, 319 + 2/3]$, et avec une résolution de 1/3. Cet intervalle correspond à une fréquence fondamentale comprise entre 50 et 533 Hz. Il couvre les fréquences fondamentales des voix d'hommes et de femmes qui vont de 80 à 450 Hz, ainsi qu'une grande partie de celles des enfants qui vont de 200 à 600 Hz. En considérant les codeurs décrits à la Section 4.3 et à l'Annexe C.1, et dont les caractéristiques sont regroupées dans le Tableau 4-1, cet intervalle semble suffisant. Avec la méthode B, la rugosité tonale disparaît

presque complètement. Naturellement, cette méthode est prohibitive en termes de complexité de calcul et coûteuse en débit. Une méthode alternative a donc été développée. Pour cela, nous avons modifié la méthode A, pour atteindre une qualité de signal reconstruit aussi proche que possible de celle obtenue avec la méthode B.

Pour analyser le problème lié à la méthode A, nous avons comparé les histogrammes des délais tonals T_0, calculés en boucle ouverte toutes les 10 ms selon la méthode A, aux histogrammes des délais tonals T_B, calculés en boucle fermée toutes les 5 ms selon la méthode B. Nous avons utilisé toute la base de données d'entraînement. Pour ne pas fausser la comparaison, nous avons éliminé la quantification des gains et utilisé une excitation innovatrice contenant 30 impulsions par sous-trame de 5 ms. Cette comparaison montre que les délais tonals T_0 sont souvent inférieurs aux délais T_B. Nous avons testé différentes limites pour les intervalles g_0, g_1 et g_2, ainsi que différents poids W_{bo}, de valeurs supérieures à 0.85. Si W_{bo} est trop grand, des multiples du délai tonal peuvent être sélectionnés, ce qui n'est pas très grave ici car le filtre de prédiction à long-terme est remplacé par un dictionnaire adaptatif. Par contre, si W_{bo} est trop petit, un délai tonal erroné peut être sélectionné. Nous avons également testé une recherche de T_0 en plus de trois étapes, soit en cherchant le maximum de la corrélation non normalisée $\tilde{R}_i$, entre le signal traité et le signal décalé, pour un nombre d'intervalles g_i supérieur à 3 ($i = 0,1,2,...,I$ avec $I > 2$). Chaque intervalle g_i a été implanté de sorte à éviter qu'il contienne un délai T_0 et un de ses multiples entier nT_0 ($n \in \mathbb{N}$ et $n > 1$). Ces tests n'ont pas été concluants.

En observant l'évolution temporelle du délai tonal pour des voix de femmes, nous avons remarqué que fréquemment ce délai évolue rapidement et qu'une extraction en boucle ouverte toutes les 10 ms est limitative. Nous avons alors implanté la "méthode C", où le délai T_0 en boucle ouverte est extrait une fois toutes les 5 ms, et où le délai tonal entier en boucle fermée T est recherché dans l'intervalle de valeurs comprises entre $T_0 - 6$ et $T_0 + 6$. Nous avons alors constaté que par rapport à la méthode A, la méthode C permet d'améliorer la qualité du signal reconstruit. Avec une telle implantation, la complexité algorithmique ne change pratiquement pas. Par contre, le débit de transmission est accru puisqu'une quantification différentielle de T_2 par rapport à T_1 n'est plus possible, T_0 étant mis à jour toutes les 5 ms.

La recherche en boucle ouverte, effectuée selon la méthode A, requiert 40000 additions, 40003 multiplications et 3 divisions par trames de 10 ms. Cette même recherche, effectuée selon la méthode C, n'exige que 3 multiplications et 3 divisions supplémentaires. Ces complexités algorithmiques sont donc pratiquement égales. Elles sont calculées pour un

délai T_0 pouvant être compris entre 40 et 286. Par contre, pour la recherche en boucle fermée du délai tonal, la complexité algorithmique pour une trame de 10 ms passe de 18082 multiplications, 18033 additions et 49 divisions avec la méthode A, à 16402 multiplications, 16360 additions et 42 divisions avec la méthode C.

Les histogrammes des délais tonals T_0, calculés avec la méthode C, et les délais tonals T_B, calculés selon la méthode B, sont illustrés à la Figure 5.13. Leur comparaison montre que les délais T_0 sont encore légèrement inférieurs aux délais T_B. Nous avons donc modifié la méthode C ainsi : si le délai tonal T_0 calculé en boucle ouverte est petit, la recherche en boucle fermée est effectuée autour de T_0 **et** autour de $2*T_0$. La valeur sélectionnée qui correspond au plus grand gain est retenue. Nous avons alors constaté que par rapport à la méthode C, cette méthode (dite méthode D) permet une nouvelle amélioration de la qualité du signal reconstruit. Cependant, avec la méthode D et pour la recherche en boucle fermée du délai tonal, la complexité algorithmique pour une trame de 10 ms peut atteindre 32804 multiplications, 32720 additions et 84 divisions. Cette complexité est élevée. Pour une implantation du codeur sur DSP, il faut considérer une implantation selon les méthodes C ou D.

Sur la base de la méthode D, nous avons étudié les intervalles sur lesquels le délai tonal en boucle fermé T doit être recherché avec une précision de 1/3, 1/2 ou 1 échantillon. Nous avons considéré une utilisation optimale des bits disponibles. Dans ce cas (cas I), nous avons défini les étendues suivantes:

- Précision de 1/3 (interpolation par 3) pour les délais tonals entiers compris entre 29 et 120 inclus. Ici, les première et dernière valeurs fractionnelles possibles sont 28+1/3 (correspondant à 29-2/3) et 120+2/3 (si la valeur entière vaut 120).

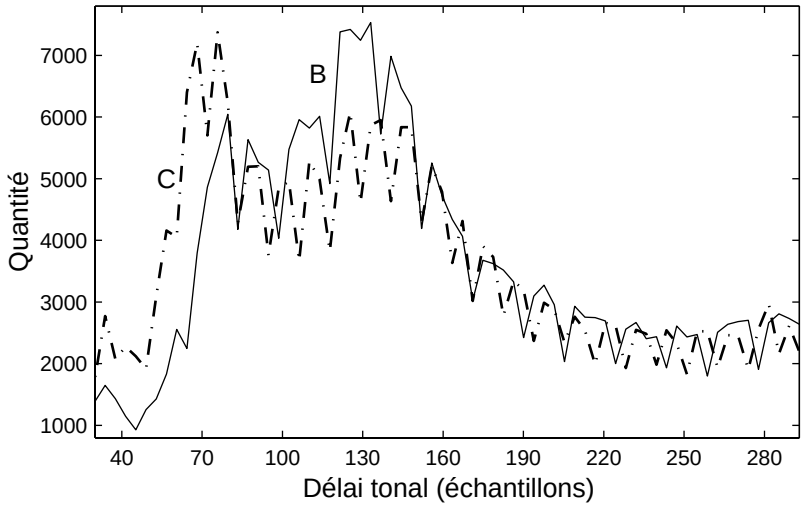

Figure 5.13 : Histogrammes des délais tonals obtenus avec les méthodes B et C.

- Précision de 1/2 (interpolation par 2) pour les délais tonals entiers compris entre 121 et 180 inclus. Ici les première et dernière valeurs fractionnelles possibles sont 120+1/2 (121-1/2) et 180+1/2.
- Précision de 1 pour un délai tonal entier compris entre 181 et 293.

Nous avons également testé une précision de 1/4 sur toute l'étendue des délais tonals entiers. Dans ce cas (cas J), l'intervalle de ces délais est compris entre 29 et 283. Ici, les première et dernière valeurs fractionnelles possibles sont 28+1/4 et 283+3/4. La différence en complexité entre les cas I et J est dans le pire des cas de 56 additions et 56 multiplications pour la recherche en boucle fermée. Par contre, la méthode J est moins complexe pour la recherche en boucle ouverte puisque son étendue totale est diminuée. Notons que la méthode J requiert plus de bits de transmission pour encoder le délai tonal.

Pour la recherche en boucle ouverte, nous avons à nouveau testé un nombre N_g d'intervalles g_i supérieur à 3. Cette fois, nous avons constaté une amélioration du signal reconstruit et avons opté pour une séparation en 10 intervalles comprenant chacun environ 25 valeurs de délais tonals T_0. La méthode E correspond à la méthode D à laquelle une telle séparation en 10 intervalles est appliquée. Par rapport à la méthode D, la complexité algorithmique de la méthode E est augmentée de 2240 additions, 2254 multiplications et 14 divisions, pour une trame de 10 ms.

Nous avons défini deux modes d'utilisation du codeur. Le mode 1 correspond à l'extraction du délai tonal selon le cas I et avec la méthode E. Il requiert 9 bits par sous-trame de 5 ms pour l'encodage du délai tonal fractionnel. Le mode 2 correspond à l'extraction du délai tonal selon le cas J avec la méthode E et requiert, 10 bits par sous-trame de 5 ms.

5.8.2 Conclusions de la section

Par trames de 10 ms, la recherche du délai tonal selon la méthode A requiert au total 58145 additions, 58197 multiplications et 53 divisions, ainsi que 15 bits pour la transmission du délai tonal fractionnel. Par contre, cette même recherche selon le mode 1 requiert au total 77864 multiplications, 77760 additions et 104 divisions, ainsi que 18 bits de transmission. Le mode 2 est à peine moins complexe que le mode 1 mais requiert 20 bits de transmission. L'augmentation en complexité des modes 1 et 2 par rapport à la méthode A est considérable mais nécessaire. En effet, pour les voix de femmes, une implantation selon la méthode A produit un signal reconstruit de nature rugueuse, alors que l'implantation selon les modes 1 et 2 permet d'obtenir un signal reconstruit de bonne qualité.

5.9 Modes d'extraction de l'excitation innovatrice

Le codeur P-MRWB-ACELP développé peut fonctionner selon trois modes différents pour extraire l'excitation innovatrice de type algébrique : les modes a1, a2 et a3. Soit le vecteur d'excitation de 80 échantillons divisé en cinq ensembles de 16 échantillons. Chaque ensemble est appelé "piste". Si l'indice $i = 0,1,...,4$ définit chaque piste et l'indice $l(i) = 0,1,...,15$ indique la position d'un échantillon dans une piste, alors la position m_k de la $k^{\text{ième}}$ impulsion sur le vecteur d'excitation sera donnée par l'équation (3.61):

$$m_k(l,i) = i + 5 \cdot l(i), \quad i = 0,1,2,3,4, \quad l(i) = 0,1,...,15.$$

Pour les modes a1, a2 et a3, 2, 3 et respectivement 4 impulsions sont attribuées à chaque piste. La position d'un échantillon non-nul dans une piste est quantifiée avec 4 bits. Si 4 impulsions sont attribuées à chaque piste, leur ordre d'encodage à l'intérieur d'une piste spécifie leur signe. Si 2 ou 3 impulsions sont attribuées à une piste, alors le signe de la première impulsion est encodé et le ou les signes restants sont spécifiés par l'ordre d'encodage. Ainsi, 45, 65 et 80 bits sont nécessaires à encoder l'excitation innovatrice des modes a1, a2 et respectivement a3 pour chacune des sous-trames de 5 ms.

5.10 Conclusions

Dans la première partie de ce chapitre nous avons décrit les contraintes et choix initiaux qui ont orienté le développement du codeur P-MRWB-ACELP, ainsi que les étapes de conception de ce codeur. Dans la seconde partie de ce chapitre nous avons présenté les contributions et innovations principales de ce travail de thèse.

Nous avons décrit l'implantation d'un quantificateur spectral transparent, peu complexe et ne nécessitant que 42 bits par trame de signal traité, soit 2.1 kbits/s. Différents modes d'utilisation du codeur P-MRWB-ACELP ont été développés : les modes 1 et 2 d'extraction de l'excitation adaptative, décrits à la Sous-section 5.8.1, ainsi que les modes d'extraction a1, a2 et a3 de l'excitation algébrique décrits à la Section 5.9.

En combinant les modes a1, a2 et a3 et les modes 1 et 2 nous obtenons trois modes d'utilisation du codeur. Le mode a1 ne permet pas d'obtenir une parole reconstruite de très bonne qualité, mais permet de réduire le débit et la complexité algorithmique. Nous l'associons au mode 1 pour former le **mode A.** Le mode A fonctionne à 14.3 kbits/s et est le mode le moins complexe du codeur P-MRWB-ACELP. Nous formons en outre le **mode B** qui associe le mode a2 au mode 2. Finalement, nous formons le **mode C** qui associe le mode a3 au mode 2. Le mode B fonctionne à 18.5 kbits/s alors que le mode C

fonctionne à 21.5 kbits/s. Ce dernier est le mode de fonctionnement proposé le plus complexe.

Le Tableau 5-10 fournit la liste du nombre de bits de quantification, attribués à chaque type de paramètres pour sa transmission, en fonction du mode de fonctionnement du codeur P-MRWB-ACELP.

Un de nos objectifs initiaux était de développer au moins un mode du codeur, fonctionnant à un débit inférieur ou égal à 14.25 kbits/s. En remplaçant le quantificateur spectral choisi, par un quantificateur spectral plus complexe mais ne nécessitant que 41 bits par trame de signal traité, par exemple le #7 du Tableau 5-8, cet objectif est atteint.

L'implantation d'un seul type de codeur nous a obligés à explorer minutieusement les limites de chaque partie du codeur, afin d'arriver à une bonne qualité de signal reconstruit. Cette méthode de travail nous a mené à développer diverses innovations. Ces innovations concernent spécialement le mode d'extraction des paramètres au niveau de l'encodeur. Elles pourraient être ajoutées à d'autres codeurs, tels que le nouveau standard WB-AMR, sans avoir à en modifier les dictionnaires et le décodeur. Il est fort probable que ces innovations en amélioreraient la qualité.

Avec du recul, je pense qu'initialement il aurait été judicieux d'implanter le squelette de deux ou trois codeurs en parallèle. Nous aurions par exemple pu développer un codeur de type SB-CELP en deux sous-bandes égales, tel que celui proposé dans l'article de McElroy et *al.* en [C-11], ainsi qu'un codeur de type SB-CELP en deux sous-bandes inégales, tel que celui proposé dans l'article de Paulus et Schnitzler [C-12] (cf. Annexe C.1). Naturellement, l'effort initial aurait considérablement augmenté, mais cet effort nous aurait permis d'évaluer les limites de chaque codeur et d'identifier le type de codeur le plus prometteur en qualité. En effet, nous aurions pu confronter une version basique des différents codeurs à l'aide d'une seule base de données de test. Dans ce cas, nous aurions réalisé un choix en évaluant les modifications à apporter à chaque type de codeur pour obtenir une bonne qualité de signal reconstruit.

Mode et débit	Paramètres	Nombre de bits par sous-trame	Nombre total de bits par trame
Mode A 14.3 kbits/s	LSP		42
	Délai tonal	9	36
	Gains adaptatif et innovateur	7	28
	Excitation innovatrice	45	180
	Total		286
Mode B 18.5 kbits/s	LSP		42
	Délai tonal	10	40
	Gains adaptatif et innovateur	7	28
	Excitation innovatrice	65	260
	Total		370
Mode C 21.5 kbits/s	LSP		42
	Délai tonal	10	40
	Gains adaptatif et innovateur	7	28
	Excitation innovatrice	80	320
	Total		430

Tableau 5-10 : Allocation des bits de quantification à chaque type de paramètres, pour les trois modes du codeur P-MRWB-ACELP.

5.11 Références

[5-1] L. Hanzo, F. Somerville et J. Woodard, "Standard forward-adaptive CELP codecs", Chapter 7, dans *Voice compression and communications*, pp. 207-278, IEEE Series on Digital & Mobile Communication, John Wiley & Sons, Inc., Publication, NY, USA, 2001.

[5-2] Document S4/SMG11 Tdoc 90/00, "AMR wideband performance requirements (WB-3) version 1.2", publié par le 3GPP, en : http://www.3gpp.org/ftp/tsg_sa/TSG_SA/TSGS_07/Docs/PDF/SP-000134.pdf Sept. 2002.

[5-3] ITU-T Recommendation G.722, "7 kHz audio – coding within 64 kbit/s", dans *Blue Book, fascicule III.4*, Melbourne, Australie, 1988.

[5-4] Document 3GPP TS 26.090 V3.1.0 (1999–12), dans ftp://ftp.3gpp.org/Specs/2000-09/R1999/26_series/ (14 nov. 2001).

[5-5] S. Pujalte et A. Moreno, "Wideband ACELP at 16 kbit/s with multi-band excitation", dans *Proc. European conference on speech communication and technology, EUROSPEECH 2001,* pp. 2001-2004, Aalborg, Danemark, 2001.

[5-6] A. Ubale et A. Gersho, "A multi-band CELP wideband speech coder", dans *Proc. IEEE Int. conf. acoustic, speech, signal processing 1997, ICASSP'97,* Vol. 2, pp. 1367-1370, Munich, Allemagne, Avr. 1997.

[5-7] J. Garofolo, L. Larnel, W. Fisher, J. Fiscus, D. Pallett et N. Dahlgren, "DARPA TIMIT, acoustic-phonetic continuous speech corpus CD-ROM", dans *US NIST Internal report 4930*, NTIS, U.S. Department of commerce, Port Royal Road, Springfield, Fév. 1993.

[5-8] "Bdsons, Base de données des sons du français", CD-ROMs enregistrés au CNET Lannion en 1985, édité par CedroM Technologies, et pressé par MPO France, Nov. 1990.

[5-9] "ITU multi-Lingual Speech Database for Telephonometry", ITU Recommendation P.50 Appendix I (02/98), http://www.itu.int/publications/itu-t/list-t-soft.html (13 Sept. 2002).

[5-10] ITU-T, "Rate-change: up- and down-sampling module", Chapter 3, dans *ITU-T Software tool library manual,* publié par l'ITU, Genève, Mai 1996.

[5-11] G. Guibé, H. T. How et L. Hanzo, "Comparative study of wideband speech spectral quantization schemes", dans *Proc. 3rd ITG Conference source and channel coding*, pp. 181-186, Munich, Allemagne, Jan. 2000.

[5-12] J.A. Gibbs et J.M. Hoskin, "LSP split vector quantization for wideband codecs with narrowband tandemming", dans *Proc. 2000 IEEE Workshop on speech coding*, pp. 120-122, Delavan, Wisconsin, USA, Sept. 2000.

[5-13] M. Ferhaoui et S. Van Gerven, "LSP quantization in wideband speech coders", dans *Proc. 1999 IEEE Workshop on speech coding*, pp. 25-27, Porvoo, Finlande, Juin 1999.

[5-14] S. Ragot, H. Lahdili et R. Lefebre, "Wideband LSF quantization by generalized voronoi codes", dans *Proc. European conference on speech communication and technology, EUROSPEECH 2001,* pp. 2319-2322, Aalborg, Danemark, Sept. 2001.

[5-15] Déjà référencé en [1-6].

[5-16] Y. Linde, A. Buzo, et R. Gray, "An algorithm for vector quantizer design", dans *IEEE Transactions on communications*, Vol. 28, pp. 84-95, Jan. 1980.

[5-17] J. Skoglund et J. Linden, "Predictive VQ for noisy channel spectrum coding: AR or MA?", dans *Proc. IEEE Int. conference on acoustics, speech and signal processing 1997, ICASSP-97*, Vol. 2, pp. 1351 -1354, Munich, Allemagne, Avr. 1997.

[5-18] H. Ohmuro, T. Moriya, K. Mano, et S. Miki, "Vector quantization of LSP parameters using moving average interframe prediction", dans *Electronics and communications in Japan,* Part 3, Vol. 77, No. 3, pp. 12-26, Mars 1994.

[5-19] S. L. Marple Jr., *Digital spectral analysis with applications*, Prentice-Hall International, Inc., 1987.

[5-20] J. G. Proakis et D. G. Manolakis, "Power spectrum estimation", Chapter 12, dans *Digital signal processing, principles, algorithms, and application*, Third Edition, pp. 896-968, Prentice-Hall International, Inc., 1996.

[5-21] K. Paliwal et S. Atal, "Efficient vector quantization of LPC Parameters at 24 bits/frame", dans *IEEE Transactions on speech and audio processing*, Vol. 1, No. 1, pp. 3-14, Jan. 1993.

[5-22] P. Kroon et S. Atal, "Strategies for improving the performance of CELP coders at low bit rates", dans *Proc. IEEE Int. conf. acoustic, speech, signal processing 1988, ICASSP'88*, pp. 151-154, New-York City, USA, 1988.

[5-23] R. Hagen, E. Ekudden, B. Johansson et W. Kleijn, "Removal of sparse-excitation artefacts in CELP", dans *Proc. IEEE Int. conf. acoustic, speech, signal processing 1998, ICASSP'98*, Vol. 1, pp. 145-148, Seattle, USA, Mai 1998.

[5-24] I. Gerson et M. Jasiuk, "Techniques for improving the performance of CELP-type speech coders", dans *IEEE Journal of selected areas in communications,* Vol.10, No. 5, pp. 858-865, Juin 1992.

[5-25] G. Biundo, M. Ansorge et B. Carnero, "Codeur de parole", brevet EP 02 015 918.2, déposé en juillet 2002.

[5-26] Communication privée à l'occasion de la visite du Dr. R. Salami (VoiceAge / Université de Scherbrooke) à Neuchâtel, le 8 Fév. 2000.

[5-27] T. Taniguchi, M. Jonhson et Y. Ohta, "Pitch sharpening for perceptually improved CELP, and the sparse-delta codebook for reduced computation", dans *Proc. IEEE Int. conf. acoustic, speech, signal processing 1991*, *ICASSP'91*, pp. 241-244, Toronto, Canada, 1991.

[5-28] S. Miki, K. Mano, T. Moriya, K. Oguchi et H. Ohmuro, "A pitch synchronous innovation CELP (PSI-CELP) coder for 2-4 kbit/s", dans *Proc. IEEE Int. conf. acoustic, speech, signal processing 1994, ICASSP'94,* pp. II-113 – II-116, Adelaide, Australie, 1994.

[5-29] P. Y. Shoham, "Constrained-stochastic excitation coding of speech at 4.8 kb", dans *Advances in speech coding*, pp. 339-348, édité par Kluwer, Boston, 1991.

[5-30] G. Biundo, M. Ansorge et B. Carnero, "Codeur de parole à gain réduit", brevet EP 02 015 920.8, déposé en juillet 2002.

[5-31] E. Harborg, J. Knudsen, A. Fuldseth et F. Johansen, "A real-time wideband CELP coder for a videophone application", dans *Proc. IEEE Int. conf. acoustic, speech, signal processing 1994, ICASSP'94,* Vol. 2, pp. 121-124, Adelaide, Australie, Avr. 1994.

[5-32] J. Paulus et J. Schnitzler, "16 kbit/s wideband speech coding based on unequal subbands", dans *Proc. IEEE Int. conf. acoustic, speech, signal processing 1996, ICASSP'96,* Vol. 1, pp. 255-258, Atlanta, USA, Mai 1996.

[5-33] G. Kubin, "Nonlinear processing of speech", Chapter 16, dans *Speech coding and synthesis*, pp. 557-610, édité par W. Kleijn et K. Paliwal, Elsevier, Amsterdam, 1995.

[5-34] C. McElroy, B. Murray et A. Fagan, "On improving wideband CELP speech coders", dans *Proceedings signal processing VII: Theories and applications, EUSIPCO'94,* Vol. 2, pp. 912-915, Edinburgh, Angleterre, Août 1994.

[5-35] A. Black, I. Atkinson, A. Kondoz et B. Evans, "High quality 14.1 kb/s wideband speech coder", dans *4th European conference on speech communication and technology, EUROSPEECH'95,* pp. 45-48, Madrid, Espagne, Sept. 1995.

Chapitre 6
Description fonctionnelle du codeur P-MRWB-ACELP

6.1 Introduction

Ce chapitre décrit les différentes fonctions algorithmiques du codeur P-MRWB-ACELP (Proprietary Multi-Rate Wide-Band ACELP) développé dans le cadre de ce travail de thèse. Ce codeur a été implanté en code ANSI C, avec une arithmétique en virgule flottante et double précision (64 bits). Les Sections 6.2 et 6.3 décrivent respectivement les fonctions de l'encodeur et du décodeur. La complexité algorithmique de chacune des fonctions est donnée pour une trame de 20 ms. Elle correspond à la complexité la plus élevée possible (pire des cas) pour le mode le plus complexe (généralement le mode C). Elle est indiquée en termes de multiplications (MU), additions (AD), soustractions (SO), divisions (DI), inversions de signe (IV), fonctions cosinus (CO) et inverses de la fonction cosinus (ACO), fonctions "logarithme en base 10" (L10), fonctions "puissance de 10" (P10), fonctions "racine carrée" (SR) et incrémentation de pointeur (I). Le nombre de valeurs à stocker en mémoire (MEM) pour le fonctionnement des différents blocs du codeur est également indiqué. La complexité algorithmique de chacune des fonctions de l'encodeur et du décodeur est regroupée dans le Tableau 6-4 à la Section 6.4. La Section 6.4 présente la complexité de l'algorithme complet. La Section 6.5 donne les conclusions de ce chapitre.

6.2 Encodeur

Cette section décrit les différentes fonctions de l'encodeur. Celles-ci sont représentées sous forme de blocs à la Figure 6.1. Les vecteurs présentant l'indice "16", sont relatifs au traitement des LSP d'ordre 16. Au cours de cette section, l'expression "le bloc *n*" définit le bloc *n* à la Figure 6.1.

6.2.1 Pré-traitement

En entrée du codeur, deux fonctions de pré-traitement combinées sont appliquées au signal de parole à traiter $s(n)$. La première est un filtrage décrit par $H_{HP}(z) = H_{HP1}(z) \cdot H_{HP2}(z)$. Le filtre $H_{HP1}(z)$ sert de protection contre les composantes en très basse fréquence du signal $s(n)$. Il a une fréquence de coupure à 10 Hz, une atténuation de –15 dB à 50 Hz et une atténuation de -0.027 dB à 100 Hz. Il est donné par :

$$H_{HP1}(z) = \frac{1 - z^{-1}}{1 - 0.9961 \cdot z^{-1}}. \tag{6.1}$$

$H_{HP2}(z)$ pré-accentue les hautes fréquences du spectre de $s(n)$. Il est défini par l'équation (2.21), où $\mu = 0.3$. La seconde fonction divise par 2 l'amplitude du signal sortant de $H_{HP}(z)$. Elle évite de possibles débordements de la dynamique du signal, si le codeur est implanté en virgule fixe.

Le pré-traitement est effectué dans le bloc *1*. Sa complexité est de 1'280 MU, 640 AD, 1'600 SO, 639 I et 8 MEM.

6.2.2 Analyse par prédiction linéaire et quantification

Les LPC sont extraits selon la méthode décrite au point 5.6.3.1 et à la Sous-section 2.3.3. La fenêtre donnée par l'équation (5.1) est utilisée pour limiter le signal pré-traité $s_{PT}(n)$. Les coefficients d'auto-corrélation r_i, $(i = 0,...,16)$, du signal fenêtré $s_w(n)$ sont calculés selon l'équation (2.13). Pour éviter des problèmes arithmétiques, la valeur 1.0 est ajoutée à r_0. Puis, la largeur de bande est expansée de 60 Hz en multipliant les coefficients d'auto-corrélation $\{r_i\}$ par les valeurs correspondantes de la fenêtre $w_{ELB}(i)$ donnée par (cf. Sous-section 3.3.1) :

$$w_{ELB}(i) = \exp\left[-\frac{1}{2}\left(\frac{2\pi f_0 i}{F_s} \right)^2 \right], \quad i = 0,...,16, \tag{6.2}$$

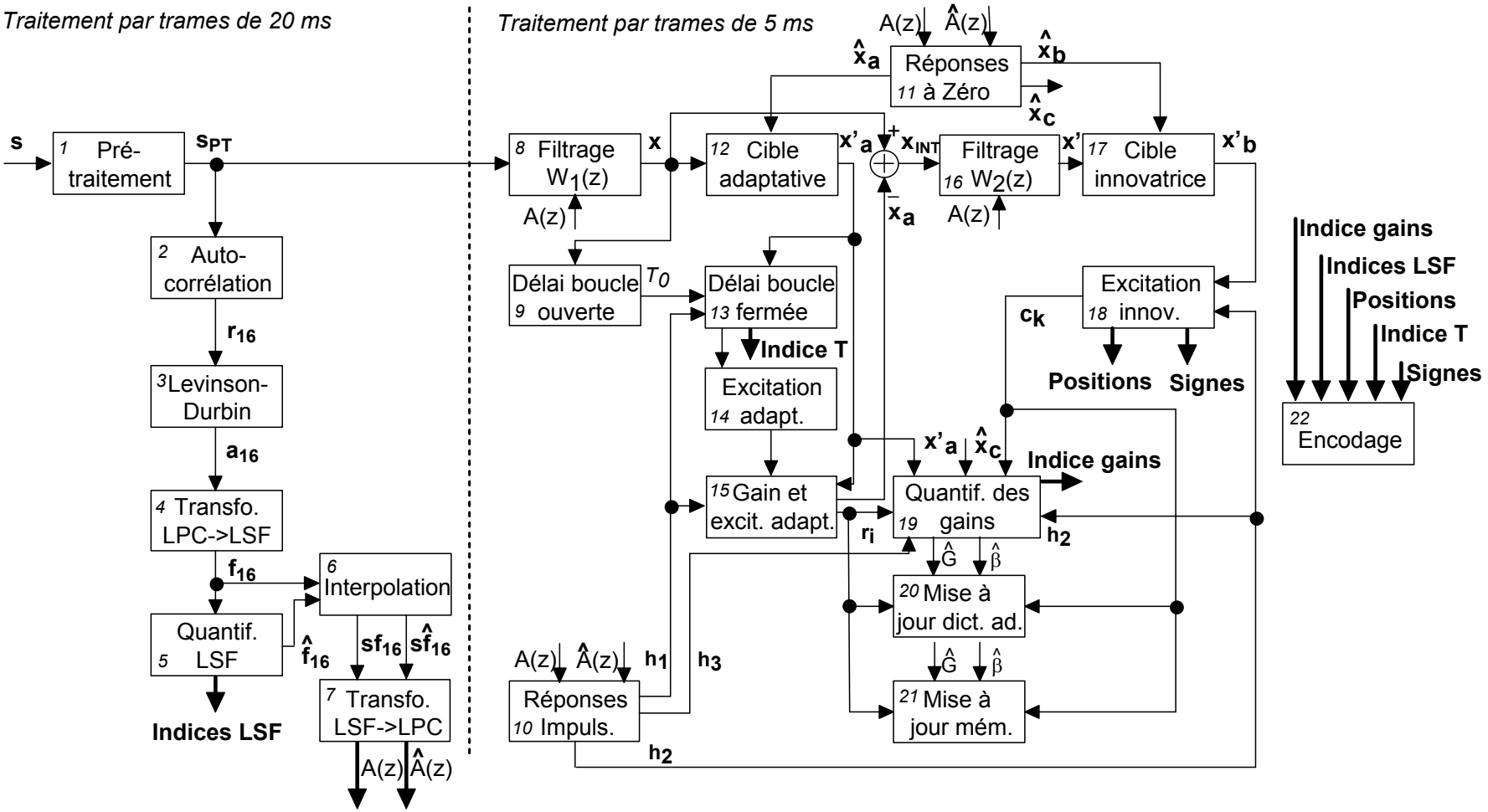

Figure 6.1 : Diagramme représentant les blocs fonctionels de l'encodeur simplifié.

où $f_0 = 60$ Hz. $F_s = 16000$ Hz est la fréquence d'échantillonnage du signal d'entrée. Finalement, le coefficient r_0 est multiplié par le facteur de correction 1.0001, ce qui correspond à l'ajout d'un bruit de fond à –40 dB. L'auto-corrélation et l'expansion de la largeur de bande sont effectuées dans le bloc *2*.

L'algorithme de Levinson-Durbin est utilisé pour extraire les coefficients LPC $a_{16}(i)$, $i = 0,...,16$, selon les équations (2.15) à (2.19) (cf. point 2.3.3.1). Cette opération est effectuée dans le bloc *3*. Les coefficients LPC sont transformés en paires de lignes spectrales (LSP) $\{\omega_i\}$, par la méthode de Kabal (cf. Sous-section 3.3.8 et Annexe A). La grille a été déterminée statistiquement : elle vaut $\Delta = 0.0077$. Ainsi, l'intervalle compris entre cos(0) et cos(π) est représenté par 260 points. En outre 4 bissections sont utilisées. En augmentant Δ, la complexité algorithmique se réduit mais le risque que deux LSP consécutifs se trouvent entre deux points de la grille augmente.

La distance minimale entre 2 LSP consécutifs est contrôlée en deux temps. Si nécessaire, les modifications suivantes sont effectuées :

$$\begin{aligned}
&\text{si } \omega_o < 0.005, \quad \text{alors } \omega_o = 0.005; \\
&\text{si } \omega_{i+1} - \omega_i - dMin < 0.0, \quad i = 1,...,15, \quad dMin = 0.0195, \\
&\text{alors } \omega_i = (\omega_i + \omega_{i+1} - dMin)/2.0, \quad \text{et} \\
&\omega_{i+1} = (\omega_i + \omega_{i+1} + dMin)/2.0; \\
&\text{si } \omega_{15} > \pi, \quad \text{alors } \omega_{15} = \pi.
\end{aligned} \tag{6.3}$$

Si l'une des conditions ci-dessus est remplie, un second contrôle est réalisé :

$$\begin{aligned}
&\text{si } \omega_o < 0.005, \quad \text{alors } \omega_o = 0.005; \\
&\text{si } \omega_{i+1} - \omega_i - dMin < 0.0, \quad i = 1,...,15, \\
&\text{alors } \omega_{i+1} = \omega_i + dMin; \\
&\text{si } \omega_{15} > \pi, \quad \text{alors } \omega_{15} = \pi.
\end{aligned} \tag{6.4}$$

Les LSP sont alors transformés dans le domaine des fréquences $\{f_i\}$:

$$f_i = \frac{F_s \cdot \omega_i}{2\pi}, \quad i = 0,...,15. \tag{6.5}$$

La transformation des paramètres LPC en $\{f_i\}$ est effectuée dans le bloc *4*.

Les paramètres $\{f_i\}$ sont quantifiés selon la méthode illustrée à la Figure 5.6, en utilisant le schéma #27, 42-[$(5,11)_{7,7;}$ $(2,4,3,3,4)_{5,6,6,6,5}$] MA-1 décrit à la Section 5.6. Soit m l'indice de la trame de signal traitée et soit $\tilde{\mathbf{z}}(m)$, le vecteur de LSP $\mathbf{f}(m)$ auquel le vecteur de LSP moyen $\overline{\mathbf{f}}$ a été retiré, alors le vecteur résiduel $\mathbf{r}(m)$ est donné par :

$$r_i(m) = \tilde{z}_i(m) - p_i(m), \quad i = 0,...,15, \tag{6.6}$$

où **p** est le vecteur de prédiction MA-1 donné par :

$$p_i(m) = \alpha_i \hat{r}_i(m-1), \quad i = 0,...,15. \tag{6.7}$$

Les $\{\alpha_i\}$ sont les coefficients de prédiction MA-1 et $\hat{\mathbf{r}}(m-1)$ est le vecteur résiduel quantifié de la trame passée. $\mathbf{r}(m)$ est quantifié en utilisant le schéma #27 et la mesure de distance Euclidienne $d(\mathbf{r}, \hat{\mathbf{r}})$, donnée par :

$$d(\mathbf{r}(m), \hat{\mathbf{r}}(m)) = \sum_{l=1}^{L} \left(r_l(m) - \hat{r}_l(m)\right)^2, \tag{6.8}$$

où L est la longueur du sous-vecteur traité. Les LSP quantifiés $\{\hat{f}_i\}$ sont transformés dans le domaine des fréquences angulaires $\{\hat{\omega}_i\}$ où l'ordre et la distance minimale entre 2 LSP consécutifs sont contrôlés selon les équations (6.3) et (6.4). La quantification des LSP est effectuée dans le bloc ***5***.

Pour chacune des sous-trames de 5 ms, un nouvel ensemble de LSP quantifié et non-quantifié est obtenu par interpolation selon l'équation (3.12). Cette opération est effectuée dans le bloc *6*. Les ensembles de LSP interpolés sont alors transformés en LPC selon la méthode de Kabal (cf. Sous-section 3.3.9)[9]. Cette opération est effectuée dans le bloc *7*.

La complexité des opérations décrites dans cette sous-section est de 16'362 MU, 24'478 AD, 9'354 SO, 92 DI, 16 IV, 112 CO, 16 ACO, 14'177 I et 3'416 MEM.

6.2.3 Analyse du délai tonal en boucle ouverte

Le signal prétraité $s_{PT}(n)$ est passé dans le filtre de pondération $W_1(z) = \tilde{W}_1(z) \cdot W_{T1}(z)$ introduit au point 5.7.1.2. Les coefficients γ_{11}, γ_{12}, μ_{11} et μ_{12} des filtres de pondération formantique $\tilde{W}_1(z)$ et de pondération de la pente spectrale $W_{T1}(z)$, décrits par les équations (5.19) et (5.22), valent respectivement 1.0, 0.2, 0.0 et 0.3. Le filtrage par $W_1(z)$ est effectué dans le bloc ***8***. Le signal de sortie du filtre $W_1(z)$, défini par $x(n)$, est utilisé pour la recherche du délai tonal T_0 en boucle ouverte. Celle-ci est réalisée toutes les 5 ms selon la méthode E décrite à la Sous-section 5.8.1. Elle est effectuée en dix étapes : les maxima $M\tilde{R}_i$ $(i = 0,1,...,9)$ de la corrélation non normalisée $\tilde{R}_g$ donnée par :

[9] Les multiplications et divisions par une puissance de 2 sont comptabilisées pour le calcul de la complexité.

$$\tilde{R}_g = \sum_{n=0}^{79} x(n)x(n-g), \tag{6.9}$$

entre le signal $x(n)$ et ce même signal décalé, sont calculés dans les intervalles de valeurs g_i suivants : $g_0 = [30,..,55]$, $g_1 = [56,..,80]$, $g_2 = [81,..,105]$, $g_3 = [106,..,130]$, $g_4 = [131,..,155]$, $g_5 = [156,..,180]$, $g_6 = [181,..,205]$, $g_7 = [206,..,230]$, $g_8 = [231,..,255]$ et $g_9 = [256,\ldots,\text{MMax}]$. MMax vaut respectivement 293 et 283 pour les modes de fonctionnement I et J (cf. Sous-section 5.8.1). Les corrélations maximales $M\tilde{R}_i$, $i = 0,...,9$, sont ensuite divisées par la valeur de $D(t_i)$:

$$D(t_i) = \sqrt{\sum_{n=0}^{79} x^2(n-t_i)}, \tag{6.10}$$

où t_i est le délai correspondant à $M\tilde{R}_i$, $t_i \in g_i$. Cette division normalise $M\tilde{R}_i$.

Le délai tonal en boucle ouverte T_0 est extrait en comparant les $R_i = M\tilde{R}_i / D(t_i)$, $i = 0,...,9$, selon l'équation (5.24) où $W_{bo} = 1.0$:

$$\begin{aligned} &R_{bo} = R_9, \quad T_0 = t_9; \\ &\text{Pour} \quad i = 8 \ \text{à} \ i = 0, \ \text{par pas de } -1, \\ &\text{si} \quad R_i \geq R_{bo}, \quad \text{alors} \quad R_{bo} = R_i \ \text{et} \ T_0 = t_i. \end{aligned} \tag{6.11}$$

Enfin, les bornes de recherche en boucle fermée sont fixées ainsi :

$$\begin{aligned} &T_{\min} = T_0 - 6; \\ &\text{si } T_{\min} < 29, \quad T_{\min} = 29; \\ &T_{\max} = T_{\min} + 12; \\ &\text{si} \quad T_{\max} > \text{MMax}, \quad T_{\max} = \text{MMax} \quad \text{et} \quad T_{\min} = \text{MMax} - 12. \end{aligned} \tag{6.12}$$

La recherche de T_0 est effectuée dans le bloc *9*. La complexité des différentes opérations décrites dans cette sous-section est de 98'695 MU, 92'815 AD, 98'256 SO, 40 DI, 40 SR, 102'348 I et 48 MEM.

6.2.4 Calcul des réponses impulsionnelles et des réponses à zéro

Les réponses impulsionnelles, $h_1(n)$, $h_2(n)$ et $h_3(n)$ des filtres de synthèse pondérée $(1/\hat{A}(z)) \cdot W_1(z)$, $(1/\hat{A}(z)) \cdot W_1(z) \cdot W_2(z)$ et du filtre $W_2(z)$ (représenté entre $x_{\text{int}}(n)$ et $x'(n)$ à la Figure 5.12) sont calculées pour chaque sous-trame de 5 ms. $W_2(z) = \tilde{W}_2(z) \cdot W_{T2}(z)$, où $\tilde{W}_2(z)$ et $W_{T2}(z)$ sont décrits par les équations (5.20) et (5.22) et où les coefficients γ_{21}, γ_{22}, μ_{21} et μ_{22} valant respectivement 0.2, 0.9, 0.3 et -0.2. Les réponses à une excitation

d'entrée nulle (réponses à zéro) de ces filtres sont également calculées. Elles sont respectivement dénommées $\hat{x}_a(n)$, $\hat{x}_b(n)$ et $\hat{x}_c(n)$. Elles tiennent compte de l'état des mémoires de ces filtres, suite au traitement de la sous-trame passée.

Les réponses impulsionelles et à zéro servent à l'extraction des excitations (cf. Sous-section 3.4.1 et Section 3.5) et à la quantification des gains, selon le schéma illustré à la Figure 5.12 (switch en position a). Leur calcul est effectué dans les blocs *10* et *11*. La complexité des opérations décrites est de 82'184 MU, 32'020 AD, 94'400 SO, 117'060 I et 37 MEM.

6.2.5 Résidu et cible pour l'extraction de l'excitation adaptative

Le résidu de prédiction à court-terme $r(n)$ du signal traité est donné par :

$$r(n) = s_{PT}(n) + \sum_{i=1}^{16} \hat{a}_p(i) \cdot s_{PT}(n-i). \qquad (6.13)$$

Il remplace l'excitation adaptative passée du dictionnaire adaptatif, pour la recherche des délais tonals τ inférieurs à la durée d'une sous-trame. L'usage du résidu réduit la complexité algorithmique de cette recherche.

Le signal cible $x'_a(n)$ pour l'extraction de l'excitation adaptative est calculé selon l'équation (3.30) :

$$x'_a(n) = x(n) - \hat{x}_a(n).$$

$x(n)$ est le signal prétraité $s_{PT}(n)$, passé à travers le premier filtre de pondération $W_1(z)$, représenté à la Figure 5.12.

Le résidu et la cible sont calculés toutes les 5 ms. Le calcul de la cible $x'_a(n)$ est effectué dans le bloc *12*. La complexité des opérations décrites dans cette sous-section est de 5'120 MU, 5'120 AD, 5'120 S0 et 9'912 I.

6.2.6 Recherche du délai tonal

Le délai tonal est recherché toutes les 5 ms. La recherche diffère selon le mode de fonctionnement du codeur. Pour le mode A, décrit à la Section 5.10 (cf. cas I de la Sous-section 5.8.1), le délai tonal en boucle fermée T est recherché avec une précision de 1/3, 1/2 ou 1 échantillon sur les intervalles de valeurs entières comprises entre 29 et 120 (1er intervalle), 121 et 180 (2ème intervalle), et respectivement 181 et 293 (3ème intervalle). Pour les modes B et C (cf. cas J de la Sous-section 5.8.1), ce délai tonal est recherché avec une précision de 1/4 sur l'intervalle compris entre 29 et 283.

Le délai tonal T est recherché en boucle fermée pour des valeurs entières, T_{INT}, comprises entre $T_{\min}$ et $T_{\max}$ (cf. équation (6.12)), en recherchant la valeur entière τ qui maximise la fonction de correspondance $\Psi(\tau)$ donnée par l'équation (3.41). Comme ici nous ne désirons pas obtenir de gains négatifs, nous ne prenons pas la valeur absolue du numérateur de $\Psi(\tau)$:

$$\Psi(\tau) = \frac{\sum_{n=0}^{N-1} x'_a(n) y_\tau(n)}{\sqrt{\sum_{n=0}^{N-1} \left[y_\tau(n)\right]^2}}, \quad \tau = T_{\min} - 4, \ldots, T_{\max} + 4. \tag{6.14}$$

Les 4 valeurs qui précèdent et suivent $T_{\min}$ et $T_{\max}$ sont nécessaires au calcul de l'interpolation décrite ci-après. $x'_a(n)$ est le signal cible discuté à la Sous-section 6.2.5. $y_\tau(n)$, en $\tau = T$, est le signal d'excitation testé, $u(-T_{INT})$, convolué à la réponse impulsionnelle de $(1/\hat{A}(z)) \cdot W_1(z)$: $h_1(n)$ (cf. Sous-section 6.2.4). Cette convolution simule le filtrage. Elle est calculée pour $T_{INT} = T_{\min}$. Puis, pour les autres valeurs de T_{INT} testées, elle est mise à jour en utilisant la récursion suivante :

$$y_\tau(n) = y_{\tau-1}(n-1) + u(-\tau) \cdot h_1(n-1). \tag{6.15}$$

Pour les valeurs de τ comprises entre 80 et MMax (cf. Sous-section 6.2.3), $u(-\tau)$ est le signal d'excitation passé contenu dans le dictionnaire adaptatif. Pour les valeurs de τ inférieures à 80, ce signal est le résidu de prédiction donné par l'équation (6.13).

Une fois le délai tonal entier T_{INT} déterminé, la fonction de correspondance $\Psi(\tau)$ est interpolée autour de $\tau = T_{INT}$, afin de calculer la fraction de T. Celle-ci correspond à la valeur interpolée maximale. Pour le mode A, les fractions –2/3, –1/3, 0.0, 1/3 et 2/3, ou –1/2, 0.0 et 1/2, sont testées en fonction de l'intervalle dans lequel se trouve T_{INT}. Si $T_{INT} > 180$, l'interpolation n'est pas réalisée ($T = T_{INT}$). Pour les modes B et C, les fractions –3/4, –2/4, –1/4, 0.0, 1/4 , 2/4 et 3/4 sont testées.

Pour le mode A, le filtre RIF (à réponse impulsionnelle finie) dénommé b_{12} est utilisé pour réaliser l'interpolation par 3 dans le premier intervalle de délais ([29,120]). b_{12} est obtenu par une fonction $\sin(s)/s$ à laquelle est appliquée une fenêtre de Hamming. b_{12} est symétrique. En posant son maximum en position 0, il est tronqué en positions +11 et –11, dénotés par +/– 11. De plus, il est complété par des zéros en positions ≤ -12 et ≥ 12, ce qui est dénoté par $b_{12}(|i| \geq 12) = 0$. Dans le second intervalle de délais ([121,180]), le filtre RIF dénommé b_8 est utilisé pour réaliser l'interpolation par 2. Ce filtre est obtenu par une fonction $\sin(s)/s$ à laquelle est appliquée une fenêtre de Hamming. Il est tronqué en +/– 7 et $b_8(|i| \geq 8) = 0$. Pour les

modes B et C, le filtre RIF dénommé b_{16} est utilisé pour réaliser l'interpolation par 4. Il est obtenu par une fonction $\sin(s)/s$ à laquelle est appliquée une fenêtre de Hamming. Il est tronqué en +/- 15 et $b_{16}(|i| \geq 16) = 0$. La fréquence de coupure (-3 dB) des filtres b_{12}, b_9 et b_{16} se situe à 6125 Hz dans le domaine sur-échantillonné.

Soit I le facteur d'interpolation ($I = 2,3,4$), soit t/I la fraction à rechercher; si la valeur t est négative, alors les modifications suivantes sont apportées :

$$\begin{aligned} \frac{t}{I} &= \frac{t}{I} + 1.0; \\ T_{INT} &= T_{INT} - 1. \end{aligned} \tag{6.16}$$

La fonction de correspondance $\Psi(\tau)$ interpolée en t/I, vaut :

$$\Psi(\tau)_t = \sum_{i=0}^{3} \Psi(\tau - i) b_I(t + I \cdot i) + \sum_{i=0}^{3} \Psi(\tau + 1 + i) b_I(I - t + I \cdot i), \tag{6.17}$$

où b_I est respectivement b_{12}, b_8 ou b_{16}.

Dans le cas où le délai tonal fractionnaire obtenu, T, est inférieur à 141, toutes les opérations décrites ci-dessus sont répétées pour $T_0 = 2 \cdot T_{INT}$, en posant $T_{\min} = 2 \cdot T_{INT} - 6$ et $T_{\max} = 2 \cdot T_{INT} + 6$, et un second délai tonal fractionnaire T' est extrait. Le délai retenu, T ou T', correspond à celui dont la valeur de la fonction de correspondance $\Psi(\tau)_t$ est maximale.

Le calcul du délai tonal en boucle fermée est effectué dans le bloc *13*.

6.2.7 Calcul de l'excitation adaptative

Soit la valeur entière T_{INT} du délai tonal T, et t/I sa composante fractionnaire ($I = 2,3,4$). Une fois T déterminé, le vecteur d'excitation adaptatif $r_i(n)$ est calculé en interpolant l'excitation contenue dans le dictionnaire adaptatif $u(n)$:

$$\begin{aligned} r_i(n) = & \sum_{k=0}^{10} u(n - T_{INT} + k) b_I(t + I \cdot k) \\ & + \sum_{k=0}^{10} u(n - T_{INT} + 1 + k) b_I(I - t + I \cdot k). \end{aligned} \tag{6.18}$$

où $n = 0,...,79$, b_I est l'un des filtres RIF respectivement dénommés b_{60}, b_{40} ou b_{80}. Ces filtres sont obtenus par une fonction $\sin(s)/s$ à laquelle est appliquée une fenêtre de Hamming. Ils sont tronqués en +/– 59, +/– 39 et

respectivement +/– 79. De plus, $b_{60}(|i| \geq 60) = 0$, $b_{40}(|i| \geq 40) = 0$ et respectivement $b_{80}(|i| \geq 80) = 0$. Ce sont des filtres tiers, demi et quart d'onde. Leur fréquence de coupure se situe à 7830 Hz dans le domaine sur-échantillonné. b_{60} et b_{40} sont utilisés pour le mode A, respectivement pour réaliser une interpolation par 3 et 2 et obtenir une résolution correspondant à 1/3 ou 1/2 échantillon. b_{80} est utilisé pour les modes B et C, pour une interpolation par 4 et une résolution correspondant à 1/4 d'échantillon. Dans un souci de complexité, la longueur des filtres b_I pourrait être réduite par un facteur 2. Dans un tel cas, la perte en qualité de signal reconstruit est minime.

Pour les valeurs de délai tonal T inférieures à 80, au fur et à mesure que $r_i(n)$ est calculé, l'égalité suivante est posée : $u(n) = r_i(n)$.

Le calcul de l'excitation adaptative est effectué dans le bloc *14*.

6.2.8 Calcul du gain adaptatif

Une fois le vecteur d'excitation adaptatif $r_i(n)$ calculé, son gain G est extrait comme suit :

$$G = \frac{\sum_{n=0}^{79} x'_a(n)y(n)}{\sum_{n=0}^{79} y(n)y(n)}, \tag{6.19}$$

où $x'_a(n)$ est le signal cible (cf. Sous-section 6.2.1), et $y(n)$ est le vecteur d'excitation adaptatif $r_i(n)$ convolué à $h_1(n)$:

$$y(n) = \sum_{k=0}^{n} r_i(k)h_1(n-k), \quad n = 0,...,79. \tag{6.20}$$

G est limité par les valeurs inférieure et supérieure de 0.0 et 1.2.

Finalement, la contribution adaptative $x_a(n)$ est obtenue en filtrant le vecteur $r_i(n)$, multiplié par son gain G, par $(1/\hat{A}(z)) \cdot W_1(z)$.

Le calcul du gain adaptatif G et de la contribution adaptative $x_a(n)$ est effectué dans le bloc *15*. La complexité des opérations décrites aux Sous-sections 6.2.6, 6.2.7 et 6.2.8, est dans le pire des cas de 108'780 MU, 104'448 AD, 73'248 SO, 172 DI, 6'644 IV, 168 SR, 110'000 I et 56 MEM. Ces valeurs tiennent compte de la possibilité d'une double recherche telle que décrite à la Sous-section 6.2.6.

6.2.9 Procédure de recherche dans le dictionnaire innovateur

Le dictionnaire d'excitations innovatrices, utilisé ici, contient des codes algébriques de type ISSP (cf. Sous-section 3.5.1). L'extraction de l'excitation innovatrice peut se faire selon les trois modes présentés à la Section 5.9 et associés aux modes de fonctionnement A, B et C du codeur (14.3, 18.5 et 21.5 kbits/s), décrits à la Section 5.10.

Pour les modes A, B et C, le dictionnaire d'excitation innovatrice de type algébrique est composé de 5 pistes contenant 2, 3 et respectivement 4 impulsions comme le décrivent les Tableaux 6-1 à 6-3. Un total de respectivement 45, 65 et 80 bits sont nécessaires pour encoder l'excitation innovatrice d'une sous-trame de 5 ms. Les 2, 3 et 4 impulsions d'une piste sont respectivement encodées avec 9, 13 et 16 bits, soit 4 bits pour la position de chacune des impulsions et 1 bit pour les signes des impulsions d'une piste pour les modes A et B. En effet, pour les 2 et respectivement 3 impulsions d'une même piste, seul le signe de la première impulsion est encodé. Les autres signes sont dérivés des positions relatives des impulsions. Pour le mode C, la position relative des impulsions d'une piste suffit à encoder leur signe (cf. Sous-section 3.5.1).

Le vecteur d'excitation innovatrice $c(n)$ est construit en prenant un vecteur de dimension 80, dont toutes les valeurs sont nulles, et en y insérant 10, 15 ou respectivement 20 impulsions aux positions m_i déterminées. Ces impulsions sont ensuite multipliées par leur signe respectif. Par exemple, pour le mode A, $c(n)$ est donné par :

$$c(n) = \sum_{i=0}^{9} s_i \delta(n - m_i) \quad n = 0,...,79. \tag{6.21}$$

Les points 6.2.9.1 et 6.2.9.2 décrivent le calcul du signal cible et l'extraction du vecteur d'excitation innovatrice de type algébrique.

Piste	Impulsions	Signes	Positions (q ∈ [0,1,2,…,15])	
1	i_0, i_5	s_0, s_5 : ±1	m_0, m_5 :	0+5q
2	i_1, i_6	s_1, s_6 : ±1	m_1, m_6 :	1+5q
3	i_2, i_7	s_2, s_7 : ±1	m_2, m_7 :	2+5q
4	i_3, i_8	s_3, s_8 : ±1	m_3, m_8 :	3+5q
5	i_4, i_9	s_4, s_9 : ±1	m_4, m_9 :	4+5q

Tableau 6-1 : Positions et signes potentiels des impulsions individuelles dans le dictionnaire algébrique pour le mode A (14.3 kbits/s).

Piste	Impulsions	Signes	Positions (q ∈ [0,1,2,…,15])	
1	i_0, i_5, i_{10}	$s_0, s_5, s_{10} : \pm 1$	m_0, m_5, m_{10} :	0+5q
2	i_1, i_6, i_{11}	$s_1, s_6, s_{11} : \pm 1$	m_1, m_6, m_{11} :	1+5q
3	i_2, i_7, i_{12}	$s_2, s_7, s_{12} : \pm 1$	m_2, m_7, m_{12} :	2+5q
4	i_3, i_8, i_{13}	$s_3, s_8, s_{13} : \pm 1$	m_3, m_8, m_{13} :	3+5q
5	i_4, i_9, i_{14}	$s_4, s_9, s_{14} : \pm 1$	m_4, m_9, m_{14} :	4+5q

Tableau 6-2 : Positions et signes potentiels des impulsions individuelles dans le dictionnaire algébrique pour le mode B (18.5 kbits/s).

Piste	Impulsions	Signes	Positions (q ∈ [0,1,2,…,15])	
1	i_0, i_5, i_{10}, i_{15}	$s_0, s_5, s_{10}, s_{15} : \pm 1$	m_0, m_5, m_{10}, m_{15} :	0+5q
2	i_1, i_6, i_{11}, i_{16}	$s_1, s_6, s_{11}, s_{16} : \pm 1$	m_1, m_6, m_{11}, m_{16} :	1+5q
3	i_2, i_7, i_{12}, i_{17}	$s_2, s_7, s_{12}, s_{17} : \pm 1$	m_2, m_7, m_{12}, m_{17} :	2+5q
4	i_3, i_8, i_{13}, i_{18}	$s_3, s_8, s_{13}, s_{18} : \pm 1$	m_3, m_8, m_{13}, m_{18} :	3+5q
5	i_4, i_9, i_{14}, i_{19}	$s_4, s_9, s_{14}, s_{19} : \pm 1$	m_4, m_9, m_{14}, m_{19} :	4+5q

Tableau 6-3 : Positions et signes potentiels des impulsions individuelles dans le dictionnaire algébrique pour le mode C (21.5 kbits/s).

6.2.9.1 Calcul de la cible pour l'extraction de l'excitation innovatrice

La recherche dans le dictionnaire algébrique est réalisée en minimisant l'erreur quadratique moyenne entre le signal $x'(n)$ et la contribution de l'excitation innovatrice $x_b(n)$, illustrés à la Figure 5.12 (switch en position a).

Le signal cible, pour l'extraction de l'excitation innovatrice $x'_b(n)$, est calculé comme suit. La contribution adaptative $x_a(n)$ est soustraite du signal $x(n)$ (cf. Sous-section 6.2.5) :

$$x_{INT}(n) = x(n) - x_a(n), \quad n = 0,...,79. \tag{6.22}$$

Le signal intermédiaire obtenu, $x_{INT}(n)$, est filtré par $W_2(z)$. Cette opération est effectuée dans le bloc *16,* et illustrée à la Figure 5.12. La sortie du filtre $W_2(z)$ est dénommée $x'(n)$. Finalement, la réponse à zéro $\hat{x}_b(n)$ de la cascade de filtres $(1/\hat{A}(z)) \cdot W_1(z) \cdot W_2(z)$ est soustraite à $x'(n)$:

$$x'_b(n) = x'(n) - \hat{x}_b(n), \quad n = 0,...,79. \tag{6.23}$$

Le calcul du signal $x'_b(n)$ est effectué dans le bloc *17*. La complexité de ce calcul est de 11'200 MU, 5'440 AD, 11'520 SO et 15'344 I.

6.2.9.2 Calcul de l'excitation innovatrice

Soit l'erreur pondérée entre le signal original et la parole reconstruite, $e(n)$, donnée par l'équation (3.43) et illustrée à la Figure 5.12 :

$$e(n) = x'_b(n) - \beta \sum_{i=0}^{n} c_k(i) \cdot h_2(n-i), \quad n = 0,...,79, \tag{6.24}$$

où $c_k(n)$ est l'excitation innovatrice testée et où β est le gain qui lui correspond. Soit E_b l'erreur quadratique moyenne pondérée, donnée par l'équation (3.46). E_b est minimisée en maximisant le second terme de l'équation (3.50), donné ici par :

$$\mathbf{\Gamma}_k = \frac{(C_k)^2}{\xi_k} = \frac{\left(\mathbf{x}'^T_b \, \mathbf{H}_2 \mathbf{c}_k\right)^2}{\mathbf{c}_k^T \mathbf{H}_2^T \mathbf{H}_2 \mathbf{c}_k} = \frac{\left(\mathbf{\Psi}^T \mathbf{c}_k\right)}{\mathbf{c}_k^T \mathbf{\Theta} \mathbf{c}_k}. \tag{6.25}$$

C_k est la cross-corrélation entre $\mathbf{x}'_b$ et le mot de code filtré $\mathbf{H}_2\mathbf{c}_k$. $\mathbf{H}_2$ est la matrice de convolution de la réponse impulsionnelle $h_2(n)$ donnée par l'équation (3.53). ξ_k est l'énergie du mot de code $\mathbf{c}_k$ filtré, et :

$$\mathbf{\Psi}^T = \mathbf{x}'^T_b \, \mathbf{H}_2; \quad \mathbf{\Theta} = \mathbf{H}_2^T \mathbf{H}_2.$$

C_k et ξ_k sont données par les équations (3.59) et (3.60) :

$$C_k = \sum_{n=0}^{79} x'_b(n)\left[c_k(n) * h_2(n)\right] = \sum_{n=0}^{79} \psi(n) c_k(n);$$

$$\xi_k = \sum_{n=0}^{79} \left[c_k(n) * h_2(n)\right]^2 =$$

$$\sum_{n=0}^{79} c_k^2(n)\phi(n,n) + 2\sum_{n=0}^{78} \sum_{i=n+1}^{79} c_k(n) c_k(i) \phi(n,i),$$

où le symbole $*$ représente la convolution. Avant l'extraction du mot de code du dictionnaire algébrique, les valeurs de $\phi(n,i)$ et de $\psi(n)$ sont calculées selon les équations (3.55) et (3.57), avec $N = 80$. Puis, la procédure de recherche décrite à la Sous-section 3.5.1, est utilisée. C_k et ξ_k sont calculées comme suit (cf. équations (3.62) et (3.63)) :

$$C_k = \sum_{i=0}^{I} s_i \psi(m_i),$$

où I vaut 10, 15 et 20 pour les modes A, B et respectivement C. De plus, $s_i = \pm 1$ est l'amplitude de l'impulsion se trouvant à la position m_i;

$$\xi_k = \sum_{i=0}^{I} \phi(m_i, m_i) + 2\sum_{i=0}^{I-1} \sum_{l=i+1}^{I} s_i s_l \phi(m_i, m_l).$$

Le signal $\psi(n)$ est alors quantifié en posant l'amplitude, de l'impulsion en position n, égale au signe de $\psi(n)$: $s_n = sign\left[\psi(n)\right]$. La valeur absolue de $\psi(n)$ est calculée et dénotée $\left|\psi(n)\right|$. $\psi(n)$ devient alors $\psi(n) = sign\left[\psi(n)\right]\left|\psi(n)\right|$ et C_k est donnée par :

$$C_k = \sum_{i=0}^{I} \left|\psi(m_i)\right|.$$

Puis, $\phi(n,i)$ est remplacée par :

$$\phi'(n,i) = sign\left[\psi(n)\right] sign\left[\psi(i)\right] \phi\left(n,i\right), \quad \begin{cases} n = 0,...,79, \\ i = n+1,...,79. \end{cases}$$

Les éléments de la diagonale principale de $\mathbf{\Theta'}$ sont multipliés par 0.5 (cf. équation (3.67)) et ξ_k devient :

$$\xi_k = \sum_{i=0}^{I} \phi'(m_i, m_i) + \sum_{i=0}^{I-1} \sum_{l=i+1}^{I} \phi'(m_i, m_l).$$

La recherche est réalisée en boucles emboîtées. A chaque nouvelle boucle une impulsion supplémentaire est positionnée. Sa contribution est prise en compte à la boucle suivante, pour la minimisation de E_b.

Les opérations présentées ici sont effectuées dans le bloc *18*. La complexité des différentes opérations pour le calcul de l'excitation innovatrice est de 109'756 MU, 97'596 AD, 91'600 SO, 29'356 DI, 320 IV, 51'184 I et 13 MEM.

6.2.10 Quantification des gains

Le gain G du vecteur d'excitation adaptative $r_i(n)$ et le gain β du vecteur d'excitation innovatrice $c_k(n)$ sont quantifiés vectoriellement sur 7 bits. L'exploration du dictionnaire de gains consiste à minimiser l'erreur quadratique pondérée E_Q, entre le signal original et le signal reconstitué, comme suit :

$$
\begin{aligned}
E_Q &= \|\mathbf{e}\|^2 = \|\mathbf{x}' - \mathbf{x}_b\|^2 = \left\|\mathbf{x}_{INT} * \mathbf{h}_3 + \hat{\mathbf{x}}_c - \hat{\beta}\mathbf{z}_1 - \hat{\mathbf{x}}_b\right\|^2 \\
&= \left\|(\mathbf{x} - \mathbf{x}_a) * \mathbf{h}_3 + \hat{\mathbf{x}}_c - \hat{\beta}\mathbf{z}_1 - \hat{\mathbf{x}}_b\right\|^2 \\
&= \left\|(\mathbf{x}'_a - \hat{G}\mathbf{y}) * \mathbf{h}_3 + \hat{\mathbf{x}}_c - \hat{\beta}\mathbf{z}_1 - \hat{\mathbf{x}}_b\right\|^2 = \left\|\mathbf{x}_1 - \hat{G}\mathbf{y}_1 - \hat{\beta}\mathbf{z}_1\right\|^2 \\
&= \left|\mathbf{x}_1^T\mathbf{x}_1 + \hat{G}^2\mathbf{y}_1^T\mathbf{y}_1 + \hat{\beta}^2\mathbf{z}_1^T\mathbf{z}_1 - 2\hat{G}\mathbf{x}_1^T\mathbf{y}_1 - 2\hat{\beta}\mathbf{x}_1^T\mathbf{z} + 2\hat{G}\hat{\beta}\mathbf{y}_1^T\mathbf{z}_1\right|; \\
\mathbf{x}_1 &= \mathbf{x}'_a * \mathbf{h}_3 + \hat{\mathbf{x}}_c - \hat{\mathbf{x}}_b; \qquad \mathbf{y}_1 = \mathbf{y} * \mathbf{h}_3,
\end{aligned}
\tag{6.26}
$$

où le symbole $*$ représente la convolution. $z_1(n)$ est donné par :

$$
z_1(n) = \sum_{i=0}^{n} c_k(i) h_2(n-i), \quad n = 0, ..., 79. \tag{6.27}
$$

Les autres vecteurs sont ceux décrits précédemment au cours de ce chapitre. $\hat{G}$ et $\hat{\beta}$ sont les valeurs de G et β quantifiées, représentées par G_j et β_l à la Figure 5.12.

En réalité, ce n'est par le gain β qui est quantifié, mais le facteur de prédiction γ décrit à la Sous-section 3.5.2 et donné par l'équation (3.72) :

$$
\beta = \gamma\beta',
$$

où β' est calculé selon les équations (3.73) à (3.79), avec $N = 80$ et $c(n) = c_k(n)$. $\overline{E} = 30$ dB et les coefficients de prédiction b_i, $i = 1, 2, 3, 4$ valent respectivement 0.62, 0.50, 0.32 et 0.19. Dans le cas présent, l'algorithme LBG ne fonctionne pas bien pour implanter le dictionnaire de quantification. Pour cette raison, cet algorithme a été utilisé sur des sous-parties du plan formé par les valeurs de $[G, \gamma]$. En effet, ce plan a été divisé en 16 zones rectangulaires, contenant un nombre égal de points pour l'entraînement. Puis, pour chaque zone, l'algorithme LBG a été utilisé afin de calculer 8 valeurs du dictionnaire final.

Les opérations présentées ici sont effectuées dans le bloc *19*. La complexité de ces opérations est de 60'328 MU, 56'116 AD, 53'216 SO, 12 DI, 8 L10, 4 P10, 54'020 I et 256 MEM.

6.2.11 Mise à jour du dictionnaire adaptatif et des mémoires

Pour la mise à jour du dictionnaire adaptatif, les vecteurs d'excitations adaptative $r_i(n)$ et innovatrice $c_k(n)$, multipliés par leur gain quantifié respectif $\hat{G}$ et $\hat{\beta}$, sont additionnés :

$$u_1(n) = \hat{G} r_i(n) + \hat{\beta} c_k(n), \quad n = 0,...,79. \tag{6.28}$$

Le vecteur $u_1(n)$ est ensuite passé dans un filtre passe-bas $H_{PB}(z)$, qui correspond au filtre de correction totale, illustré à la Figure 5.12. $H_{PB}(z)$ est un filtre RIF d'ordre 20, implanté en utilisant une fenêtre de Hamming. Sa fréquence de coupure se situe à 5680 Hz. La sortie de ce filtre est d'abord insérée en $u(n)$, pour $n = 0,...,79$, puis toutes les valeurs du dictionnaire adaptatif sont décalées de 80 positions :

$$u(-n) = u(-n+80), \quad n = T_{\max},....,1. \tag{6.29}$$

La mise à jour du dictionnaire adaptatif est effectuée dans le bloc *20*.

Finalement, les états des mémoires des filtres illustrés à la Figure 5.12 sont mis à jour. Ceci permet de calculer les réponses à zéro et les vecteurs cibles pour le traitement de la sous-trame suivante, en simulant le fonctionnement du décodeur. Pour cela, le switch est positionné en b. Les vecteurs $\hat{G} \cdot r_i(n)$ et $\hat{\beta} \cdot c_k(n)$ ainsi que la cible intermédiaire $x_{INT}(n)$ sont passés dans les filtres de synthèse pondérée $(1/\hat{A}(z)) \cdot W_1(z)$, $F(z) \cdot (1/\hat{A}(z)) \cdot W_1(z) \cdot W_2(z)$ et dans le filtre $W_2(z)$ (cf. Sous-section 6.2.4). Le filtre $F(z)$ est donné par :

$$F(z) = \frac{1+\eta \cdot z^{-1}}{1+\eta}, \quad \begin{cases} \eta = 0.5 \cdot \min(\hat{G}, 1.0), & \text{si } \hat{G} > 0.8, \\ \eta = 0, & \text{ailleurs.} \end{cases} \tag{6.30}$$

La mise à jour des états des mémoires des filtres est effectuée dans le bloc *21*. La complexité des opérations présentées ici est de 52'324 MU, 23'844 AD, 66'840 SO, 4 DI, 4 IV, 39'236 I et 21 MEM.

6.2.12 Encodage des différents paramètres

Les paramètres à encoder et à transmettre sont : les indices des sous-dictionnaires de LSP une fois toutes les 20 ms, puis toutes les 5 ms : l'indice du dictionnaire de gains; les positions et amplitudes des impulsions; le délai tonal fractionnaire. A l'exception du délai tonal fractionnaire T, tous les autres paramètres sont des entiers. Ceux-ci sont simplement transformés en base deux, pour obtenir un code binaire. Par contre, la valeur du délai T doit être transformée en valeur entière, puis en code binaire.

L'encodage des paramètres est effectué dans le bloc *22*. Sa complexité est de 434 MU, 4 AD, 438 SO, et 430 DI, 411 I et 12 MEM.

6.3 Décodeur

Cette section décrit les différentes fonctions du décodeur. Celles-ci sont représentées sous forme de blocs à la Figure 6.2. Au cours de cette section, l'expression "le bloc *n*" définit le bloc *n* à la Figure 6.2.

6.3.1 Décodage des paramètres

Le code binaire reçu de l'encodeur est d'abord transformé en indices, qui sont les indices des sous-dictionnaires de LSP résiduels, reçus toutes les 20 ms, et toutes les 5 ms l'indice correspondant au délai tonal fractionnaire, les indices définissant les positions et les amplitudes des impulsions de l'excitation innovatrice ainsi que l'indice du dictionnaire des gains.

Le bloc *1* effectue la transformation du code binaire en indices, dont la complexité est de 434 MU, 446 AD, 12 SO, 12 DI, 411 I et 28 MEM.

6.3.2 Traitement des paramètres LSP

Les indices des sous-dictionnaires de LSP résiduels servent à reconstruire le vecteur de LSP de la 4ème sous-trame de 5 ms, d'une trame de signal de 20 ms. Soit m l'indice de la trame traitée, le vecteur de LSP quantifiés dans le domaine des fréquences, auquel le vecteur de LPS moyens $\overline{\mathbf{f}}$ a été soustrait, $\tilde{z}(m)$, est reconstruit comme suit :

$$\tilde{z}_i(m) = \hat{r}_i(m) + p_i(m), \quad i = 0,...,15, \tag{6.31}$$

où $\hat{r}(m)$ est le vecteur de LSP résiduels reçu et reconstruit, et où $p(m)$ est le vecteur de prédiction MA-1 donné par l'équation (6.7). Les LSP $\{\hat{f}_i\}$, s'obtiennent en additionnant à $\tilde{z}(m)$ le vecteur $\overline{\mathbf{f}}$. Ils sont transformés dans le domaine des fréquences angulaires $\{\hat{\omega}_i\}$, puis un contrôle de l'ordre et de la distance minimale entre 2 LSP consécutifs est réalisé en deux temps, selon les équations (6.3) et (6.4). Ce contrôle est nécessaire en cas d'erreurs de transmission.

Les LSP sont interpolés selon l'équation (3.12). Un nouvel ensemble de LSP pour chaque sous-trame de 5 ms est obtenu. Les LSP interpolés sont transformés en coefficients LPC $\{a_i\}$, selon la méthode de Kabal (cf. Sous-section 3.3.9). Le filtre de synthèse obtenu, $1/\hat{A}(z)$, est utilisé pour la reconstruction du signal de parole $\hat{s}(n)$.

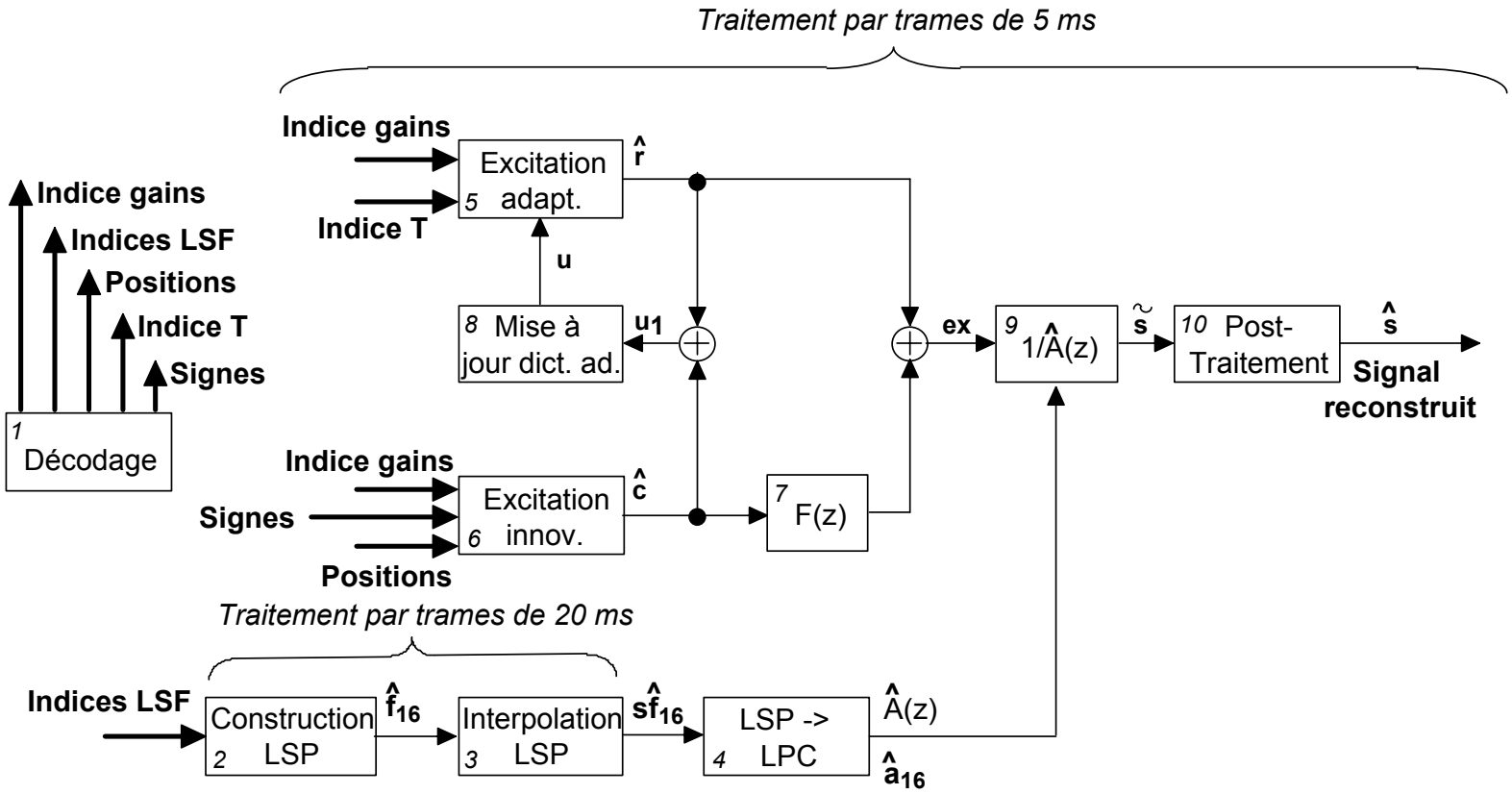

Figure 6.2 : Diagramme représentant les blocs fonctionnels du décodeur simplifié.

Les opérations présentées ici sont effectuées dans les blocs *2, 3* et *4*. Leur complexité est de 577 MU, 896 AD, 344 SO, 30 DI, 64 CO et 320 I.

Les opérations présentées ci-après sont réalisées 4 fois par trame.

6.3.3 Reconstruction de l'excitation adaptative

Le vecteur d'excitation adaptatif $r_i(n)$ est obtenu à partir de l'équation (6.18) selon le délai tonal fractionnaire T décodé. Puis, ce vecteur est multiplié par le gain adaptatif décodé $\hat{G}$:

$$\hat{r}(n) = \hat{G} \cdot r_i(n), \quad n = 0, ..., 79. \tag{6.32}$$

Ces opérations sont effectuées dans le bloc *5*. Leur complexité est de 13'440 MU, 1'312 AD, 324 SO, 6'408 IV et 6'712 I.

6.3.4 Reconstruction de l'excitation innovatrice et décodage de son gain

Les indices définissant les positions et les signes des impulsions de l'excitation innovatrice sont utilisés pour reconstruire cette excitation $c_k(n)$. Le gain de l'excitation innovatrice $\hat{\beta}$ est obtenu à partir de l'équation :

$$\hat{\beta} = \hat{\gamma}\beta',$$

où β' est calculé selon les équations (3.73) à (3.79), avec $N = 80$ et $c(n) = c_k(n)$. $\bar{E} = 30$ dB et les coefficients de prédiction b_i, $i = 1, 2, 3, 4$ valent respectivement 0.62, 0.50, 0.32 et 0.19. Finalement, l'excitation $c_k(n)$ est multipliée par son gain $\hat{\beta}$:

$$\hat{c}(n) = \hat{\beta} \cdot c_k(n), \quad n = 0, ..., 79. \tag{6.33}$$

Les opérations présentées ici sont effectuées dans le bloc *6*. Leur complexité est de 432 MU, 504 AD, 16 SO, 8 DI, 8 L10, 4 P10 et 736 I.

6.3.5 Construction de l'excitation totale et mise à jour du dictionnaire adaptatif

L'excitation totale $ex(n)$ est obtenue en passant $\hat{c}(n)$ dans le filtre $F(z)$ (bloc *7*), donné par l'équation (6.30) et en lui additionnant $\hat{r}(n)$.

Pour la mise à jour du dictionnaire adaptatif, les vecteurs $\hat{c}(n)$ et $\hat{r}(n)$ sont additionnés :

$$u_1(n) = \hat{G} \cdot r_i(n) + \hat{\beta} \cdot c_k(n), \quad n = 0,...,79.$$

Le vecteur $u_1(n)$ est alors passé dans un filtre passe-bas $H_{PB}(z)$ qui correspond au filtre de correction totale illustré à la Figure 5.12, et décrit à la Sous-section 6.2.11. La sortie de ce filtre est copiée dans le vecteur $u(n)$ pour $n = 0,...,79$, puis toutes les valeurs du dictionnaire adaptatif sont déplacées de 80 positions, selon l'équation (6.29). La mise à jour du dictionnaire adaptatif est effectuée dans le bloc *8*.

La complexité des opérations présentées ici est de 8'164 MU, 8'164 AD, 14'360 SO, 4 DI, 4 IV et 14'624 I.

6.3.6 Calcul de la parole reconstruite

Le signal de parole est synthétisé en passant l'excitation totale $ex(n)$ dans le filtre $1/\hat{A}(z)$ (bloc *9*). Le signal obtenu est dénoté $\tilde{s}(n)$. La complexité de cette opération est de 5124 MU, 4 AD, 9'920 SO et 9'596 I.

6.3.7 Post-traitement

Le signal de parole reconstruit $\hat{s}(n)$ est calculé en post-traitant le signal $\tilde{s}(n)$ (bloc *10*). Le post-traitement est effectué en trois étapes. La première étape consiste en un filtrage adaptatif qui combine un filtre formantique $H_{FPF}(z)$ et un filtre de pondération de la pente spectrale $H_{TPF}(z)$. Ce dernier compense la pente spectrale introduite par $H_{FPF}(z)$. Sans celui-ci, le signal de parole reconstruit donnerait un effet "étouffé" [4-1]. Le signal sortant de cette combinaison de filtres est dénommé $\tilde{s}'(n)$. $H_{FPF}(z)$ est donné par :

$$H_{FPF}(z) = \frac{\hat{A}(z/\alpha_{P1})}{\hat{A}(z/\alpha_{P2})}, \quad \alpha_{P1} = 0.5, \quad \alpha_{P2} = 0.6. \tag{6.34}$$

$H_{TPF}(z)$ est donné par :

$$H_{TPF}(z) = 1 - 0.4 \cdot z^{-1}. \tag{6.35}$$

La première étape se termine par une remise à l'échelle de l'énergie du signal $\tilde{s}'(n)$: soit $E1$, l'énergie du signal $\tilde{s}(n)$, et $E2$ l'énergie du signal $\tilde{s}'(n)$; soit $G_E = E2/E1$, la remise à l'échelle s'effectue ainsi :

$$\text{pour } n = 0:79$$
$$\left\{ \begin{array}{ll} \tilde{s}'(n) = (0.85 \cdot G_{EP} + 0.15 \cdot G_E) \cdot \tilde{s}'(n); & \\ G_{EP} = 0.85 \cdot G_{EP} + 0.15 \cdot G_E; & \end{array} \right\} \quad (6.36)$$

La dernière valeur de G_{EP} est conservée d'une sous-trame à l'autre.

La seconde étape inverse le processus de pré-traitement que subit le signal à l'entrée de l'encodeur (cf. Sous-section 6.2.1). Elle consiste en un filtrage par $1/\tilde{H}_{HP}(z)$ suivi par une multiplication par 2, où :

$$\tilde{H}_{HP}(z) = H_{HP1}(z) \cdot \tilde{H}_{HP2}(z) \quad (6.37)$$

$H_{HP1}(z)$ est décrit par l'équation (6.1), alors que $\tilde{H}_{HP2}(z)$ est donné par l'équation (2.21) où $\mu = 0.3 \cdot 0.9 = 0.27$.

Finalement, la troisième étape consiste en un filtrage passe-bas du signal, qui élimine toutes ses composantes de fréquences supérieures à 7 kHz. Le filtre utilisé est un filtre RII d'ordre 7. Ce filtre est un filtre elliptique (Cauer) calculé avec le programme Matlab.

La complexité des opérations présentées ici est de 18'688 MU, 8'960 AD, 25'920 SO, 4 DI, 4 SR, 22756 I et 18 MEM.

6.4 Complexité algorithmique totale

Le Tableau 6-4 regroupe la complexité algorithmique de chacun des blocs de l'encodeur et du décodeur ainsi que leur complexité algorithmique totale. La taille de la mémoire fixe de l'encodeur et du décodeur est y est également donnée.

En admettant que l'encodeur et le décodeur fonctionnent en même temps, la complexité par seconde est de 29'666'100 MU, 23'140'350 AD, 27'824'400 SO, 1'508'200 DI, 669'800 IV, 800 ACO, 8'800 CO, 800 L10, 400 P10, 28'477'000 I et 10'600 SR.

Comme l'illustre la Figure 6.1, au niveau de l'encodeur, seuls quelques blocs peuvent « fonctionner » en parallèle. Il s'agit dans un premier temps des blocs *8*, *10* et *11*, puis des blocs *9* et *12*, et enfin des blocs *20* et *21*. Au niveau du décodeur, les blocs *2*, *3* et *4* peuvent fonctionner en parallèle aux blocs *5* et *6*, lorsque la première sous-trame de signal est traitée. De plus les blocs *8* et *9* peuvent fonctionner en parallèle.

Fig.	# blocs	MU	AD	SO	DI	IV	ACO	CO	L10	P10	I	SR	MEM
6.1	***1***	1’280	640	1’600	0	0	0	0	0	0	693	0	8
	2,3,4,5,6,7	16’362	24’478	9’354	92	16	16	112	0	0	14’177	0	3’416
	8,9	98’695	92’815	98’256	40	0	0	0	0	0	102’348	40	48
	10,11	82’184	32’020	94’400	0	0	0	0	0	0	117’060	0	37
	12	5’120	5’120	5’120	0	0	0	0	0	0	9’912	0	0
	13,14,15	108’780	104’448	73’248	172	6’644	0	0	0	0	110’000	168	56
	16,17	11’200	5’440	11’520	0	0	0	0	0	0	15’344	0	0
	18	109’756	97’596	91’600	29’356	320	0	0	0	0	51’184	0	13
	19	60’328	56’116	53’216	12	0	0	0	8	4	54’020	0	256
	20,21	52’324	23’844	66’840	4	4	0	0	0	0	39’236	0	21
	22	434	4	438	430	0	0	0	0	0	411	0	12
6.2	***1***	434	446	12	12	0	0	0	0	0	411	0	28
	2,3,4	577	896	344	30	0	0	64	0	0	320	0	0
	5	13’440	1’312	324	0	6’408	0	0	0	0	6’712	0	0
	6	432	504	16	8	0	0	0	8	4	736	0	0
	7,8	8’164	8’164	14’360	4	4	0	0	0	0	14’624	0	0
	9	5’124	4	9’920	0	0	0	0	0	0	9’596	0	0
	10	18’688	8’960	25’920	4	0	0	0	0	0	22’756	4	18

Encodeur complet	546’463	442’521	505’592	30’106	6’984	16	112	8	4	514’385	208	3’867
Décodeur complet	46’859	20’286	50’896	58	6’412	0	64	8	4	55’155	4	46

Tableau 6-4 : Complexité algorithmique, par trame de 20 ms, des différents blocs du codeur et du décodeur.

Le Tableau 6-4 peut servir de base pour établir un profil détaillé de la charge temporelle en complexité de calcul. Cependant une étude fine et complète du graphe de dépendances temporelles, des différents blocs fonctionnels, pourrait être effectuée. Pour cela il faudrait indexer temporellement tous les signaux à l'entrée et à la sortie des blocs fonctionnels, ainsi qu'à l'intérieur de ces blocs. Une telle étude est un travail conséquent. Toutefois, elle permettrait d'évaluer les degrés de liberté pour exploiter au maximum les parallélismes, et choisir ainsi l'architecture à utiliser pour une implantation matérielle.

6.5 Conclusions

Les différentes fonctions algorithmiques du codeur P-MRWB-ACELP, encodeur et décodeur, ont été décrites au cours de ce chapitre. De plus une analyse détaillée de la complexité relative à ces fonctions à été donnée dans le pire des cas et pour le mode le plus complexe qui en général est le mode C.

L'algorithme de l'encodeur est beaucoup plus complexe que celui du décodeur. Par exemple, il nécessite environ 12, 22 et respectivement 10 fois plus de multiplications, additions et soustractions. De plus, les fonctions les plus complexes de l'encodeur sont l'extraction de l'excitation adaptative (boucle ouverte et fermée), l'extraction de l'excitation innovatrice, suivies du calcul des réponses impulsionnelles et à zéro, de la quantification des gains et de la mise à jour des mémoires. L'extraction de l'excitation adaptative est respectivement 2, 2.5, 3.5 et 4 fois plus complexe (multiplications) que les fonctions précitées et 13 fois plus complexe que l'analyse LPC !

Au niveau du décodeur, le bloc de post-traitement et la composition de l'excitation adaptative sont les fonctions les plus complexes. Elles sont suivies de la composition de l'excitation totale associée à la mise à jour du dictionnaire adaptatif.

Il est clair que les innovations décrites au Chapitre 5, qui concernent les filtres de pondération perceptuelle et l'extraction du délai tonal T ont une influence considérable sur la complexité algorithmique. Le problème de la complexité algorithmique est discuté au Chapitre 8.

6.6 Référence

[6-1] R. Salami, L. Hanzo, R. Steele, K. Wong et I. Wassell, "Speech coding", Chapter 3, dans *Mobile radio communications*, pp. 186-346, Raymond Steel Ed., Pentech Press Publishers, London, 1992.

Chapitre 7
Tests et résultats

7.1 Introduction

Ce chapitre décrit les tests et les résultats permettant de qualifier les modes A, B et C du codeur P-MRWB-ACELP (Proprietary Multi-Rate Wide-Band ACELP), fonctionnant à 14.3, 18.5 et respectivement 21.5 kbits/s. Les tests sont décrits à la Section 7.2, alors que les résultats sont donnés et discutés à la Section 7.3. La Section 7.4 propose de futurs travaux visant à améliorer davantage la qualité du codeur.

7.2 Tests

Pour qualifier les modes A, B et C du codeur P-MRWB-ACELP, des tests auditifs informels, dits en "double aveugle avec référence cachée" (cf. Section 2.6) ont été réalisés par 15 auditeurs.

Pour permettre une comparaison, ces tests ont également été effectués pour les différents modes du codeur G.722 de l'ITU [4-1], ainsi que pour trois modes du nouveau standard WB-AMR (Wide-Band Adaptive Multi-Rate) de l'ETSI [7-2]. Les modes du codeur G.722 correspondent à des débits de 64, 56 et 48 kbits/s. Ils sont introduits à la Section 4.1 et sont décrits par les sigles G.722 A, G.722 B et respectivement G.722 C. Le standard WB-AMR est introduit à la Sous-section 4.4.1. Il fonctionne à de multiples débits compris entre 6.6 et 23.85 kbits/s. Ici, les modes testés correspondent aux débits de 14.25, 18.25 et 23.05 kbits/s, puisque ces débits sont les plus proches des débits du codeur P-MRWB-ACELP.

Les tests auditifs ont été effectués pour chacun des 9 codeurs sus-mentionnés, avec 12 séquences de parole, T01 à T12, d'une durée de 9.4

secondes. Ces séquences proviennent essentiellement des bases de données BD-TEST et BD-IMT, décrites à la Sous-section 5.5.2.

Les séquences T01, T02 et T03 proviennent de la base de données BD-IMT. Elles ont été enregistrées en "amateur" et sont parfois bruitées ou contiennent des artefacts tels que des souffles. T01 et T02 sont des voix d'enfants : Danilo (11 ans) pour T01, et Pamela (8 ans) pour T02. Danilo et Pamela sont les enfants les plus âgés dont les voix sont contenues dans la base de données BD-IMT. Les autres enfants présentent des difficultés à prononcer une phrase complète. Les séquences T01 et T02 sont prononcées en français. Elles ont été retenues car elles racontent des fables et sont agréables à l'auditeur. Selon l'auteur de cette thèse, la qualité du signal reconstruit obtenu avec le codeur P-MRWB-ACELP est similaire quelle que soit la voix d'enfant testée, de la base de données BD-IMT. Le 3ème signal, T03, est prononcé en italien et correspond à la voix d'une collègue dont la fréquence fondamentale moyenne est très élevée.

La 4ème séquence, T04, est prononcée en anglais par un homme. Elle provient de la base de données TIMIT non-contenue dans BD-TRAIN (cf. Sous-section 5.5.2).

Les 8 autres séquences proviennent de la base de donnée BD-TEST, qui contient 30 signaux de parole. T05, T07, T09 et T011 sont des voix de femmes, alors que T06, T08, T10 et T12 sont des voix d'hommes. Un premier test auditif a été réalisé par l'auteur de cette thèse avec les modes A et C du codeur P-MRWB-ACELP. Les 30 signaux de la base de données BD-TEST répartis en deux classes, voix de femmes et voix d'hommes, ont été divisés en deux sous-classes : l'une contenant les signaux relativement bien encodés et l'autre les signaux relativement mal encodés. Chacune des 4 sous-classes obtenues contient 7 à 8 signaux. Dans chaque sous-classe de signaux bien encodés, le signal de meilleure qualité d'encodage a été retiré. De même, dans chaque sous-classe de signaux mal encodés, le signal de moins bonne qualité d'encodage a été retiré. Finalement, deux signaux ont été retenus au hasard dans chacune des 4 sous-classes.

Les signaux T01 à T12, encodés avec chacun des 9 codeurs sus-mentionnés, ont été testés par les 15 auditeurs. Pour chacun des 108 signaux encodés, l'auditeur entend trois séquences. La première est la séquence originale. Puis viennent à l'aveugle, soit la séquence encodée suivie de l'originale, soit la séquence originale suivie de la séquence encodée. L'auditeur compare la deuxième et la troisième séquence à la première et les note en fonction des dégradations constatées. Il utilise l'échelle des DMOS spécifiée à la Section 2.6. La réalisation des tests auditifs décrits ici requiert plus d'une heure par auditeur. Pour une statistique plus significative, il aurait

fallu un nombre d'auditeurs plus important, ainsi qu'un nombre plus élevé de séquences de test dans différentes langues.

Les résultats obtenus sont présentés à la Section 7.3.

7.3 Résultats

L'évaluation par les différents auditeurs des dégradations des 108 séquences encodées est regroupée dans les Tableaux D-1, D-2 et D-3 de l'Annexe D. Les auditeurs sont indiqués par les lettres A1 à A15. La dernière colonne de ces tableaux indique la note moyenne obtenue par la séquence encodée. Le Tableau 7-1 indique le sexe et l'âge des différents auditeurs.

Auditeur	Sexe	Age
A1	F	25
A2	F	29
A3	F	37
A4	F	42
A5	H	24
A6	H	25
A7	H	25
A8	H	27
A9	H	29
A10	H	30
A11	H	30
A12	H	31
A13	H	32
A14	H	35
A15	H	39

Tableau 7-1 : Sexe et âge des auditeurs (F=Femme; H=Homme).

Le Tableau 7-2 regroupe les DMOS (moyenne globale) obtenus pour chacun des 9 codeurs cités à la Section 7.2.

Codeur	DMOS
P-MRWB-ACELP, 14.3 kbits/s	3.51
P-MRWB-ACELP, 18.5 kbits/s	3.85
P-MRWB-ACELP, 21.5 kbits/s	4.13
G.722, 48 kbits/s	3.72
G.722, 56 kbits/s	4.21
G.722, 64 kbits/s	4.44
WB-AMR, 14.25 kbits/s	4.00
WB-AMR, 18.25 kbits/s	4.31
WB-AMR, 23.05 kbits/s	4.49

Tableau 7-2 : DMOS pour les différents codeurs testés.

L'objectif principal du travail présenté ici était d'implanter un codeur de parole en bande élargie à débit adaptatif variable, et d'obtenir une bonne qualité du signal reconstruit. De plus, un codeur à faible complexité était visé. Ces objectifs ont été atteints.

En se basant sur les contraintes et performances requises par l'ETSI pour le nouveau standard WB-AMR (cf. Section 4.2), nous désirions en outre obtenir un codeur ayant au moins deux modes de fonctionnement permettant une implantation sur le GSM-FR. Un mode de fonctionnement à un débit inférieur ou égal à 14.25 kbits/s était visé, ainsi qu'un mode dépassant le débit de 16 kbits/s. En l'absence d'erreurs de transmission, ces modes auraient dû permettre d'obtenir une qualité du signal reconstruit similaire à celle du G.722 C (48 kbits/s) et respectivement à celle du G.722 B (56 kbits/s). Nous visions également un mode fonctionnant à un débit supérieur, permettant d'obtenir la qualité du G.722 A (64 kbits/s).

Selon l'étude de l'état de l'art présentée à la Section 4.3 et à l'Annexe C.1, et selon la discussion présentée à la Section 5.3, nous pensions atteindre les objectifs sus-mentionnés en implantant un codeur de type ACELP en une seule bande de fréquences. Selon la littérature, un tel codeur aurait permis d'obtenir non seulement une qualité de signal reconstruit similaire à celle du G.722 C pour un débit inférieur à 14.25 kbits/s, mais également une qualité de signal reconstruit similaire à celle du G.722 B (voire G.722 A) à un débit de 16 kbits/s.

Malgré les améliorations en qualité de signal reconstruit, obtenues en appliquant les contributions décrites à la Section 5.7, et malgré l'utilisation de la méthode E (cf. Section 5.8) pour l'extraction de l'excitation adaptative, le Tableau 7-2 montre que les résultats obtenus n'atteignent pas tous les objectifs fixés en termes de qualité de signal reconstruit. En effet, bien que la qualité du signal reconstruit obtenue avec les différents modes du codeur P-MRWB-ACELP est bonne, la qualité du G.722 C n'est atteinte que pour un débit de 18.5 kbits/s, alors que la qualité du G.722 B n'est approchée que pour un débit de 21.5 kbits/s. De plus, les résultats laissent supposer qu'il faudrait un débit supérieur à 24.5 kbits/s pour atteindre la qualité du G.722 A.

Le problème relevé par les auditeurs est lié à la présence d'un léger bruit de type "balai" encore persistant dans certains signaux (principalement des voix de femmes), spécialement pour les débits inférieurs. Bien que ce bruit ne soit pas perçu par certains auditeurs, d'autres le qualifient de légèrement gênant, voire de gênant à très gênant. Certains auditeurs ont parfois l'impression que le signal codé est légèrement filtré par un passe-bas.

La Sous-section 7.3.1 décrit les limites des tests réalisés et les incohérences des résultats obtenus. Ces limites et ces incohérences expliquent

partiellement le fait que la qualité visée, en se basant sur l'étude de l'état de l'art, n'a pas été atteinte.

7.3.1 Limites des tests auditifs

Les tableaux présentés à l'Annexe D montrent les limites des tests auditifs réalisés. Les auditeurs ne sont pas spécifiquement entraînés pour attribuer des notes en se basant sur le DMOS obtenu par d'autres codeurs. Certains résultats sont incohérents puisque les mêmes fichiers sont évalués à 5 et à 1 par différents auditeurs. Bien qu'aucun des codeurs testés n'introduise une perte (même partielle) de l'intelligibilité du signal encodé, et que, quel que soit le codeur évalué ici, le timbre de la voix des locuteurs est inchangé, certains auditeurs attribuent la note 1 à un signal tout à fait intelligible, mais partiellement bruité. Toutefois, la note 1 ne devrait correspondre qu'à une qualité de signal dont une partie de l'intelligibilité est perdue ou affectée.

Le Tableau 7-3 regroupe les DMOS en fonction du sexe des auditeurs, ainsi que les valeurs extrêmes attribuées à chaque codeur. Il montre que les femmes et les hommes ont une perception différente des codeurs testés. Aucune conclusion ne peut être relevée quant à la perception en fonction de l'âge des auditeurs. Le but du Tableau 7-3 est de relever le fait que les tests auditifs sont subjectifs et que leurs résultats dépendent des auditeurs choisis.

Le Tableau 7-4 montre que les résultats dépendent également de la base de données de test. En effet, pour les séquences de parole prononcées par des hommes (adultes), tous nos objectifs sont pratiquement atteints !

Codeur	**Auditeurs femmes**			**Auditeurs hommes**		
	Min.	**DMOS**	**Max.**	**Min.**	**DMOS**	**Max**
P-MRWB-ACELP, 14.3 kbits/s	3	3.85	5	1	3.38	5
P-MRWB-ACELP, 18.5 kbits/s	3	4.27	5	1	3.69	5
P-MRWB-ACELP, 21.5 kbits/s	3	4.35	5	3	4.04	5
G.722, 48 kbits/s	2	3.77	5	1	3.70	5
G.722, 56 kbits/s	3	4.31	5	3	4.17	5
G.722, 64 kbits/s	3	4.54	5	3	4.41	5
WB-AMR, 14.25 kbits/s	3	4.08	5	1	3.97	5
WB-AMR, 18.25 kbits/s	3	4.40	5	2	4.28	5
WB-AMR, 23.05 kbits/s	3	4.65	5	3	4.44	5

Tableau 7-3 : Note minimale attribuée, DMOS, et note maximale attribuée, pour les différents codeurs testés en fonction du sexe des auditeurs.

Le Tableau 7-5 présente les résultats en fonction du sexe des locuteurs et du sexe des auditeurs. Il montre que pour les auditeurs femmes et pour les séquences de parole prononcées par des hommes (adultes), nos objectifs sont dépassés. De même, le Tableau 7-6 montre que nos objectifs sont dépassés en ne considérant que le seul signal prononcé en anglais, T04.

Il aurait été intéressant de tester plus de séquences de parole en anglais et de choisir plus d'auditeurs femmes. Cela pourrait changer la statistique sans rendre les DMOS des modes A, B et C du codeur P-MRWB-ACELP supérieurs ou égaux à ceux des modes C, B et respectivement A du G.722. En effet, le codeur P-MRWB-ACELP présente quelques limites quant à l'encodage des voix de femmes et notamment lorsque le mode A est utilisé. Ces limites sont liées à la technique ACELP appliquée à un signal en bande élargie.

Codeur	**DMOS (Locuteurs femmes)**	**DMOS (Locuteurs hommes)**
P-MRWB-ACELP, 14.3 kbits/s	3.15	3.66
P-MRWB-ACELP, 18.5 kbits/s	3.56	4.00
P-MRWB-ACELP, 21.5 kbits/s	3.88	4.31
G.722, 48 kbits/s	3.55	3.76
G.722, 56 kbits/s	4.24	4.09
G.722, 64 kbits/s	4.51	4.35
WB-AMR, 14.25 kbits/s	3.70	4.22
WB-AMR, 18.25 kbits/s	4.00	4.53
WB-AMR, 23.05 kbits/s	4.27	4.64

Tableau 7-4 : DMOS pour les différents codeurs testés en fonction du sexe des locuteurs (adultes).

Codeur	**DMOS (Locuteurs femmes)**		**DMOS (Locuteurs hommes)**	
	Aud. F	**Aud. H**	**Aud. F**	**Aud. H**
P-MRWB-ACELP, 14.3 kbits/s	3.55	3.01	4.00	3.54
P-MRWB-ACELP, 18.5 kbits/s	4.00	3.4	4.35	3.93
P-MRWB-ACELP, 21.5 kbits/s	4.00	3.83	4.55	4.22
G.722, 48 kbits/s	3.50	3.56	3.80	3.74
G.722, 56 kbits/s	4.30	4.21	4.35	4.00
G.722, 64 kbits/s	4.50	4.51	4.55	4.28
WB-AMR, 14.25 kbits/s	3.75	3.68	4.45	4.13
WB-AMR, 18.25 kbits/s	4.20	3.92	4.60	4.51
WB-AMR, 23.05 kbits/s	4.45	4.21	4.80	4.58

Tableau 7-5 : DMOS pour les différents codeurs testés en fonction du sexe des locuteurs (adultes) et du sexe des auditeurs (Aud. F=Femme, H=Homme), pour les séquences en français.

Codeur	DMOS
P-MRWB-ACELP, 14.3 kbits/s	4.1
P-MRWB-ACELP, 18.5 kbits/s	4.5
P-MRWB-ACELP, 21.5 kbits/s	4.6
G.722, 48 kbits/s	3.5
G.722, 56 kbits/s	4.0
G.722, 64 kbits/s	4.5
WB-AMR, 14.25 kbits/s	4.6
WB-AMR, 18.25 kbits/s	4.6
WB-AMR, 23.05 kbits/s	4.8

Tableau 7-6 : DMOS pour la séquence T04, prononcée en anglais par un homme.

La discussion qui précède montre à quel point les évaluations des codeurs présentés dans l'état de l'art dépendent de la base de données de test utilisée et des auditeurs ayant réalisés les tests. Les résultats décrits par les auteurs des articles qui y sont présentés, sont des tests parfois formels, parfois informels. Ils sont obtenus avec diverses bases de données, soit en anglais, soit en norvégien ou en japonais. Certains tests sont réalisés avec des haut-parleurs et d'autres avec un casque. De plus, certains tests sont effectués avec la méthode en "double aveugle avec référence cachée" utilisée ici, alors que d'autres tests ne sont réalisés que sur la base de la séquence encodée. Dans ce cas, seule la séquence encodée est écoutée par l'auditeur et aucune comparaison avec la séquence originale n'est réalisée. L'échelle est alors dénommée MOS. L'auditeur attribue les notes 1 à 5 s'il considère que la séquence encodée est qualifiée de mauvaise, faible, correcte, bonne et respectivement excellente. Finalement certains tests sont réalisés en comparant directement les séquences encodées avec deux codeurs différents. Considérant toutes ces différences, il n'est pas impossible qu'un codeur évalué similaire au G.722 B dans certaines conditions, soit considéré de qualité supérieure ou de qualité inférieure au G.722 B dans d'autres conditions.

Il est dit en [7-3] que le codeur G.729 [7-4] est évalué à un MOS de 3.9. Nous avons sous-échantillonné le signal T11 à 8 kHz, puis encodé le nouveau signal avec le G.729. T11 encodé avec le P-MRWB-ACELP mode A (DMOS de 3.51) est légèrement bruité et contient un très léger effet métallique, il présente parfois un effet résonnant (tel un écho) similaire à un bruit de balai frappé sur le sol. Ce signal est toutefois totalement intelligible et le timbre de voix du locuteur est conservé. Par contre, T11 encodé avec le G.729 présente continuellement une forte composante bruitée et une consonance métallique. Il contient un effet résonant et présente un caractère étouffé tel un effet passe-bas. Le timbre de voix du locuteur est modifié par rapport à celui de la séquence originale (à 8 kHz). Cet exemple montre que toutes les évaluations résultant des tests auditifs regroupés dans les tableaux en Annexe D et dans le

Tableau 7-2 auraient pu être rehaussées si les tests avaient été faits en MOS. Il est alors possible que les écarts entre les différentes évaluations ainsi qu'entre les différents DMOS diminuent.

7.4 Futurs travaux et conclusions

Le codeur P-MRWB-ACELP montre quelques limites quant à l'encodage des voix de femmes et notamment lorsque le mode A est utilisé. Ces limites sont certainement liées à la technique ACELP codant le signal en bande élargie en une seule bande de fréquences.

Seule l'extraction de l'excitation innovatrice n'a pas fait l'objet d'une étude plus approfondie dans le cadre de ce travail de recherche. Afin d'apporter d'ultérieures améliorations au codeur P-MRWB-ACELP, il serait intéressant d'explorer d'autres méthodes d'extraction de cette excitation. Ces méthodes pourraient être inspirées de l'état de l'art présenté au Chapitre 4. Une extraction de l'excitation innovatrice toutes les 2.5 ms pourrait également être envisagée. De plus, un réglage plus fin des différents coefficients des filtres du codeur P-MRWB-ACELP pourrait être effectué. Cependant, même avec diverses améliorations, il semble difficile que le codeur P-MRWB-ACELP puisse atteindre la qualité du G.722 à 64 kbits/s.

Le codeur P-MRWB-ACELP nous a permis d'atteindre pratiquement tous nos objectifs. C'est un codeur ACELP de très bonne qualité. Cette qualité est d'autant plus satisfaisante que le développement du codeur n'a été réalisé initialement que par 2 personnes, dont l'auteur de cette thèse, puis uniquement par ce dernier. Un tel codeur ne peut naturellement pas concurrencer un codeur développé par des entreprises spécialisées, disposant d'une large équipe de chercheurs travaillant dans le domaine.

Bien que le codeur WB-AMR soit basé sur un codeur de type ACELP, il permet d'obtenir, à débit égal, une qualité de signal reconstruit supérieure à celle du P-MRWB-ACELP. Ceci s'explique aisément puisque le codeur WB-AMR est un codeur en sous-bandes, qui n'encode qu'un signal échantillonné à 12.8 kHz et non pas à 16 kHz (cf. Sous-section 4.4.1). La bande de fréquences comprise entre 6.4 et 7 kHz est recomposée artificiellement sur la base du signal décodé et sur-échantillonné (16 kHz).

7.5 Références

[7-1] ITU-T Recommendation G.722, "7 kHz audio – coding within 64 kbit/s", dans *Blue Book*, *fascicule III.4*, Melbourne, Australie, 1988.

[7-2] 3GPP TS 26.190 V5.0.0 (2001–03) document, dans ftp://ftp.3gpp.org/Specs/2001-09/Rel-5/26_series/ (14 Nov. 2001).

[7-3] http://www.adaptivedigital.com/prod/wh_pap/jan_07.htm (14 Nov. 2001).

[7-4] L. Hanzo, F. Somerville et J. Woodard, "Standard forward-adaptive CELP codecs", Chapter 7, dans *Voice compression and communications*, pp. 207-278, IEEE Series on Digital & Mobile Communication, John Wiley & Sons, Inc., Publication, NY, USA, 2001.

Chapitre 8
Conclusions générales

Le travail de recherche présenté dans ce rapport de thèse a consisté en l'étude et l'implémentation algorithmique d'un codeur de parole en bande élargie, fonctionnant à différents débits. Ce codeur, dénommé le P-MRWB-ACELP (Proprietary Multi-Rate Wide-Band ACELP), a été implanté en code ANSI-C, avec une arithmétique en virgule flottante. L'objectif principal de cette recherche était d'obtenir une bonne qualité du signal reconstruit, tout en visant un codeur à faible complexité, sans toutefois négliger le débit nécessaire à la transmission de l'information.

L'élaboration complète d'un codeur de parole entièrement original dépasse le travail d'une seule thèse. Ainsi l'algorithme implanté se base sur le codeur G.729 de l'ITU, pour la parole en bande étroite. Le choix d'un codeur de type ACELP a été motivé par une étude de l'état de l'art et par les besoins du marché, respectivement antérieurs et simultanés au début de ce travail de thèse.

Le codeur P-MRWB-ACELP fonctionne à 14.3, 18.5 et 21.5 kbits/s. Il permet d'obtenir une bonne qualité de signal reconstruit. Cette qualité est d'autant plus satisfaisante que son développement n'a été réalisé que par une très petite équipe de travail.

Un de nos objectifs initiaux était de développer au moins un mode du codeur, fonctionnant à un débit inférieur ou égal à 14.25 kbits/s. Cet objectif est atteint en remplaçant le quantificateur spectral choisi, par un quantificateur plus complexe, mais ne requérant que 41 bits par 20 ms de signal traité.

Le développement du codeur P-MRWB-ACELP a permis la publication de trois contributions scientifiques et a mené à la description de diverses innovations faisant l'objet de trois brevets. Ces contributions et innovations sont citées à la Section 8.1.

Les deux premières publications, [8-1] et [8-2], concernent l'implantation d'un quantificateur vectoriel séparé et multi-étages, pour l'encodage des paramètres spectraux (LSP). L'implantation de ce quantificateur a mené à la définition d'un nouveau critère de transparence, pour qualifier la quantification des paramètres spectraux d'un signal de parole en bande élargie. Il s'agit du critère C, décrit à la Sous-section 5.6.4. La méthodologie développée pour implanter ce quantificateur est utilisable pour le développement de tout codeur de parole de type LPAS.

Une troisième publication, [8-3], concerne l'utilisation de deux filtres de pondération différents pour l'extraction des excitations adaptative et innovatrice.

Les innovations faisant l'objet de brevets permettent une nette amélioration de la qualité du signal reconstruit, obtenue avec un codeur ACELP encodant de la parole en bande élargie en une seule bande de fréquences. Ces innovations peuvent être ajoutées à d'autres codeurs de type CELP, tels que le nouveau standard WB-AMR, pour en améliorer la qualité. Elles concernent :

- L'introduction d'un pré-filtre pour le dictionnaire adaptatif [8-4].
- Le contrôle de la contribution de l'excitation innovatrice en fonction de l'importance de l'excitation adaptative [8-5].
- L'introduction de deux filtres de pondération formantique différents mais liés, utilisés pour l'extraction des excitations adaptative et innovatrice [8-6].

Les innovations précitées entraînent une forte augmentation de la complexité d'exécution de l'algorithme ACELP usuel. Cette augmentation est nécessaire pour assurer une bonne qualité de signal reconstruit. Toutefois, elle ne semble pas contraignante, puisque les plates-formes de traitement du signal digital (DSP) sont toujours plus rapides et permettent d'implanter des algorithmes toujours plus complexes.

Le codeur P-MRWB-ACELP a de nombreuses applications. Il pourrait être utilisé pour la téléphonie numérique, mobile ou fixe, mais également pour des applications telles que la transmission de la voix sur les réseaux par paquets (par exemple de type Internet), la transmission du signal de parole pour la vidéo-conférence et l'audio-conférence, ainsi que pour la radiodiffusion et la télévision. Le codeur P-MRWB-ACELP peut également être utilisé pour des applications de stockage de la parole ayant pour but, soit un archivage, soit une utilisation différée de l'information enregistrée. Le stockage de séquences de parole pré-enregistrées, permettant de définir une interface "hommes-machines" basée sur la parole, est également envisageable.

De telles séquences peuvent être utilisées dans des jeux informatiques interactifs.

L'utilisation du codeur P-MRWB-ACELP dans des applications telles que la transmission sur les réseaux par paquets, qui à la base ne sont pas prévues pour garantir la même qualité de service que des applications comme la téléphonie, requiert l'introduction de nouveaux blocs fonctionnels. Un tel bloc fonctionnel est notamment un recouvreur d'information pour les paquets perdus. Dans le cas de la téléphonie mobile, un détecteur de l'activité de la voix (VAD) conjugué à un mode d'émission discontinue (DTX), ainsi qu'à un générateur de bruit de confort (CNG) permettent de réduire le débit de transmission pendant les périodes où le signal ne contient pas de parole.

Le fait de se limiter dès le départ à un type unique d'implantation dans cette étude, basée sur l'état de l'art, a été très contraignant. Ceci est d'autant plus vrai que l'état de l'art ne présente pas une méthodologie de tests subjectifs uniforme, permettant de comparer la qualité des codeurs. Une méthodologie uniforme serait utile pour toutes les publications à venir.

Avec du recul, je pense qu'initialement il aurait été judicieux d'implanter en parallèle le squelette de plusieurs types de codeurs. L'effort initial aurait considérablement plus grand, mais il nous aurait permis de confronter une version basique de différents codeurs à l'aide d'une seule base de données de test. Nous aurions ainsi réalisé un choix, en évaluant les modifications à apporter à chaque type de codeur, pour obtenir une bonne qualité de signal reconstruit. Notons que l'évolution de l'état de l'art, de l'année 2000 à l'année 2002, montre un grand intérêt pour les codeurs de type CELP en sous-bandes.

Toutefois, l'implantation d'un seul type de codeur, nous a obligés à explorer minutieusement les limites de ce codeur, afin d'obtenir une bonne qualité de signal reconstruit. Cette méthode de travail nous a mené à développer les diverses innovations sus-mentionnées.

8.1 Publications et brevets

8.1.1 Publications

[8-1] G. Biundo, S. Grassi, M. Ansorge et F. Pellandini, "Spectral quantization for wideband speech coding", dans *Proc. of 1st COST 276 Workshop on information and knowledge management for integrated media communication* (CD-ROM), Leganés (Madrid), Espagne, Nov. 2001.

[8-2] G. Biundo, S. Grassi, M. Ansorge, F. Pellandini et P.-A. Farine, "Design techniques for spectral quantization in wideband speech coding", dans *Proc. of 3rd COST 276 Workshop on information and knowledge management for integrated media communication* (CD-ROM), Budapest, Hongrie, Oct. 2002.

[8-3] G. Biundo, M. Ansorge, F. Pellandini et P.-A. Farine, "Perceptual weighting for ACELP wideband speech coder", dans *Proc. of 4th COST 276 Workshop on information and knowledge management for integrated media communication*, pp. 105-110, Bordeaux, France, Mars-Avril 2003.

8.1.2 Brevets

[8-4] G. Biundo, M. Ansorge et B. Carnero, "Codeur de parole", brevet EP 02 015 918.2, déposé en juillet 2002.

[8-5] G. Biundo, M. Ansorge et B. Carnero, "Codeur de parole à gain réduit", brevet EP 02 015 920.8, déposé en juillet 2002.

[8-6] G. Biundo, M. Ansorge et B. Carnero, "Codeur de parole avec 2 filtres formantiques", brevet EP 02 015 919.0, déposé en juillet 2002.

Annexe A
Extraction des paramètres LSP

A.1 Méthode de Kabal

La méthode de Kabal exploite la symétrie des polynômes $P'(z)$ et $Q'(z)$ donnés par l'équation (3.13). On peut exprimer ces polynômes de la façon suivante, en groupant leurs termes :

$$\begin{aligned} P'(z) &= z^{-8} \cdot \left[(z^{+8} + z^{-8}) + p'_1 (z^{+7} + z^{-7}) + \ldots + p'_8 \right], \\ Q'(z) &= z^{-8} \cdot \left[(z^{+8} + z^{-8}) + q'_1 (z^{+7} + z^{-7}) + \ldots + q'_8 \right], \end{aligned} \tag{A.1}$$

où:

$$\begin{aligned} & p'_i = a_{16}(i) + a_{16}(16 - i + 1) - p'_{i-1}, \\ & q'_i = a_{16}(i) - a_{16}(16 - i + 1) + q'_{i-1}, \\ & \text{avec} \quad i = 1, \ldots, 8, \quad \text{et} \quad p'_0 = 1, \quad q'_0 = 1; \end{aligned} \tag{A.2}$$

et où les $a_{16}(k)$ sont les coefficients LPC d'ordre 16. Une division polynomiale, qui ne demande que des additions et des soustractions, a été réalisée sur les coefficients de $P(z)$ et $Q(z)$ de l'équation (3.13), afin d'éliminer les solutions triviales en $z = \pm 1$. Si l'on considère les polynômes $P'(z)$ et $Q'(z)$ sur le cercle unité en posant $z = e^{j\omega}$, et que l'on normalise par un facteur 2, on a :

$$\tilde{P}'(\omega) = \frac{P'(\omega)}{2e^{-8j\omega}} = \cos(8\omega) + p'_1 \cos(7\omega) + \ldots + p'_7 \cos(\omega) + \frac{p'_8}{2},$$
$$\tilde{Q}'(\omega) = \frac{Q'(\omega)}{2e^{-8j\omega}} = \cos(8\omega) + q'_1 \cos(7\omega) + \ldots + q'_7 \cos(\omega) + \frac{q'_8}{2}. \quad \text{(A.3)}$$

Les polynômes de Tchebychev de type I sont donnés par [A-1]:

$$T_n(x) = \cos(n \cos^{-1} x), \quad \text{si } |x| \le 1.0. \quad \text{(A.4)}$$

Ils peuvent être générés grâce à l'équation récursive suivante :

$$T_n(x) = 2xT_{n-1}(x) - T_{n-2}(x). \quad \text{(A.5)}$$

En posant $x = \cos(\omega)$ dans l'équation (A.4), on obtient pour n compris entre 0 et 8, les relations suivantes :

$$\begin{aligned}
T_0(x) &= \cos(0) = 1, \\
T_1(x) &= \cos(\omega) = x, \\
T_2(x) &= \cos(2\omega) = 2x^2 - 1, \\
T_3(x) &= \cos(3\omega) = 4x^3 - 3x, \\
T_4(x) &= \cos(4\omega) = 8x^4 - 8x^2 + 1, \\
T_5(x) &= \cos(5\omega) = 16x^5 - 20x^3 + 5x, \\
T_6(x) &= \cos(6\omega) = 32x^6 - 48x^4 + 18x^2 - 1, \\
T_7(x) &= \cos(7\omega) = 64x^7 - 112x^5 + 56x^3 - 7x, \\
T_8(x) &= \cos(8\omega) = 128x^8 - 256x^6 + 160x^4 - 32x^2 + 1.
\end{aligned} \quad \text{(A.6)}$$

En utilisant ces relations dans l'équation (A.3), on obtient les deux polynômes d'ordre 8 , $\tilde{P}'(x)$ et $\tilde{Q}'(x)$ suivants :

$$\begin{aligned}
\tilde{P}'(x) = \; & 128x^8 + 64p'_1x^7 + (-256 + 32p'_2)x^6 + \\
& (-112p'_1 + 16p'_3)x^5 + (160 - 48p'_2 + 8p'_4)x^4 + \\
& (56p'_1 - 20p'_3 + 4p'_5)x^3 + (-32 + 18p'_2 - 8p'_4 + 2p'_6)x^2 + \\
& (-7p'_1 + 5p'_3 - 3p'_5 + p'_7)x + (1 - p'_2 + p'_4 - p'_6 + \frac{p'_8}{2}),
\end{aligned} \quad \text{(A.7)}$$

$$\begin{aligned}\tilde{Q}'(x) = \; & 128x^8 + 64q'_1x^7 + (-256 + 32q'_2)x^6 + \\ & (-112q'_1 + 16q'_3)x^5 + (160 - 48q'_2 + 8q'_4)x^4 + \\ & (+56q'_1 - 20q'_3 + 4q'_5)x^3 + (-32 + 18q'_2 - 8q'_4 + 2q'_6)x^2 + \\ & (-7q'_1 + 5q'_3 - 3q'_5 + q'_7)x + (1 - q'_2 + q'_4 - q'_6 + \frac{q'_8}{2}).\end{aligned}$$

Les racines de $\tilde{P}'(x)$ et $\tilde{Q}'(x)$ sont les LSP dans le domaine "x", soit $\{x_i\}$, avec $x_i = \cos(\omega_i)$. L'équation (3.16) dans ce même domaine devient :

$$+1.0 > x_1 > x_2 > \ldots > x_{16} > -1.0. \qquad \text{(A.8)}$$

Comme $\tilde{P}'(x)$ et $\tilde{Q}'(x)$ sont des polynômes du 8ème degré, leurs racines ne peuvent être calculées de façon exacte. En [A-2], Kabal propose d'utiliser le "zero crossing", ou recherche du passage par le niveau zéro. La recherche se fait en partant de $x = +1.0$, en décrémentant itérativement cette valeur de $\Delta = 0.02$. Lorsque l'on détecte un intervalle contenant un passage par zéro, on redéfinit la position de celui-ci. On utilise d'abord 4 bissections consécutives. Ensuite, on réalise une interpolation linéaire. Etant donnée la propriété d'ordre des racines, la recherche est réalisée alternativement entre $\tilde{P}'_{16}(x)$ et $\tilde{Q}'_{16}(x)$ en partant de la position de la dernière racine trouvée. Le choix de la largeur de la grille, Δ, et du nombre de bissections à tester est fonction du nombre de LPC et de la fréquence d'échantillonnage. Ces valeurs sont évaluées statistiquement. Elles doivent permettre d'éviter l'oubli d'une intersection par zéro. Kabal considère une analyse LPC d'ordre 10 pour de la parole en bande étroite. Pour le codeur présenté ici, Δ a été fixé à 0.0077.

A.2 Références

[A-1] J. Proakis et D. Manolakis, "Design of digital filters", Chapter 8, dans *Digital signal processing. Principles, algorithms, and applications*, 3rd ed., pp. 852-895, Prentice Hall International Editions, Inc., 1996.

[A-2] P. Kabal et P. Ramachandran, "The computation of line spectral frequencies using chebyshev polynomials", dans *IEEE Trans. on acoustics, speech and signal processing*, Vol. 34, No. 6, pp. 1419-1426, 1986.

Annexe B
Transformation de LSP en LPC

B.1 Méthode d'expansion directe

Les polynômes symétriques et anti-symétriques $P(z)$ et $Q(z)$ sont donnés par :

$$\begin{aligned} P(z) &= (1+z^{-1}) \prod_{i=1,3,5,7,9,11,13,15} \left[1 + c_i \cdot z^{-1} + z^{-2}\right] = \sum_{i=0}^{17} p_i z^{-i}, \\ Q(z) &= (1-z^{-1}) \prod_{i=2,4,6,8,10,12,14,16} \left[1 + c_i \cdot z^{-1} + z^{-2}\right] = \sum_{i=0}^{17} q_i z^{-i}, \\ &\text{avec } c_i = -2\cos(\omega_i), \end{aligned} \tag{B.1}$$

où les $\{\omega_i\}$ sont les coefficients LSP. On trouve les coefficients $\{p_i\}$ et $\{q_i\}$ en multipliant les termes des produits de l'équation (B.1). On a :

$$\begin{aligned} p_0 &= p_{17} = 1, \\ p_1 &= p_{16} = 1 + s_1, \\ p_2 &= p_{15} = 8 + s_1 + s_2, \\ p_3 &= p_{14} = 8 + 7s_1 + s_2 + s_3, \\ p_4 &= p_{13} = 28 + 7s_1 + 6s_2 + s_3 + s_4, \\ p_5 &= p_{12} = 28 + 21s_1 + 6s_2 + 5s_3 + s_4 + s_5, \\ p_6 &= p_{11} = 56 + 21s_1 + 15s_2 + 5s_3 + 4s_4 + s_5 + s_6, \\ p_7 &= p_{10} = 56 + 35s_1 + 15s_2 + 10s_3 + 4s_4 + 3s_5 + s_6 + s_7, \\ p_8 &= p_9 = 70 + 35s_1 + 20s_2 + 10s_3 + 6s_4 + 3s_5 + 2s_6 + s_7 + s_8; \end{aligned} \tag{B.2}$$

$$
\begin{aligned}
q_0 &= -q_{17} = 1,\\
q_1 &= -q_{16} = -1 + s'_1,\\
q_2 &= -q_{15} = 8 - s'_1 + s'_2,\\
q_3 &= -q_{14} = -8 + 7s'_1 - s'_2 + s'_3,\\
q_4 &= -q_{13} = 28 - 7s'_1 + 6s'_2 - s'_3 + s'_4,\\
q_5 &= -q_{12} = -28 + 21s'_1 - 6s'_2 + 5s'_3 - s'_4 + s'_5,\\
q_6 &= -q_{11} = 56 - 21s'_1 + 15s'_2 - 5s'_3 + 4s'_4 - s'_5 + s'_6,\\
q_7 &= -q_{10} = -56 + 35s'_1 - 15s'_2 + 10s'_3 - 4s'_4 + 3s'_5 - s'_6 + s'_7,\\
q_8 &= -q_9 = 70 - 35s'_1 + 20s'_2 - 10s'_3 + 6s'_4 - 3s'_5 + 2s'_6 - s'_7 + s'_8,
\end{aligned}
\tag{B.3}
$$

où les $\{s_i\}$ et les $\{s'_i\}$ sont les sommes des produits des termes c_i de suffixes impaires et respectivement paires donnés par :

$$
s_1 = c_1 + c_3 + c_5 + c_7 + c_9 + c_{11} + c_{13} + c_{15};
\tag{B.4}
$$

$$
\begin{aligned}
s_2 = {} & c_1c_3 + c_1c_5 + c_1c_7 + c_1c_9 + c_1c_{11} + c_1c_{13} + c_1c_{15} + c_3c_5 + c_3c_7\\
& + c_3c_9 + c_3c_{11} + c_3c_{13} + c_3c_{15} + c_5c_7 + c_5c_9 + c_5c_{11} + c_5c_{13}\\
& + c_5c_{15} + c_7c_9 + c_7c_{11} + c_7c_{13} + c_7c_{15} + c_9c_{11} + c_9c_{13} + c_9c_{15}\\
& + c_{11}c_{13} + c_{11}c_{15} + c_{13}c_{15};
\end{aligned}
\tag{B.5}
$$

$$
\begin{aligned}
s_3 = {} & c_1c_3c_5 + c_1c_3c_7 + c_1c_3c_9 + c_1c_3c_{11} + c_1c_3c_{13} + c_1c_3c_{15}\\
& + c_1c_5c_7 + c_1c_5c_9 + c_1c_5c_{11} + c_1c_5c_{13} + c_1c_5c_{15} + c_1c_7c_9\\
& + c_1c_7c_{11} + c_1c_7c_{13} + c_1c_7c_{15} + c_1c_9c_{11} + c_1c_9c_{13} + c_1c_9c_{15}\\
& + c_1c_{11}c_{13} + c_1c_{11}c_{15} + c_1c_{13}c_{15} + c_3c_5c_7 + c_3c_5c_9 + c_3c_5c_{11}\\
& + c_3c_5c_{13} + c_3c_5c_{15} + c_3c_7c_9 + c_3c_7c_{11} + c_3c_7c_{13} + c_3c_7c_{15}\\
& + c_3c_9c_{11} + c_3c_9c_{13} + c_3c_9c_{15} + c_3c_{11}c_{13} + c_3c_{11}c_{15} + c_3c_{13}c_{15}\\
& + c_5c_7c_9 + c_5c_7c_{11} + c_5c_7c_{13} + c_5c_7c_{15} + c_5c_9c_{11} + c_5c_9c_{13}\\
& + c_5c_9c_{15} + c_5c_{11}c_{13} + c_5c_{11}c_{15} + c_5c_{13}c_{15} + c_7c_9c_{11} + c_7c_9c_{13}\\
& + c_7c_9c_{15} + c_7c_{11}c_{13} + c_7c_{11}c_{15} + c_7c_{13}c_{15} + c_9c_{11}c_{13} + c_9c_{11}c_{15}\\
& + c_9c_{13}c_{15} + c_{11}c_{13}c_{15};
\end{aligned}
\tag{B.6}
$$

$$
\begin{aligned}
s_4 = {} & c_1c_3c_5c_7 + c_1c_3c_5c_9 + c_1c_3c_5c_{11} + c_1c_3c_5c_{13} + c_1c_3c_5c_{15} \\
& + c_1c_3c_7c_9 + c_1c_3c_7c_{11} + c_1c_3c_7c_{13} + c_1c_3c_7c_{15} + c_1c_3c_9c_{11} \\
& + c_1c_3c_9c_{13} + c_1c_3c_9c_{15} + c_1c_3c_{11}c_{13} + c_1c_3c_{11}c_{15} + c_1c_3c_{13}c_{15} \\
& + c_1c_5c_7c_9 + c_1c_5c_7c_{11} + c_1c_5c_7c_{13} + c_1c_5c_7c_{15} + c_1c_5c_9c_{11} \\
& + c_1c_5c_9c_{13} + c_1c_5c_9c_{15} + c_1c_5c_{11}c_{13} + c_1c_5c_{11}c_{15} + c_1c_5c_{13}c_{15} \\
& + c_1c_7c_9c_{11} + c_1c_7c_9c_{13} + c_1c_7c_9c_{15} + c_1c_7c_{11}c_{13} + c_1c_7c_{11}c_{15} \\
& + c_1c_7c_{13}c_{15} + c_1c_9c_{11}c_{13} + c_1c_9c_{11}c_{15} + c_1c_9c_{13}c_{15} + c_1c_{11}c_{13}c_{15} \\
& + c_3c_5c_7c_9 + c_3c_5c_7c_{11} + c_3c_5c_7c_{13} + c_3c_5c_7c_{15} + c_3c_5c_9c_{11} \\
& + c_3c_5c_9c_{13} + c_3c_5c_9c_{15} + c_3c_5c_{11}c_{13} + c_3c_5c_{11}c_{15} + c_3c_5c_{13}c_{15} \\
& + c_3c_7c_9c_{11} + c_3c_7c_9c_{13} + c_3c_7c_9c_{15} + c_3c_7c_{11}c_{13} + c_3c_7c_{11}c_{15} \\
& + c_3c_7c_{13}c_{15} + c_3c_9c_{11}c_{13} + c_3c_9c_{11}c_{15} + c_3c_9c_{13}c_{15} + c_3c_{11}c_{13}c_{15} \\
& + c_5c_7c_9c_{11} + c_5c_7c_9c_{13} + c_5c_7c_9c_{15} + c_5c_7c_{11}c_{13} + c_5c_7c_{11}c_{15} \\
& + c_5c_7c_{13}c_{15} + c_5c_9c_{11}c_{13} + c_5c_9c_{11}c_{15} + c_5c_9c_{13}c_{15} + c_5c_{11}c_{13}c_{15} \\
& + c_7c_9c_{11}c_{13} + c_7c_9c_{11}c_{15} + c_7c_9c_{13}c_{15} + c_7c_{11}c_{13}c_{15} + c_9c_{11}c_{13}c_{15};
\end{aligned}
\tag{B.7}
$$

$$
\begin{aligned}
s_5 = {} & c_1c_3c_5c_7c_9 + c_1c_3c_5c_7c_{11} + c_1c_3c_5c_7c_{13} + c_1c_3c_5c_7c_{15} \\
& + c_1c_3c_5c_9c_{11} + c_1c_3c_5c_9c_{13} + c_1c_3c_5c_9c_{15} + c_1c_3c_5c_{11}c_{13} \\
& + c_1c_3c_5c_{11}c_{15} + c_1c_3c_5c_{13}c_{15} + c_1c_3c_7c_9c_{11} + c_1c_3c_7c_9c_{13} \\
& + c_1c_3c_7c_9c_{15} + c_1c_3c_7c_{11}c_{13} + c_1c_3c_7c_{11}c_{15} + c_1c_3c_7c_{13}c_{15} \\
& + c_1c_3c_9c_{11}c_{13} + c_1c_3c_9c_{11}c_{15} + c_1c_3c_9c_{13}c_{15} + c_1c_3c_{11}c_{13}c_{15} \\
& + c_1c_5c_7c_9c_{11} + c_1c_5c_7c_9c_{13} + c_1c_5c_7c_9c_{15} + c_1c_5c_7c_{11}c_{13} \\
& + c_1c_5c_7c_{11}c_{15} + c_1c_5c_7c_{13}c_{15} + c_1c_5c_9c_{11}c_{13} + c_1c_5c_9c_{11}c_{15} \\
& + c_1c_5c_9c_{13}c_{15} + c_1c_5c_{11}c_{13}c_{15} + c_1c_7c_9c_{11}c_{13} + c_1c_7c_9c_{11}c_{15} \\
& + c_1c_7c_9c_{13}c_{15} + c_1c_7c_{11}c_{13}c_{15} + c_1c_9c_{11}c_{13}c_{15} + c_3c_5c_7c_9c_{11} \\
& + c_3c_5c_7c_9c_{13} + c_3c_5c_7c_9c_{15} + c_3c_5c_7c_{11}c_{13} + c_3c_5c_7c_{11}c_{15} \\
& + c_3c_5c_7c_{13}c_{15} + c_3c_5c_9c_{11}c_{13} + c_3c_5c_9c_{11}c_{15} + c_3c_5c_9c_{13}c_{15} \\
& + c_3c_5c_{11}c_{13}c_{15} + c_3c_7c_9c_{11}c_{13} + c_3c_7c_9c_{11}c_{15} + c_3c_7c_9c_{13}c_{15} \\
& + c_3c_7c_{11}c_{13}c_{15} + c_3c_9c_{11}c_{13}c_{15} + c_5c_7c_9c_{11}c_{13} + c_5c_7c_9c_{11}c_{15} \\
& + c_5c_7c_9c_{13}c_{15} + c_5c_7c_{11}c_{13}c_{15} + c_5c_9c_{11}c_{13}c_{15} + c_7c_9c_{11}c_{13}c_{15};
\end{aligned}
\tag{B.8}
$$

$$
\begin{aligned}
s_6 &= c_1c_3c_5c_7c_9c_{11} + c_1c_3c_5c_7c_9c_{13} + c_1c_3c_5c_7c_9c_{15} + c_1c_3c_5c_7c_{11}c_{13} \\
&+ c_1c_3c_5c_7c_{11}c_{15} + c_1c_3c_5c_7c_{13}c_{15} + c_1c_3c_5c_9c_{11}c_{13} + c_1c_3c_5c_9c_{11}c_{15} \\
&+ c_1c_3c_5c_9c_{13}c_{15} + c_1c_3c_5c_{11}c_{13}c_{15} + c_1c_3c_7c_9c_{11}c_{13} + c_1c_3c_7c_9c_{11}c_{15} \\
&+ c_1c_3c_7c_9c_{13}c_{15} + c_1c_3c_7c_{11}c_{13}c_{15} + c_1c_3c_9c_{11}c_{13}c_{15} + c_1c_5c_7c_9c_{11}c_{13} \\
&+ c_1c_5c_7c_9c_{11}c_{15} + c_1c_5c_7c_9c_{13}c_{15} + c_1c_5c_7c_{11}c_{13}c_{15} + c_1c_5c_9c_{11}c_{13}c_{15} \\
&+ c_1c_7c_9c_{11}c_{13}c_{15} + c_3c_5c_7c_9c_{11}c_{13} + c_3c_5c_7c_9c_{11}c_{15} + c_3c_5c_7c_9c_{13}c_{15} \\
&+ c_3c_5c_7c_{11}c_{13}c_{15} + c_3c_5c_9c_{11}c_{13}c_{15} + c_3c_7c_9c_{11}c_{13}c_{15} + c_5c_7c_9c_{11}c_{13}c_{15};
\end{aligned}
\tag{B.9}
$$

$$
\begin{aligned}
s_7 &= c_1c_3c_5c_7c_9c_{11}c_{13} + c_1c_3c_5c_7c_9c_{11}c_{15} + c_1c_3c_5c_7c_9c_{13}c_{15} \\
&+ c_1c_3c_5c_7c_{11}c_{13}c_{15} + c_1c_3c_5c_9c_{11}c_{13}c_{15} + c_1c_3c_7c_9c_{11}c_{13}c_{15} \\
&+ c_1c_5c_7c_9c_{11}c_{13}c_{15} + c_3c_5c_7c_9c_{11}c_{13}c_{15};
\end{aligned}
\tag{B.10}
$$

$$
s_8 = c_1c_3c_5c_7c_9c_{11}c_{13}c_{15}. \tag{B.11}
$$

Les $s'_1, s'_2, \ldots, s'_8,$ sont respectivement les équivalents de $s_1, s_2, \ldots, s_8,$ où les indices 1, 3, 5, 7, 9, 11, 13 et 15 sont respectivement remplacés par les indices 2, 4, 6, 8,10,12,14 et 16. Finalement, le filtre LPC est donné par :

$$A_{16}(z) = (P(z) + Q(z))/2. \tag{B.12}$$

La complexité totale de calcul est donc de 618 additions et 519 multiplications. Par rapport à la complexité calculée en [B-1] pour un ordre $p = 10$, on a environ 7 fois plus d'additions et 10 fois plus de multiplications !

B.2 Méthode de Kabal

Les polynômes $P(z)$ et $Q(z)$ donnés par :

$$
\begin{aligned}
P(z) &= (1 + z^{-1}) \sum_{i=0}^{16} p'_i \, z^{-i}, \\
Q(z) &= (1 - z^{-1}) \sum_{i=0}^{16} q'_i \, z^{-i},
\end{aligned}
\tag{B.13}
$$

sont symétriques et anti-symétriques. $P'(z)$ et $Q'(z)$ valent :

$$P'(z) = \sum_{i=0}^{16} p'_i \, z^{-i}; \qquad Q'(z) = \sum_{i=0}^{16} q'_i \, z^{-i}. \tag{B.14}$$

Comme $P'(z)$ et $Q'(z)$ sont symétriques, seuls leurs 8 premiers coefficients sont nécessaires. Les coefficients $\{p'_i\}$ et $\{q'_i\}$ s'obtiennent comme suit :

$$
\begin{aligned}
c_{10} &= -x_1, & c'_{10} &= -x_2,\\
c_{20} &= -2\cdot x_3\cdot c_{10}+1, & c'_{20} &= -2\cdot x_4\cdot c'_{10}+1,\\
c_{21} &= 2\cdot c_{10}-2\cdot x_3, & c'_{21} &= 2\cdot c'_{10}-2\cdot x_4,\\
c_{30} &= -2\cdot x_5\cdot c_{20}+c_{21}, & c'_{30} &= -2\cdot x_6\cdot c'_{20}+c_{21},\\
c_{31} &= 2\cdot c_{20}-2\cdot x_5\cdot c_{21}+1, & c'_{31} &= 2\cdot c_{20}-2\cdot x_6\cdot c'_{21}+1,\\
c_{32} &= c_{21}-2\cdot x_5, & c'_{32} &= c'_{21}-2\cdot x_6,\\
c_{40} &= -2\cdot x_7\cdot c_{30}+c_{31}, & c'_{40} &= -2\cdot x_8\cdot c'_{30}+c'_{31},\\
c_{41} &= 2\cdot c_{30}-2\cdot x_7\cdot c_{31}+c_{32}, & c'_{41} &= 2\cdot c'_{30}-2\cdot x_8\cdot c'_{31}+c'_{32},\\
c_{42} &= c_{31}-2\cdot x_7\cdot c_{32}+1, & c'_{42} &= c'_{31}-2\cdot x_8\cdot c_{32}+1,\\
c_{43} &= c_{32}-2\cdot x_7, & c'_{43} &= c'_{32}-2\cdot x_8,\\
c_{50} &= -2\cdot x_9\cdot c_{40}+c_{41}, & c'_{50} &= -2\cdot x_{10}\cdot c'_{40}+c'_{41},\\
c_{51} &= 2\cdot c_{40}-2\cdot x_9\cdot c_{41}+c_{42}, & c'_{51} &= 2\cdot c'_{40}-2\cdot x_{10}\cdot c'_{41}+c'_{42},\\
c_{52} &= c_{41}-2\cdot x_9\cdot c_{42}+c_{43}, & c'_{52} &= c'_{41}-2\cdot x_{10}\cdot c'_{42}+c'_{43},\\
c_{53} &= c_{42}-2\cdot x_9\cdot c_{43}+1, & c'_{53} &= c'_{42}-2\cdot x_{10}\cdot c'_{43}+1,\\
c_{54} &= c_{43}-2\cdot x_9, & c'_{54} &= c'_{43}-2\cdot x_{10},\\
c_{60} &= -2\cdot x_{11}\cdot c_{50}+c_{51}, & c'_{60} &= -2\cdot x_{12}\cdot c'_{50}+c'_{51},\\
c_{61} &= 2\cdot c_{50}-2\cdot x_{11}\cdot c_{51}+c_{52}, & c'_{61} &= 2\cdot c'_{50}-2\cdot x_{12}\cdot c'_{51}+c'_{52},\\
c_{62} &= c_{51}-2\cdot x_{11}\cdot c_{52}+c_{53}, & c'_{62} &= c'_{51}-2\cdot x_{12}\cdot c'_{52}+c'_{53},\\
c_{63} &= c_{52}-2\cdot x_{11}\cdot c_{53}+c_{54}, & c'_{63} &= c'_{52}-2\cdot x_{12}\cdot c'_{53}+c'_{54},\\
c_{64} &= c_{53}-2\cdot x_{11}\cdot c_{54}+1, & c'_{64} &= c'_{53}-2\cdot x_{12}\cdot c'_{54}+1,\\
c_{65} &= c_{54}-2\cdot x_{11}, & c'_{65} &= c'_{54}-2\cdot x_{12},\\
c_{70} &= -2\cdot x_{13}\cdot c_{60}+c_{61}, & c'_{70} &= -2\cdot x_{14}\cdot c'_{60}+c'_{61},\\
c_{71} &= 2\cdot c_{60}-2\cdot x_{13}\cdot c_{61}+c_{62}, & c'_{71} &= 2\cdot c_{60}-2\cdot x_{14}\cdot c'_{61}+c'_{62},\\
c_{72} &= c_{61}-2\cdot x_{13}\cdot c_{62}+c_{63}, & c'_{72} &= c'_{61}-2\cdot x_{14}\cdot c'_{62}+c'_{63},\\
c_{73} &= c_{62}-2\cdot x_{13}\cdot c_{63}+c_{64}, & c'_{73} &= c'_{62}-2\cdot x_{14}\cdot c'_{63}+c'_{64},\\
c_{74} &= c_{63}-2\cdot x_{13}\cdot c_{64}+c_{65}, & c'_{74} &= c'_{63}-2\cdot x_{14}\cdot c'_{64}+c'_{65},\\
c_{75} &= c_{64}-2\cdot x_{13}\cdot c_{65}+1, & c'_{75} &= c'_{64}-2\cdot x_{14}\cdot c'_{65}+1,\\
c_{76} &= c_{65}-2\cdot x_{13}, & c'_{76} &= c'_{65}-2\cdot x_{14},\\
c_{80} &= -2\cdot x_{15}\cdot c_{70}+c_{71}, & c'_{80} &= -2\cdot x_{16}\cdot c'_{70}+c'_{71},\\
c_{81} &= 2\cdot c_{70}-2\cdot x_{15}\cdot c_{71}+c_{72}, & c'_{81} &= 2\cdot c'_{70}-2\cdot x_{16}\cdot c'_{71}+c'_{72},\\
c_{82} &= c_{71}-2\cdot x_{15}\cdot c_{72}+c_{73}, & c'_{82} &= c'_{71}-2\cdot x_{16}\cdot c'_{72}+c'_{73},\\
c_{83} &= c_{72}-2\cdot x_{15}\cdot c_{73}+c_{74}, & c'_{83} &= c'_{72}-2\cdot x_{16}\cdot c'_{73}+c'_{74},\\
c_{84} &= c_{73}-2\cdot x_{15}\cdot c_{74}+c_{75}, & c'_{84} &= c'_{73}-2\cdot x_{16}\cdot c'_{74}+c'_{75},\\
c_{85} &= c_{74}-2\cdot x_{15}\cdot c_{75}+c_{76}, & c'_{85} &= c'_{74}-2\cdot x_{16}\cdot c'_{75}+c'_{76},\\
c_{86} &= c_{75}-2\cdot x_{15}\cdot c_{76}+1, & c'_{86} &= c'_{75}-2\cdot x_{16}\cdot c'_{76}+1,\\
c_{87} &= c_{76}-2\cdot x_{15}; & c'_{87} &= c'_{76}-2\cdot x_{16};
\end{aligned}
\qquad \text{(B.15)}
$$

où les $\{x_i\}$ sont les LSP dans le domaine "x", avec $x_i = \cos(\omega_i)$. Les derniers termes de cette récursion donnent les coefficients $\{p'_i\}$ et $\{q'_i\}$:

$$\begin{aligned} p'_i &= c_{88-i}, \ i = 1,...,7, & q'_i &= c'_{88-i}, \ i = 1,...,7, \\ p'_8 &= 2 \cdot c_{80}; & q'_8 &= 2 \cdot c'_{80}. \end{aligned} \tag{B.16}$$

Selon les équations (B.12) et (B.13), les coefficients LPC sont donnés par :

$$\begin{aligned} a_{16}(1) &= (p'_1 + q'_1)/2, \\ a_{16}(i) &= \frac{(p'_i + p'_{i-1}) + (q'_i - q'_{i-1})}{2}, \quad i = 2,...,8; \\ a_{16}(i) &= \frac{(p'_{17-i} + p'_{16-i}) - (q'_{17-i} - q'_{16-i})}{2}, \quad i = 9,...,15; \\ a_{16}(16) &= (p'_1 - q'_1 + 2)/2. \end{aligned} \tag{B.17}$$

La complexité totale de calcul de la méthode de Kabal est de 88 multiplications et de 143 additions. Cette complexité est beaucoup plus petite que la complexité de la méthode d'expansion directe, décrite à la Section B.1.

B.3 Références

[B-1] S. Grassi, "Line spectrum pairs and the CELP FS1016 speech coder", Chapter 5, dans *Optimized implementation of speech processing algorithms*, pp. 73-112, Thèse éditée par la Faculté des Sciences de l'Université de Neuchâtel, 1998.

Annexe C
Catégories et sous-catégories des codeurs de l'état de l'art

C.1 Catégories et sous-catégories des codeurs de l'état de l'art jusqu'en 1999

L'état de l'art jusqu'en 1999, présenté ici et discuté à la Section 4.3, fait apparaître trois catégories de codeurs : les codeurs de type CELP, les codeurs par transformée et les codeurs mixtes. Les codeurs de type CELP se divisent en deux sous-catégories : les codeurs encodant le signal en une seule bande de fréquences, et les codeurs séparant le signal en plusieurs sous-bandes de fréquences avec encodage séparé de chaque sous-bande. L'état de l'art est présenté ici par catégories et sous-catégories. A l'intérieur d'une même catégorie, les articles sont classés selon leur ordre chronologique de parution.

C.1.1 Codeurs de type CELP

1. L'article de Laflamme et *al.* [C-1] décrit un codeur de type ACELP, similaire à celui du G.729. Une technique de recherche partielle est proposée pour réaliser une recherche efficace dans le dictionnaire algébrique. Cette technique permet de réduire la complexité algorithmique du codeur. A 13 kbits/s, la "qualité" est dite "haute".
2. L'article de Salami et *al.* [C-2] décrit un codeur de type ACELP, qui permet une implantation en temps réel sur le DSP TMS320C30. Ce codeur est basé sur celui de Laflamme et *al.* décrit ci-dessus. Il fonctionne à 9.6 et à 14 kbits/s. Son dictionnaire innovateur est formé d'un dictionnaire algébrique et d'un dictionnaire à excitations

impulsionnelles binaires régulières. La recherche de l'excitation innovatrice se fait par étages, en commençant par le dictionnaire algébrique. A 14 kbits/s, la "qualité" est celle du G.722 B.

3. L'article de Harborg et *al.* [C-3] décrit un codeur de type CELP fonctionnant à un débit de 16 kbits/s. Le codeur présenté est basé sur celui décrit en [C-4]. Le dictionnaire innovateur est composé de vecteurs non recouvrants. Ces vecteurs ne contiennent que trois valeurs non-nulles, générées par un processus gaussien. Trois versions du codeur sont présentées et chacune permet une implantation en temps réel sur le DSP TMS320C31. La "qualité" est celle du G.722 A.
4. L'article de McElroy et *al.* [C-5] décrit un codeur de type CELP, utilisant de multiples dictionnaires innovateurs. L'excitation innovatrice résulte de la combinaison de chaque vecteur sélectionné dans les différents dictionnaires. Chaque dictionnaire est à bande de fréquences limitée. Une telle structure de dictionnaires permet d'accorder plus d'importance aux fréquences du signal les plus perceptibles pour l'oreille. Ce codeur fonctionne à 16 et à 24 kbits/s. A 16 kbits/s, la "qualité" est celle du G.722 B et C, pour respectivement les voix d'hommes et de femmes.
5. L'article de Black et *al.* [C-6] décrit un codeur à 14.1 kbits/s, par prédiction linéaire excitée par un signal résiduel composé d'impuslions : PRELP (Pulsed Residual Excited Linear-Prediction). Ce codeur est une variante d'un codeur CELP. Il contient deux dictionnaires d'excitations innovatrices. Le premier dictionnaire est formé d'impulsions répétées périodiquement. Il modélise la périodicité dans les phases de transition du signal, là où le filtre de prédiction à long-terme se montre inefficace. Le second dictionnaire ne contient que quelques valeurs non-nulles, qui sont des séquences gaussiennes. L'excitation innovatrice est choisie dans l'un des deux dictionnaires. La "qualité" est celle du G.722 C.
6. L'article de Sasaki et *al.* [C-7] décrit un codeur de type CELP à 16 kbits/s, où l'excitation innovatrice est quantifiée sur deux étages. Pour chaque étage, le dictionnaire est divisé en deux sous-dictionnaires à structures conjuguées. Les vecteurs des dictionnaires ne contiennent que quelques valeurs non-nulles. La "qualité" est celle du G.722 B.
7. L'article de Serizawa et *al.* [C-8] décrit un codeur de type MP-CELP (Multi-Pulse based CELP) à 16 kbits/s, qui fait usage d'une prédiction LPC ordinaire ("forward"), d'une prédiction LTP et d'une prédiction calculée de façon "backward", basée sur le signal résiduel (LPC) quantifié de la trame précédente. La quantification de l'excitation innovatrice est réalisée différemment en fonction du caractère voisé ou non-voisé du signal traité. La "qualité" est celle du G.722 B.

8. L'article de Koishida et *al.* [C-9] décrit un codeur de type ACELP à 16 kbits/s. L'analyse de l'enveloppe spectrale courante est remplacée ici par une analyse cepstrale, selon l'échelle des Mels généralisée : MGC (Mel-Generalized Cepstral analysis). La "qualité" est celle du G.722 A.

C.1.2 Codeurs de type SB-CELP

9. L'article de Roy et Kabal [C-10] décrit un codeur de type SB-CELP à 16 kbits/s. L'enveloppe spectrale du signal est calculée à partir du signal en une seule bande. Le signal résiduel est séparé en deux sous-bandes de fréquences (0-4 et 4-8 kHz), et les paramètres qualifiant l'excitation sont extraits pour chacune des sous-bandes, en accordant plus d'importance à la bande inférieure. Les auteurs ne donnent aucune comparaison qualitative par rapport à un autre codeur en bande élargie.
10. L'article de McElroy et *al.* [C-11] décrit un codeur de type SB-CELP. La bande inférieure (0-4 kHz) est encodée à l'aide d'un codeur CELP typiquement utilisé pour la bande étroite. La bande supérieure est encodée à l'aide d'une prédiction linéaire d'ordre 2 et d'un dictionnaire non-recouvrant de petites dimensions. Ce codeur fonctionne à des débits variant de 7.2 à 14.4 kbits/s. Sa "qualité" est celle du G.722 B, voire inférieure à celle-ci, en fonction du caractère de la voix du locuteur.
11. L'article de Paulus et Schnitzler [C-12] décrit un codeur de type SB-ACELP à 16 kbits/s. La bande inférieure (0-6 kHz) est encodée par un codeur de type ACELP. La bande supérieure (6-7 kHz) est simplement représentée par un bruit blanc dont on ajuste l'énergie toutes les 2.5 ms. La "qualité" est celle du G.722 C. Cet article est très intéressant puisque le principe d'encodage qu'il décrit est repris par le nouveau standard WB-AMR.
12. L'article de Ubale et Gersho [C-13] décrit un codeur de type SB-CELP à 16 kbits/s. L'enveloppe spectrale du signal, ainsi que sa périodicité, sont calculées à partir du signal en une seule bande de fréquences. Ce codeur utilise la combinaison de deux dictionnaires d'excitations innovatrices. Ces dictionnaires ont été pré-entraînés ("off-line") sur la base d'excitations séparées en deux sous-bandes de fréquences. Chaque dictionnaire correspond à l'une de ces bandes. L'excitation est encodée sans séparation en sous-bandes, en combinant les excitations extraites de chacun des dictionnaires. L'extraction se fait d'abord dans le dictionnaire correspondant aux hautes fréquences. Grâce au pré-entraînement "off-line", aucun filtrage en sous-bandes n'est nécessaire en temps réel et aucun délai dû à celui-ci n'est introduit. La "qualité" est celle du G.722 C.

13. L'article de Schnitzler [C-14] est basé sur celui présenté ci-dessus au point 11. Il décrit un codeur de type SB-ACELP fonctionnant à 13 kbits/s. La bande inférieure (0-6 kHz) est encodée par un codeur de type ACELP. La bande supérieure (6-7 kHz) n'est pas encodée, mais générée par une synthèse des hautes fréquences (HFR : High-Frequency-Resynthesis). La synthèse se base sur les caractéristiques de la bande inférieure. La "qualité" est celle du G.722 C.
14. L'article de Combescure et *al.* [C-15] décrit un algorithme pour l'encodage de la parole et du signal audio. Le codeur fonctionne à 16, 24 et 32 kbits/s. Il combine un codeur par transformée adaptative (ATC Adaptive Transform Codec) et un codeur de type SB-CELP. Ce codeur s'appelle ATCELP. Le codeur SB-CELP traite le signal de parole et le codeur ATC (32 kbits/s) le signal audio. Le codeur de parole sépare le signal en deux bandes de fréquences. La bande inférieure (0-5 kHz) est encodée par un algorithme de type ACELP. La bande supérieure n'est pas traitée à 16 kbits/s. A 24 kbits/s elle est encodée par un algorithme de type CELP. Pour les débits de 16 et 24 kbits/s, la "qualité" est respectivement celle du G.722 C et B.
15. L'article de Black et Kondoz [C-16] propose un codeur de type SB-CELP à 16 kbits/s et à faible délai de traitement. Sur la bande inférieure (0-4 kHz), l'analyse LPC est faite de façon "backward", basée sur le signal quantifié de la trame passée. Sur la bande supérieure (4-8 kHz), l'analyse LPC est réalisée de façon ordinaire. Les trames de signal traitées sont très courtes afin de réduire le délai de traitement du codeur. La "qualité" est celle du G.722 C.

C.1.3 Codeur par transformée

16. L'article de Moriya et *al.* [C-17] décrit un codeur à 16 kbits/s, appelé TwinVQ (Transform domain Weighted Interleave Vector Quantization), basé sur une quantification vectorielle entrelacée et pondérée, dans le domaine fréquentiel. L'extraction des paramètres LPC et une transformée discrète en cosinus modifiée (MDCT : Modified Discrete Cosine Transform) sont réalisées en parallèle sur le signal. Les LSF sont quantifiés. Les coefficients MDCT sont aplanis dans le domaine des fréquences par le filtre d'analyse LPC et la périodicité du résidu de prédiction LPC est extraite puis encodée. Puis la structure fine de la MDCT lissée par analyse LPC, et exempte de composante périodique, est normalisée selon l'échelle des Barks. La quantification de cette structure fine est effectuée par quantification vectorielle entrelacée et pondérée. La

"qualité" est celle du G.722 C. Ici, le délai de traitement est trop grand pour une application en temps réel.

C.1.4 Codeurs mixtes

17. L'article de Lefebvre et *al.* [C-18] décrit un codeur similaire à un codeur CELP, où l'excitation innovatrice est encodée par transformée dans le domaine des fréquences, sans utiliser l'analyse par synthèse. Un tel codeur est appelé TCX-CELP (Transform Codec Excitation CELP). Il se prête à l'encodage de la parole et de la musique. Pour le codage de la parole à 16 kbits/s, la qualité est dite "très bonne".
18. L'article de Xie et Adoul [C-19] propose de reprendre le codeur TCX-CELP décrit par Lefebvre et *al.* [C-18] et d'encoder l'excitation innovatrice dans le domaine des fréquences à l'aide de dictionnaires de quantification vectorielle algébriques et "emboîtés", appelés EAVQ (embedded algebraic vector quantizers). Ce codeur fonctionne à 16 kbits/s. La "qualité" est dite "plus que très bonne".
19. L'article de Salami et *al.* [C-20] décrit un algorithme de type ACELP à 16 kbits/s, qui passe en mode ACELP/TCX pour fonctionner à 24 kbits/s. Cet algorithme utilise une analyse LPC hybride, qui réalise une prédiction "forward" (ordinaire) ou "forward/backward". Les coefficients LPC sont encodés sous forme de ISP. Dans le mode ACELP/TCX, l'excitation innovatrice est soit encodée selon une structure algébrique, soit selon une structure de codage par transformée TCX. A 16 kbits/s et respectivement à 24 kbits/s, la "qualité" est pratiquement celle du G.722 C et celle du G.722 B.
20. L'article de Chen [C-21] décrit un codeur utilisant la MDCT pour encoder le résidu de prédiction linéaire à court- et long-terme. Ce résidu est transformé et séparé en douze bandes de fréquences, puis quantifié. Le codeur fonctionne à un débit variant de 16 à 32 kbits/s. La "qualité" est dite "haute".
21. L'article de Schnitzler et *al.* [C-22] décrit un codeur basé sur l'analyse par synthèse, où l'analyse LPC est réalisée par prédiction "forward" ou "forward/backward". L'excitation innovatrice est encodée soit dans le domaine temporel en utilisant un dictionnaire algébrique, soit dans le domaine fréquentiel en réalisant une transformée de Fourier discrète et en utilisant une quantification scalaire. Ce codeur fonctionne à 15.5 et à 20 kbits/s. Les auteurs ne donnent aucune comparaison qualitative par rapport un autre codeur en bande élargie.

C.1.5 Autre codeur

22. L'article de Abreu et Docampo [C-23] décrit un codeur similaire à un codeur de type CELP, mais où l'excitation innovatrice s'obtient par déconvolution multi-impulsionnelle. Ce codeur est appelé MP-Dec (MultiPulse-Deconvolution). Il utilise un algorithme de déconvolution pointu, pour obtenir l'excitation peuplée de rares valeurs non-nulles. Ce codeur fonctionne à des débits variant de 16 à 33 kbits/s. A 33 kbits/s, la "qualité" est celle du G.722 A, alors qu'à 16 kbits/s, elle est inférieure à celle du G.722 C.

C.2 Catégories et sous-catégories des codeurs de l'état de l'art de 2000 à 2002

L'état de l'art de 2000 à 2002 pour le codage de la parole en bande élargie, discuté à la Section 4.5, est présenté ici par catégories et sous-catégories. A l'intérieur d'une même catégorie, les articles sont classés selon leur ordre chronologique de parution. Le nouveau standard WB-AMR de l'ETSI est décrit à la Section 4.5.

C.2.1 Codeur de type CELP

23. L'article de Pujatle et Moreno [C-24] décrit deux codeurs de type ACELP fonctionnant à 16 kbits/s, dont l'excitation innovatrice est codée en utilisant deux dictionnaires algébriques et donc deux gains. L'un des codeurs filtre le vecteur-cible passe-haut, pour la recherche dans le second dictionnaire algébrique. Ce codeur porte ainsi une attention particulière aux hautes fréquences de l'excitation. La "qualité" est celle du G.722 C et respectivement celle du G.722 B.

C.2.2 Codeur de type SB-CELP

24. Les articles de Erdmann et *al.* [C-25] et [C-26] décrivent un codeur de type SB-ACELP, fonctionnant à un débit variable, allant de 9.1 à 24 kbits/s. La bande inférieure (0-6 kHz) est encodée à l'aide d'un algorithme de type ACELP. Cet algorithme n'utilise le dictionnaire adaptatif que si la parole a un caractère voisé. Pour les bas débits, la bande supérieure (6-7 kHz) est encodée à l'aide d'un algorithme dit "d'expansion de la bande de fréquences", avec un débit de 300 bits/s. Toutefois, si le débit utilisé est grand, la bande supérieure est encodée par ADPCM en utilisant 4 kbits/s. Ce codeur est basé sur les codeurs

proposés par Paulus et Schnitzler en [C-12] et [C-14]. Il a été l'un des candidats pour le standard WB-AMR de l'ETSI.

C.2.3 Codeurs en sous-bandes, où la bande de fréquences inférieure est encodée à l'aide d'un algorithme pour la bande étroite

25. L'article de Taori et *al.* [C-27] décrit un codeur qui traite le signal original, $\mathbf{s_o}$, en deux bandes de fréquences. Ce codeur est illustré à la Figure 8.1. Un filtrage passe-bas (0-4 kHz) du signal original est d'abord effectué. D'une part, le signal ainsi filtré est soustrait du signal original retardé : on obtient le signal supérieur, $\mathbf{s_s}$. D'autre part le signal filtré est décimé par un facteur 2, puis codé à l'aide d'un codeur existant pour la bande étroite (codeur de base). Le signal supérieur est encodé par une technique appelée "injection de hautes fréquences" (Hi-BIN : High Band Injection). Cette technique modélise le signal par un bruit dont l'enveloppe spectrale est mise en forme et dont le niveau d'énergie est adapté. L'enveloppe spectrale est obtenue par analyse LPC. Si le codeur de base est le GSM-EFR [C-28], alors ce codeur fonctionne à 16 kbits/s. Dans ce cas, la "qualité" est légèrement inférieure à celle du G.722 B.
26. L'article de Koishida et *al.* [C-29] décrit un codeur qui fonctionne à 16 kbits/s et qui utilise deux algorithmes d'encodage : un algorithme de base et un algorithme dit "d'amélioration". L'algorithme de base est celui du standard G.729. Il traite le signal original filtré passe-bas (0-4 kHz) et décimé par 2. L'algorithme d'amélioration est un codeur CELP qui encode le signal original en utilisant de l'information extraite du codeur de base. Cet algorithme utilise une prédiction linéaire (LPC) d'ordre 16. Il extrait les excitations en tenant compte d'une partie du signal encodé par le codeur de base. L'article propose quatre structures différentes pour le codeur d'amélioration. La "qualité" est légèrement supérieure à celle du MPEG-4 CELP à 16 kbits/s.
27. L'article de McCree [C-30] décrit un codeur de type SB-ACELP fonctionnant à 14 kbits/s. La bande inférieure (0-4 kHz) est traitée par le codeur G.729 annexe E [C-31], fonctionnant à 11.8 kbits/s. Le codage de la bande supérieure (4-8 kHz) ne nécessite que 2.2 kbits/s. L'excitation de la bande supérieure est produite à l'aide d'un bruit, modulé par l'enveloppe temporelle du signal de la bande inférieure, compris entre 3 et 4 kHz. La "qualité" est celle du G.722 B.
28. L'article de McCree et *al*. [C-32] décrit un codeur de type SB-ACELP, à débit variable allant de 8 à 32 kbits/s. La bande de fréquences inférieure (0-4 kHz) est encodée en utilisant le standard NB-AMR [C-33]. Comme pour le codeur présenté au point 27, l'excitation de la bande supérieure

(4-8 kHz) est produite à l'aide d'un bruit, modulé par l'enveloppe temporelle du signal de la bande inférieure, compris entre 3 et 4 kHz. Cette modulation introduit une structure périodique dans le temps. Au niveau du décodeur, le bruit modulé est d'abord lissé afin d'éviter des rapides variations en amplitude, si le signal est bruité. Il est ensuite passé dans un filtre de synthèse LPC. Les coefficients de ce filtre sont extraits et encodés au niveau de l'encodeur. Pour les débits les plus élevés, le codage dans la bande inférieure est modifié en réduisant la durée de la sous-trame et en augmentant le nombre d'impulsions du dictionnaire algébrique. Si le codeur fonctionne à 16 kbits/s et si le signal d'entrée ne contient pas de bruit, alors la "qualité" est celle du G.722 A.

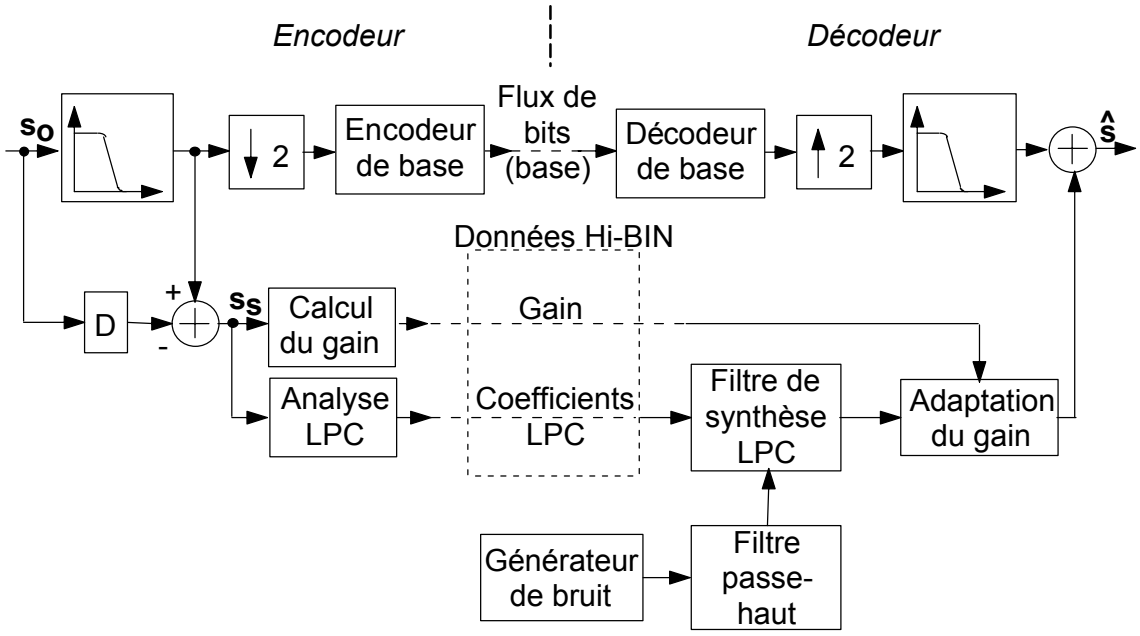

Figure 8.1 : Codeur N° 25 (Taori et *al.* [C-27]) incorporant une injection des hautes-fréquences. Le bloc "D" représente le retard appliqué au signal original.

C.2.4 Codeurs en sous-bandes, où la bande de fréquences inférieure est encodée à l'aide d'un algorithme pour la bande étroite et où la bande de fréquences supérieure est encodée par transformée

29. L'article de Lee et Bae [C-34] décrit un algorithme de type SB-ACELP qui fonctionne à 18.9 kbits/s. La bande inférieure (0-4 kHz) est encodée par le codeur GSM-EFR [C-28], basé sur le modèle ACELP. La bande supérieure (4-8 kHz) est encodée en utilisant la transformée discrète en ondelettes [C-35]. Les auteurs obtiennent une qualité de signal reconstruit similaire à celle du G.722 B.
30. L'article de Kim et *al.* [C-36] décrit un algorithme de type SB-ACELP. Ce codeur est illustré à la Figure 8.2. La bande de fréquences inférieure (0-4 kHz) est encodée à l'aide du G.729 annexe E [C-31]. Le signal

reconstruit de la bande inférieure est sur-échantillonné et soustrait au signal original pour donner le signal de la bande supérieure. Le signal de la bande supérieure est alors encodé en utilisant un banc de filtres "gammatone" avec un modèle auditif inversible. Les filtres "gammatone" simulent le mouvement de la membrane basilaire. Le signal de la bande supérieure est séparé en bandes de fréquences critiques. Ces bandes de fréquences sont approchées par des impulsions masquées, qui sont ensuite transformées dans le domaine des fréquences et quantifiées à l'aide de dictionnaires de quantification. A 18 kbits/s, la "qualité" est celle du G.722.1 à 24 kbits/s.

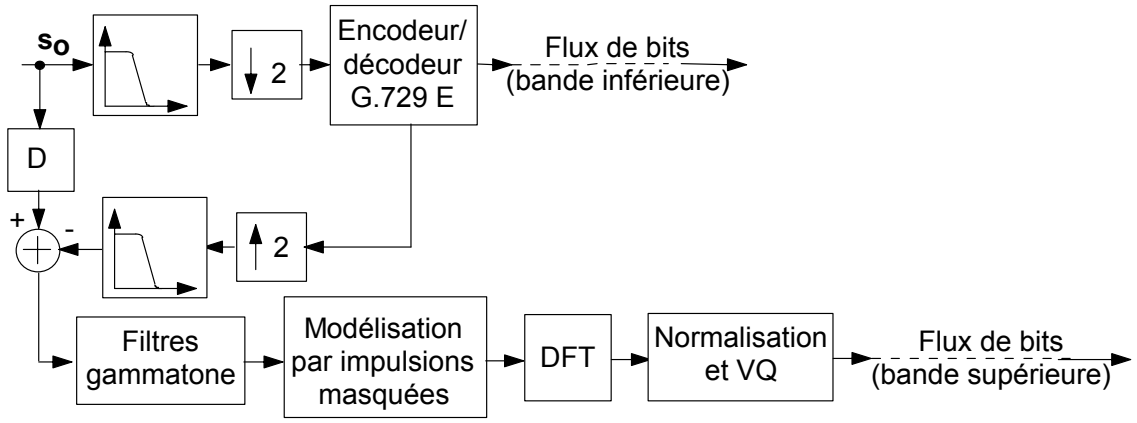

Figure 8.2 : Codeur N° 30 (Kim et *al.* [C-36]) de type SB-ACELP. Le bloc "D" représente le retard appliqué au signal original.

C.2.5 Codeur de type MELP

31. L'article de Lin et *al.* [C-37] décrit un codeur type MELP (Mixed Excitation Linear Prediction) [C-38]. Il s'agit d'un codeur d'analyse par synthèse, où l'excitation mixée est construite à l'aide d'un générateur de bruit et d'un générateur d'impulsions. Les impulsions peuvent avoir un caractère périodique ou apériodique. Si ce codeur fonctionne à 8.4 kbits/s, la "qualité" est celle du G.722 C.

C.2.6 Codeur par transformée

32. L'article de Aguilar et *al.* [C-39] décrit un codeur par transformées sinusoïdales. Ce codeur fonctionne aussi bien pour la bande étroite que pour la bande élargie. Pour la bande élargie, le débit est de 9.6 kbits/s et la séquence de bits relative à la bande étroite est encapsulée dans la séquence de bits relative à la bande élargie. Les auteurs ne donnent aucune comparaison qualitative par rapport à un autre codeur de parole en bande élargie.

33. L'article de Kokes et Gibson [C-40] décrit un codeur par transformée, où les blocs de base de la transformée sont adaptatifs et varient dans le temps. Les auteurs proposent d'utiliser une modulation non-uniforme des transformées bi-orthogonales (NMLBT : nonuniform modulated lapped biorthogonal transforms). Un mécanisme automatique permet de déterminer les combinaisons des bandes de fréquences pour produire une trame de parole à modulation NMLBT. Les auteurs ne donnent aucune comparaison qualitative par rapport à un autre codeur de parole en bande élargie.

C.3 Références

[C-1] C. Laflamme, J-P. Adoul, R. Salami, S. Morissette et P. Mabilleau, "16 kbps wideband speech coding technique based on algebraic CELP", dans *Proc. IEEE Int. conf. acoustic, speech, signal processing 1991, ICASSP'91,* Vol. 1, pp. 13-16, Toronto, Canada, Mai 1991.

[C-2] R. Salami, C. Laflamme, et J-P. Adoul, "Real-time implementation of a 9.6 kbit/s ACELP wideband speech coder", dans *Proc. IEEE Global telecommunications conference 1992, GLOBECOM'92*, pp. 447-451, Floride, USA, Déc. 1992.

[C-3] E. Harborg, J. Knudsen, A. Fuldseth et F. Johansen, "A real-time wideband CELP coder for a videophone application", dans *Proc. IEEE Int. conf. acoustic, speech, signal processing 1994, ICASSP'94,* Vol. 2, pp. 121-124, Adelaide, Australie, Avr. 1994.

[C-4] A. Fuldseth et E. Harborg, "Wideband speech coding at 16 kbit/s for a videophone application", dans *Speech communication*, Vol. 11, pp. 139-148, Elsevier science publishers B. V., 1992.

[C-5] C. McElroy, B. Murray et A. Fagan, "On improving wideband CELP speech coders", dans *Proceedings signal processing VII: Theories and applications, EUSIPCO'94,* Vol. 2, pp. 912-915, Edinburgh, Grande Bretagne, Août 1994.

[C-6] A. Black, I. Atkinson, A. Kondoz et B. Evans, "High quality 14.1 kb/s wideband speech coder", dans *4th European conference on speech communication and technology, EUROSPEECH'95,* pp. 45-48, Madrid, Espagne, Sept. 1995.

[C-7] S. Sasaki, A. Kataoka et T. Moriya, "Wideband CELP coder at 16 kbits/s with 10-ms frame", dans *4th European conference on speech communication and technology, EUROSPEECH'95,* pp. 41-44, Madrid, Espagne, Sept. 1995.

[C-8] M. Serizawa, A. Murashima et K. Ozawa, "A 16 kbit/s wideband CELP coder with a high-order backward predictor and its fast coefficient calculation", dans *Proc. IEEE Workshop on speech coding proceedings 1997,* pp. 107-108, Pocono Manor, USA, Sept. 1997.

[C-9] K. Koishida, G. Hirabayashi, K. Tokuda et T. Kobayashi, "A wideband CELP speech coder at 16 kbit/s based on mel-generalized cepstral analysis", dans *Proc. IEEE Int. conf. acoustic, speech, signal processing 1998, ICASSP'98,* Vol. 1, pp. 161-164, Seattle, USA, Mai 1998.

[C-10] G. Roy et P. Kabal, "Wideband CELP coding at 16 kbits/sec", dans *Proc. IEEE Int. conf. acoustic, speech, signal processing 1991, ICASSP'91,* Vol. 1, pp. 17-20, Toronto, Canada, 1991.

[C-11] C. McElroy, B. Murray et A. Fagan, "Wideband speech coding in 7.2 kb/s", dans *Proc. IEEE Int. conf. acoustic, speech, signal processing 1993, ICASSP'93,* Vol. 2, pp. 620-623, Minneapolis, USA, Avr. 1993.

[C-12] J. Paulus et J. Schnitzler, "16 kbit/s wideband speech coding based on unequal subbands", dans *Proc. IEEE Int. conf. acoustic, speech, signal processing 1996, ICASSP'96,* Vol. 1, pp. 255-258, Atlanta, USA, Mai 1996.

[C-13] A. Ubale et A. Gersho, "A multi-band CELP wideband speech coder", dans *Proc. IEEE Int. conf. acoustic, speech, signal processing 1997, ICASSP'97,* Vol. 2, pp. 1367-1370, Munich, Allemagne, Avr. 1997.

[C-14] J. Schnitzler, "A 13.0 kbit/s wideband speech coder based on SB-ACELP", dans *Proc. IEEE Int. conf. acoustic, speech, signal processing 1998, ICASSP'98,* Vol. 1, pp. 157-160, Seattle, USA, Mai 1998.

[C-15] P. Combescure, J. Schnitzler, K. Fischer, R. Kirchherr, C. Lamblin, A. Le Guyader, D. Massaloux, C. Quinquis, J. Stegmann, et P. Vary, "A 16, 24, 32 kbit/s wideband speech codec based on ATCELP", dans *Proc. IEEE Int. conf. acoustic, speech, signal processing 1999, ICASSP'99,* Vol 1, pp. 5-8, Phoenix, USA, Mai 1999.

[C-16] A. Black et A. Kondoz, "High quality low delay wideband speech coding at 16 kb/s", Chapter 17, dans *Insights into Mobile multimedia communications,* pp. 271-283, édité par D. Bull, C. Canagarjah et A. Nix, Academic Press, London, 1999.

[C-17] T. Moriya, N. Iwakami, A. Jin, K. Ikeda et S. Miki, "A design of transform coder for both speech and audio signals at 1 bit/sample", dans *Proc. IEEE Int. conf. acoustic, speech, signal processing 1997, ICASSP'97,* Vol. 2, pp. 1371-1374, Munich, Allemagne, Avr. 1997.

[C-18] R. Lefebvre, R. Salami, C. Laflamme, et J-P. Adoul, "High quality coding of wideband audio signals using transform coded excitation (TCX)", dans *Proc. IEEE Int. conf. acoustic, speech, signal processing 1994, ICASSP'94,* Vol. 1, pp. 193-196, Adelaide, Australie, Avr. 1994.

[C-19] M. Xie et J.-P. Adoul, "Embedded algebraic vector quantizers (EAVQ) with application to wideband speech coding", dans *Proc. IEEE Int. conf. acoustic, speech, signal processing 1996, ICASSP'96,* Vol. 1, pp. 240-243, Atlanta, USA, Mai 1996.

[C-20] R. Salami, R. Lefebvre et C. Laflamme, "A wideband codec at 16/24 kbit/s with 10 ms frame", dans *IEEE Workshop on speech coding proceedings 1997,* pp. 103-104, Pocono Manor, USA, Sept. 1997.

[C-21] J. Chen, "A candidate for the ITU-T's new wideband speech coding standard", dans *Proc. IEEE Int. conf. acoustic, speech, signal processing 1997, ICASSP'97,* Vol. 2, pp. 1359-1362, Munich, Allemagne, Avr. 1997.

[C-22] J. Schnitzler, J. Eggers, C. Erdmann et P. Vary, "Wideband speech coding using forward/backward adaptive prediction with mixed time/frequency domain excitation", dans *IEEE Workshop on speech coding proceedings 1999,* pp. 4-6, Provoo, Finlande, Juin 1999.

[C-23] V. Abreu et D. Docampo, "A multiple-deconvolution codec for wideband speech", dans *4th European conference on speech communication and technology, EUROSPEECH'95,* pp. 49-52, Madrid, Espagne, Sept. 1995.

[C-24] S. Pujalte et A. Moreno, "Wideband ACELP at 16 kbit/s with multi-band excitation", dans *Proc. European conference on speech communication and technology, EUROSPEECH 2001,* pp. 2001-2004, Aalborg, Danemark, 2001.

[C-25] C. Erdmann, P. Vary, K. Fischer, J. Stegmann, C. Quinquis, D. Massaloux et B. Kövesi, "An adaptive multi-rate wideband speech codec with adaptive gain re-quantization", dans *IEEE Workshop on speech coding proceedings 2000,* pp. 145-147, Delavan, USA, Sept. 2000.

[C-26] C. Erdmann, P. Vary, K. Fischer, W. Xu, M. Marke, T. Fingscheidt, I. Varga, M. Kaindl, C. Quinquis, B. Kövesi et D. Massaloux, "A candidate proposal for a 3 GPP adaptive multi-rate wideband speech codec", dans *Proc. IEEE Int. conf. acoustic, speech, signal processing 2000, ICASSP'2001,* Vol. 2, pp. 757-760, Salt Lake City, USA, Mai 2001.

[C-27] R. Taori, R. Sluijter et A. Gerrits, "Hi-Bin : an alternative approach to wideband speech coding", dans *Proc. IEEE Int. conf. acoustic, speech, signal processing 2000, ICASSP'2000,* Vol. 2, pp. 1157-1160, Istanbul, Turquie, Juin 2000.

[C-28] L. Hanzo, F. Somerville et J. Woodard, "Standard forward-adaptive CELP codecs", Chapter 7, dans *Voice compression and communications*, pp. 207-278, IEEE Series on Digital & Mobile Communication, John Wiley & Sons, Inc., Publication, NY, USA, 2001.

[C-29] K. Koishida, V. Cuperman et A. Gersho, "A 16-kbits/s bandwitdth scalable audio coder based on the G.729 standard", dans *Proc. IEEE Int. conf. acoustic, speech, signal processing 2000, ICASSP'2000,* Vol. 2, pp. 1149-1152, Istanbul, Turquie, Juin 2000.

[C-30] A. McCree, "A 14 kb/s wideband speech coder with a parametric highband model", dans *Proc. IEEE Int. conf. acoustic, speech, signal processing 2000, ICASSP'2000,* Vol. 2, pp. 1153-1156, Istanbul, Turquie, Juin 2000.

[C-31] UIT-T Recommendation G.729 – Annexe E, "Codage de la parole à 8 kbit/s en utilisant la prédiction linéaire à excitation par séquences codées à structure algébrique conjuguée (CS-ACELP), Annexe E: Algorithme de codage vocal CS-ACELP à 11,8 kbit/s", Sept. 1998.

[C-32] A. McCree, T. Unno, A. Anandakumar, A. Bernard et E. Pakosoy, "An embedded adaptive multi-rate wideband speech coder", dans *Proc. IEEE Int. conf. acoustic, speech, signal processing 2001, ICASSP'2001,* Vol. 4, pp. 2613-2616, Salt Lake City, USA, Mai 2001.

[C-33] 3GPP TS 26.090 V3.1.0 (1999–12) document, dans ftp://ftp.3gpp.org/Specs/2000-09/R1999/26_series/ (14 Nov. 2001).

[C-34] S. Lee et K. Bae, "Wideband speech coding algorithm with application of discrete wavelet transform to upper band", dans *Proc. European conference on speech communication and technology, EUROSPEECH 2001,* pp. 2005-2008, Aalborg, Danemark, Sept. 2001.

[C-35] O. Rioul et M. Vetterli, "Wavelet and signal processing", dans *IEEE Signal processing magazine*, pp. 14-38, Oct. 1991.

[C-36] K. Kim, S. Jung, Y. Park et D. Youn, "A new bandwidth scalable wideband speech/audio coder", dans *Proc. IEEE Int. conf. acoustic, speech, signal processing 2002, ICASSP'2002,* Vol. 1, pp. 657-660, Orlando, USA, Mai 2002.

[C-37] W. Lin, S. Koh et X. Lin, "Mixed excitation linear prediction coding of wideband speech at 8 kbps", dans *Proc. IEEE Int. conf. acoustic, speech, signal processing 2000, ICASSP'2000,* Vol. 2, pp. 1137-1140, Istanbul, Turquie, Juin 2000.

[C-38] A. McCree et T. Barnwell, "A mixed excitation LPC vocoder model for low bit rate speech coding", dans *IEEE Trans. on speech and audio processing,* Vol. 3, N° 4, pp. 242-250, Juil. 1995.

[C-39] G. Aguilar, J. Chen, R. Dunn, R. McAulay, X. Sun, W. Wang et R. Zopf, "An embedded sinusoidal transform codec with measured phases and sampling rate scalability", dans *Proc. IEEE Int. conf. acoustic, speech, signal processing 2000, ICASSP'2000,* Vol. 2, pp. 1141-1144, Istanbul, Turquie, Juin 2000.

[C-40] M. Kokes et J. Gibson, "Frequency selectivity via the SpEnt Methodology for wideband speech compression", dans *Proc. IEEE Int. conf. acoustic, speech, signal processing 2001, ICASSP'2001,* (CD-ROM), Salt Lake City, USA, Mai 2001.

Annexe D
Résultats des tests auditifs

Codeur	Séquence	Evaluation selon l'auditeur															Moyenne
		A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	
P-MRWB-ACELP, 14.3 kbits/s	T01	4	4	4	5	4	3	5	4	4	4	4	4	4	3	4	4.0
	T02	4	4	5	4	5	4	5	4	4	4	4	3	3	3	4	4.0
	T03	3	3	4	3	4	3	4	4	3	3.5	3	3	1	1	4	3.1
	T04	5	4	5	4	4	4	4	4	4	5	4	3	3	4	4	4.1
	T05	4	4	5	4	4	3	5	3	2	3	3	3	1	1	4	3.3
	T06	4	4	4	5	5	4	4	4	4	4.5	4	4	3	2	4	4.0
	T07	3	3	3	4	3	3	4	4	3	3	3	3	2	2	4	3.1
	T08	4	3	5	3	5	3	3	4	3	4.5	4	4	3	2	4	3.6
	T09	3	3	4	5	3	3	4	3	2	3	3	4	2	2	5	3.3
	T10	4	3	4	3	5	3	3	4	2	3	3	3	2	2	4	3.2
	T11	3	3	3	4	3	3	4	4	2	3	3	2	2	2	4	3.0
	T12	4	3	4	5	5	3	4	3	3	4	3	3	1	2	5	3.5
P-MRWB-ACELP, 18.5 kbits/s	T01	5	4	5	5	4	4	4	4	4	4.5	4	4	4	3	5	4.2
	T02	5	4	5	5	3	4	4	4	5	3.5	4	4	3	1	4	3.9
	T03	4	4	4	5	4	4	3	4	3	3	3	4	2	2	4	3.5
	T04	5	4	5	5	5	5	4	5	4	4	4	5	4	3	5	4.5
	T05	5	4	4	4	4	4	4	3	3	3	4	2	1	2	4	3.4
	T06	5	4	4	5	4	4	5	4	4	3	4	5	4	2	5	4.1
	T07	4	4	4	5	3	3	4	4	5	3.5	3	4	4	3	5	3.9
	T08	4	3	5	4	5	4	4	4	4	4	4	4	2	3	4	3.9
	T09	3	3	4	4	4	4	4	4	3	3.5	3	3	2	3	4	3.4
	T10	3	3	5	4	5	3	4	3	4	4	3	3	2	3	4	3.5
	T11	3	4	4	4	4	3	4	4	3	3	4	3	3	3	4	3.5
	T12	5	4	5	5	5	3	5	4	5	3.5	4	4	3	4	4	4.2
P-MRWB-ACELP, 21.5 kbits/s	T01	4	5	5	5	5	4	5	4	5	4	4	4	5	4	4	4.5
	T02	5	5	5	4	5	4	4	5	3	3	4	4	3	3	4	4.1
	T03	4	4	4	4	4	5	3	4	3	3.5	5	3	3	3	4	3.8
	T04	5	4	5	5	5	5	5	5	5	4.5	3	5	5	3	5	4.6
	T05	4	3	5	4	5	4	5	4	3	3.5	4	5	3	3	4	4.0
	T06	5	4	5	4	5	4	5	5	4	4	4	5	4	4	5	4.5
	T07	4	4	4	5	4	3	5	4	5	3.5	4	4	3	4	5	4.1
	T08	4	4	5	4	5	4	4	4	4	4	4	4	4	3	4	4.1
	T09	3	4	5	4	4	4	4	4	4	3.5	4	4	3	3	4	3.8
	T10	5	4	5	5	5	4	4	4	4	4	3	4	3	4	4	4.1
	T11	3	3	5	4	4	4	4	4	4	4	4	3	3	3	4	3.7
	T12	5	4	5	4	5	4	4	4	4	5	3	4	4	4	5	4.3

Tableau D-1 : Evaluation des séquences de test pour les différents modes du codeur P-MRWB-ACELP.

Codeur	Séquence	Evaluation selon l'auditeur															Moyenne
		A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	
G.722 48 kbits/s	T01	5	3	5	5	5	4	4	4	4	4.5	4	3	2	3	5	4.0
	T02	5	4	4	4	4	4	4	4	5	4	4	4	4	3	4	4.1
	T03	5	3	3	5	5	4	4	4	4	4	4	4	1	2	5	3.8
	T04	4	3	3	4	4	4	4	4	3	4	3	3	2	3	4	3.5
	T05	5	4	4	4	4	4	4	4	4	3	4	3	4	3	4	3.9
	T06	4	3	4	5	4	4	4	4	4	4	4	4	2	3	4	3.8
	T07	3	3	3	4	4	3	3	4	4	3.5	3	3	3	4	4	3.4
	T08	4	3	3	4	4	4	4	4	3	4	4	3	3	3	4	3.6
	T09	3	2	3	3	4	3	5	4	3	4	3	3	3	3	4	3.3
	T10	4	4	3	4	4	4	4	4	3	4	4	3	3	2	4	3.6
	T11	3	3	3	4	4	3	3	4	3	3.5	4	3	3	2	4	3.3
	T12	4	5	3	5	5	4	5	5	4	4	4	4	4	4	5	4.3
G.722, 56 kbits/s	T01	4	4	4	4	4	4	5	5	5	4.5	4	3	4	3	4	4.1
	T02	5	4	4	5	4	5	5	5	5	4.5	5	5	5	5	5	4.8
	T03	4	5	3	5	5	5	4	5	4	4.5	4	5	4	4	4	4.4
	T04	4	4	4	5	4	3	4	4	4	4	4	4	5	3	4	4.0
	T05	4	4	4	4	4	4	4	5	4	4.5	4	4	5	4	4	4.2
	T06	5	4	4	5	4	4	5	4	4	4.5	4	5	4	3	4	4.2
	T07	5	4	5	5	4	4	3	5	5	3.5	4	5	4	5	4	4.4
	T08	4	4	5	4	4	4	5	4	4	3.0	4	4	4	3	4	4.0
	T09	4	4	5	4	4	4	4	4	4	4	4	5	4	4	4	4.1
	T10	4	4	5	5	4	4	4	5	4	4	4	3	4	3	4	4.1
	T11	5	4	4	4	5	4	4	5	4	4.5	4	4	4	3	4	4.2
	T12	4	3	5	5	4	4	4	4	4	4.5	4	4	4	4	5	4.2
G.722 64 kbits/s	T01	5	4	5	5	5	3	4	5	5	3.5	4	3	5	3	5	4.3
	T02	5	4	4	5	5	5	5	5	5	4.5	5	4	5	4	5	4.7
	T03	3	4	4	5	5	5	4	5	5	5	5	5	5	5	4	4.6
	T04	5	4	5	5	4	5	5	4	4	4.5	4	4	4	4	4	4.5
	T05	5	5	4	5	5	5	4	5	4	4.5	4	4	5	4	5	4.6
	T06	5	4	4	5	3	4	4	5	5	4.5	4	4	4	4	4	4.2
	T07	5	4	4	4	4	4	4	4	5	4.5	4	4	4	4	5	4.2
	T08	4	4	5	5	5	4	5	5	4	3.5	4	3	4	3	5	4.2
	T09	5	4	5	5	5	4	4	5	5	4	4	5	5	4	5	4.6
	T10	5	4	4	5	5	4	5	5	5	4	4	3	5	4	5	4.5
	T11	5	4	5	5	5	4	5	5	5	4	4	4	5	3	5	4.5
	T12	5	4	4	5	5	5	4	5	5	4	4	4	5	4	4	4.5

Tableau D-2 : Evaluation des séquences de test pour les différents modes du codeur G.722.

Codeur	Séquence	Evaluation selon l'auditeur															Moyenne
		A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	
WB-AMR 14.25	T01	5	3	4	4	4	4	5	4	5	4.5	4	5	4	3	5	4.2
	T02	4	3	5	4	5	4	5	5	5	3.5	4	4	4	2	5	4.2
	T03	3	4	3	3	5	4	3	4	4	3	4	3	2	2	4	3.4
	T04	5	4	5	5	5	4	5	5	5	4	4	5	4	4	5	4.6
	T05	4	3	4	4	4	4	4	4	4	3.5	3	3	4	3	4	3.7
	T06	5	4	4	4	5	5	5	3	4	5	4	4	4	3	4	4.2
	T07	4	4	5	5	4	4	4	4	4	3	4	5	4	2	5	4.1
	T08	5	4	4	4	4	4	4	4	4	5	5	4	3	3	4	4.1
	T09	3	4	4	4	5	4	4	5	4	3.5	3	4	5	1	4	3.8
	T10	4	4	5	5	4	4	5	4	4	4	4	4	4	2	4	4.1
	T11	3	4	4	3	4	3	4	4	4	3.5	4	3	3	2	4	3.5
	T12	4	4	5	5	4	4	4	4	5	3.5	4	4	5	3	4	4.2
WB-AMR 18.25	T01	4	5	5	5	5	5	5	4	5	5	4	5	5	4	5	4.7
	T02	4	4	4	4	5	4	5	5	5	4	4	4	5	3	5	4.3
	T03	4	3	4	5	4	4	3	4	5	4	4	4	4	2	4	3.9
	T04	5	4	4	5	5	5	4	5	5	5	4	4	4	5	5	4.6
	T05	4	4	5	5	4	4	4	4	4	3.5	4	4	5	3	4	4.1
	T06	5	4	5	5	5	5	5	5	5	4.5	4	5	5	4	5	4.8
	T07	4	4	4	5	5	5	3	5	4	4	4	5	4	3	4	4.2
	T08	5	4	5	4	5	4	5	5	5	4	4	4	3	3	4	4.3
	T09	3	4	4	4	4	4	5	5	4	4	4	3	4	2	4	3.9
	T10	5	4	4	5	5	4	5	4	5	3.5	4	3	4	4	4	4.2
	T11	5	4	5	4	4	3	5	4	4	3.5	3	5	3	3	4	4.0
	T12	4	5	5	5	5	4	5	5	5	5	5	5	5	4	5	4.8
WB-AMR 23.05	T01	5	5	4	5	5	5	4	4	5	5	4	5	5	5	5	4.7
	T02	5	5	4	5	5	5	5	4	5	5	4	4	5	3	5	4.6
	T03	3	5	4	5	5	5	3	4	4	3.5	4	4	5	3	4	4.1
	T04	5	5	5	5	5	5	4	5	5	5	5	5	5	4	4	4.8
	T05	4	4	5	5	5	3	5	4	5	4.5	4	5	5	4	5	4.5
	T06	5	4	5	3	5	5	5	4	4	5	5	5	4	5	4	4.5
	T07	4	5	4	5	5	4	5	4	5	4	4	5	4	3	5	4.4
	T08	5	4	5	5	5	4	4	5	5	4	4	4	5	3	4	4.4
	T09	4	4	5	4	4	4	4	4	5	3	4	4	5	3	4	4.1
	T10	5	5	5	5	5	5	5	5	5	4.5	4	5	5	4	5	4.8
	T11	5	4	5	5	4	4	5	5	4	3.5	4	4	4	3	5	4.3
	T12	5	5	5	5	5	5	5	4	5	4.5	4	4	5	4	4	4.6

Tableau D-3 : Evaluation des séquences de test pour les différents modes testés, du codeur WB-AMR.

Annexe E
Glossaire

3G :	Troisième Génération de téléphonie mobile
ACELP :	Algebraic-Code-Excited Linear-Prediction
ACO :	Inverse de la fonction "cosinus"
AD :	Additions
ADPCM :	Adaptive-Differential Pulse-Code Modulation
AMR :	Adaptive Multi-Rate
AR :	Auto-Regressive
ATC :	Adaptive Transform Codec
ATCELP :	Adaptive Transform CELP
BI :	Bande de fréquences Inférieure
BS :	Bande de fréquences Supérieure
CELP :	Code-Excited Linear-Prediction
CNG :	Confort Noise Generation
CO :	Fonction "cosinus"
CS-ACELP :	Conjugate-Structure Algebraic Code Excited Linear-Prediction
DE :	Distance Euclidienne (non pondérée)
DEP :	Distance Euclidienne Pondérée
DI :	Divisions
DMOS :	Degradation Mean Opinion Score
DSP :	Digital Signal Processing
DTX :	Discontinous Transmission
ETSI :	European Telecommunications Standards Institute
GPRS :	General Packet Radio Service
GSM EFR :	GSM Enhanced Full-Rate
GSM FR :	GSM Full-Rate
GSM HR :	GSM Half-Rate

HF :	Haute Fréquence
Hi-BIN :	High-Band Injection
I :	Incrémentations de pointeur
IS :	Inverse Sine
ISP :	Immittance Spectrum Pairs
ISPP :	Interleaved Single-Pulse Permutation
ITU (ou UIT) :	International Telecommunications Union
ITU-T :	ITU – Telecommunications Standards Sector
IV :	Inversions de signe
L10 :	Fonctions "logarithme en base 10"
LAR :	Log Area Ratio
LBG :	Linde-Buzo-Gray
LMS :	Least-Mean-Square
LPAS :	Linear Predictive Analysis-by-Synthesis
LPC :	Linear Prediction Coefficients
LSP :	Line Spectrum Pairs
LTP :	Long-Term Prediction
MA :	Moving Average
MDCT :	Modified Discrete Cosine Transform
MELP :	Mixed Excitation Linear Prediction
MEM :	Nombre de valeurs à stocker en mémoire
MGC :	Mel-Generalized Cepstral analysis
MLT :	Modulated Lapped Transform
MP-CELP :	Multi-Pulse based CELP
MP-Dec :	MultiPulse-Deconvolution
MPEG4 :	Moving Picture Experts Group – 4
MSVQ :	Multi-Stage Vector Quantization
MU :	Multiplications
NB-AMR :	Narrow-Band Adaptive Multi-Rate
NU :	Non-Uniforme
NMLBT :	Nonuniform Modulated Lapped Biorthogonal Transforms
P10 :	Fonctions "puissance de 10"
P-MRWB-ACELP :	Proprietary Multi-Rate Wide-Band ACELP
PRELP :	Pulsed Residual Excited Linear-Prediction
$\overline{SD}$:	Moyenne de la distorsion spectrale
SR :	Fonctions "racine carrée"
RAM :	Random Access Memory
RC :	Reflexion Coefficients
ROM :	Read-Only Memory
RIF :	A Réponse Impulsionnelle Finie

RPE :	Regular-Pulse Excitation
SB-CELP :	Sub-Band CELP
S-MSVQ :	Split Multi-Stage Vector Quantization
SNR :	Signal-to-Noise Ratio
SNVQ :	Safety-Net Vector Quantization
SPL :	Sound Pressure Level
SVQ :	Split Vector Quantization
TCX-CELP :	Transform Codec Excitation CELP
TwinVQ :	Transform domain Weighted Interleave Vector Quantization
VAD :	Voice Activity Detector
VQ :	Vector Quantization
WB-AMR :	Wide-Band Adaptive Multi-Rate
wMOPS :	weighted Million Operations Per Second
WQ :	Weighted Quantization

Giuseppina Biundo Lotito

Laboratoire d'électronique et de traitement du signal ESPLAB,
Institut de microtechnique IMT, Université de Neuchâtel UNI-NE,
Rue A.-L. Breguet 2, CH-2000 Neuchâtel, Suisse.
giuseppina.biundo@unine.ch

Date et lieu de naissance :	6 mai 1973, 2000 Neuchâtel
Originaire de :	Cressier (NE), Castelbuono, Palerme (Italie)
Domaine de recherche actuel :	Codage du signal de parole
Langues :	Français, italien, anglais, allemand

Etudes :

1993-1998	Diplôme en Electronique-Physique, Faculté des Sciences de l'Université de Neuchâtel
1992-1993	Cours de secrétariat, Ecole Supérieure de Commerce de Neuchâtel
1988-1991	Baccalauréat et Maturité C, Gymnase Cantonal, Neuchâtel

Expérience professionnelle :

Dès 1998, assistante de recherche et collaboratrice scientifique au Laboratoire d'électronique et de traitement du signal ESPLAB.

Révision d'articles pour des conférences internationales, dont :

IEEE International symposium on signals, circuits, and systems, Iasi, Roumanie, SCS 2001 et 2003 (10-11 juillet, 2001 et 2003).

Contribution à l'organisation de *International COST 254 Workshop on intelligent communication technologies and applications, with emphasis on mobile communications*, Neuchâtel, Suisse, 5-7 mai 1999.

Enseignement :

Dès 1998, définition et supervision de cinq projets d'étudiants et de stagiaires au ESPLAB, IMT.

Coordination des travaux d'étudiants pour l'IMT UNI-NE, ainsi que pour le Laboratoire commun de microtechnique entre l'IMT et le STI - Section de microtechnique, de l'Ecole Polytechnique Fédérale de Lausanne EPFL.

De 1996 à 1998, assistante volontaire à l'Université de Neuchâtel pour le cours d'électronique générale du Professeur A. Shah, IMT.

Prix :

Prix Omega Etudiant, pour *Etude et implémentation des algorithmes de détection du silence dans les signaux de parole*, travail de diplôme, 1998.

Sociétés professionnelles :

Membre IEEE et ISCA.

Publications et brevets :

Voir Section 8.1.