N° D'ORDRE : 16/2010-M/EL

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE DES SCIENCES ET DE LA TECHNOLOGIE HOUARI BOUMEDIENE

FACULTE D'ELECTRONIQUE ET D'INFORMATIQUE

MEMOIRE

Présenté pour l'obtention du diplôme de MAGISTER

EN : ELECTRONIQUE

Spécialité: Communication parlée

Par: M. Farah Faris

Thème

Codage de la parole pour des transmissions très bas débit

Soutenu publiquement le: 30/05/2010 devant le jury composé de :

M. S. CHITROUB	**Professeur**	**à USTHB**	**Président**
M. B. BOUDRAA	**Maître de Conférences A**	**à USTHB**	**Directeur de Mémoire**
M. F. ALLILAT	**Maître de Conférences A**	**à USTHB**	**Examinateur**
M. M. DEBIECHE	**Maître de Conférences A**	**à USTHB**	**Examinateur**
Mme. L. FALEK	**Maître de Conférences A**	**à USTHB**	**Examinatrice**

Remerciements

Le travail présenté dans ce mémoire a été effectués au sein du laboratoire de la communication parlée et de traitement du signal (LCPTS) de la faculté d'électronique et d'informatique de l'USTHB.

J'adresse tout particulièrement ma reconnaissance à mon directeur de thèse : Mr. Bachir BOUDRAA. Il a su me faire profiter de ses nombreuses connaissances en traitement de la parole. Je tiens également à le remercier pour la confiance et la sympathie qu'il m'a témoignés, et aussi de la liberté d'action et d'autonomie dont j'ai bénéficié durant ce travail.

Je tiens à exprimer mes remerciements aux membres du jury, qui ont accepté d'évaluer mon travail.

Je remercie le Professeur Salim CHETROUB, d'avoir accepté de présider le jury de ce mémoire.

Je tiens particulièrement à remercier Mr. Farid ALILAT, Maître de Conférences à l'USTHB, pour l'honneur qu'il me fait en acceptant de participer à mon jury.

Je remercie aussi Mr. Mohamed DEBYECHE, Maître de Conférences à l'USTHB, pour avoir accepté d'être dans mon jury.

J'adresse mes vifs remerciements à Mme. Leila FALEK, Maître de Conférences à l'USTHB, pour l'honneur qu'elle me fait en acceptant de juger mon travail.

Je tiens également à faire part de toute ma gratitude et ma sympathie à tous les membres de notre laboratoire et mes collègues d'ODM, notamment Salim, Hakim et Chahinez.

Un spécial remerciement aux membres du G82, Hamza, Nacer, et Abdelhak. Merci pour votre amitié et pour tous les moments de détente et fous rires inoubliables passés ensemble.

J'adresse un grand merci à toute ma famille qui a toujours été présente lorsque j'en ai eu besoin, en particulier à mes parents, à mon frère et à mes sœurs.

Enfin merci à toutes les personnes que je n'ai pas citées ici et qui se reconnaîtront dans ces quelques lignes.

Dédicaces

Je dédie ce travail à mes chers parents qui tiennent une place immense dans mon cœur.

A une personne unique au monde, Hafid mon frère et à mes sœurs. Mon amour pour vous est sans limite.

A Hamza et à tous les membres de l'équipe EXP et ITO

Ames chèrs amis, Fateh, Riad, Khaled et a tous mes collègues de la post graduation.

A tous ceux qui me sont chèrs.

TABLE DES MATIERES

Chapitre II : CODAGE PREDICTIF LINEAIRE

Chapitre III : MISE EN ŒUVRE D'UN CODEUR MELP 1.2

Chapitre IV : RESULTATS ET EVALUATIONS

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABREVIATIONS

AC	*Auto-corrélation Coefficients*
ACELP	*Algebraic Code Exited Linear Prediction*
AMR	*Adaptive Multi-Rate*
AR	*Auto-Regressive*
CELP	*Code Exited Linear Prediction*
CS-ACELP	*Conjugate Structure Algebraic Code Exited Linear Prediction*
DoD	*Department of Defense*
DRT	*Diagnostic Rhyme Test*
DSP	*Digital Signal Processing*
EFR	*Enhanced Full Rate*
ETSI	*European Telecommunications Standards Institute*
FFT	*Fast Fourier transform*
FS1015	*Federal Standard 1015*
FS1016	*Federal Standard 1016*
FS1017	*Federal Standard 1017*
GSM	*Global System for Mobil communication*
JVQ	*Joint Vector Quantization*
LBG	*Linde Buzo Gray*
LD-CELP	*Low Delay Code Exited Linear Prediction*
LLR	*Log Likelihood Ratio*
LP	*Linear Prediction*
LPC	*Linear Predictive Coding*
LSF	*Line Spectral Frequency*
MBE	*Multi-Band Excitation*
MDF	*Magnitude Difference Function*
MELP	*Mixed Excitation Linear Prediction*

MFCC	*Mel-Frequency Cepstral Coefficients*
MNB	*Measuring Normalizing Block*
MOS	*Mean Opinion Score*
MP-MLQ	*Multi Pulse Maximum Likelihood Quantization*
MRT	*Modified Rhyme Test*
MSVQ	*Multi Stage Vector Quantization*
PAPE	*Phrases Arabes Phonétiquement Equilibrées*
PESQ	*Perceptual Evalution Speech Quality*
PLP	*Perceptual Linear Predictive*
PSQM	*Perceptual Speech Quality Mesure*
SQ	*Scalar Quantization*
VQ	*Vector Quantization*
FIR	*Finate Impulse Response*
RC	*Reflection Coefficients*
RCR	*Radio Center for Research*
SNR	*Signal to Noise Ratio*
SD	*Spectral Distortion*
STC	*Sinusoidal Transform Coder*
TIA	*Telecommunications Industry Association*
ITU	*International Telecommunications Union*
VBR	*Variable Bite Rate*
VLBR	*Very Low Bit Rate*
Vqdtool	*Vector Quantization Design Tool*
VSELP	*Vector Sum Exited Linear Prediction*
WI	*Waveform Interpolation*
WSS	*Weighted Spectral Slope*

INTRODUCTION GENERALE

Le développement de codeurs de parole de haute qualité, fonctionnant à un débit faible, a été motivé par le marché croissant des systèmes digitaux de télécommunications et d'enregistrement, où les applications les plus importantes sont les systèmes de radiocommunications avec les mobiles, les systèmes de communications par satellite, les systèmes de communications pour les multimédia, la téléphonie par internet et la visiophonie. Ce progrès a été rendu possible grâce au développement de processeurs DSP (Digital Signal Processing) et de circuits VLSI (Very Large Scale Integration) performants. En outre, des modèles mathématiques très proches du système phonatoire humain ont pu être mis au point et utilisés dans des algorithmes de codage de plus en plus complexes.

Tout système de codage de parole réalise un compromis parmi plusieurs contraintes. Idéalement, on voudrait disposer d'un système capable de représenter le signal de parole avec un débit très faible, produisant un signal synthétisé d'une qualité transparente et ceci même en présence de différentes formes de bruits. Ce codeur utiliserait en plus un algorithme de faible complexité et de faible exigence en mémoire. Pour les applications de télécommunications, il faudrait aussi maintenir le délai du codage très court et assurer une parfaite robustesse contre les erreurs de transmission. Pour la qualité de restitution du signal vocal, une dégradation limite du signal doit être évaluée lors du développement d'un système de codage. Une évaluation objective réalisée par des mesures de distorsions du signal ou de rapport signal à bruit restant insuffisante, les exigences et les opinions subjectives des utilisateurs potentiels doivent être prises en compte.

Dès la fin des années 60 apparaît la technique de Prédiction Linéaire (LPC : **L**inear **P**redictive **C**oding), cette dernière suppose que le signal de parole peut être considéré comme la sortie d'un filtre tout-pôle. Un codeur LPC sera particulièrement performant lorsqu'il utilise des techniques d'analyse par synthèse (LPAS : **L**inear **P**rediction **A**nalysis-by-**S**ynthesis). Dans un système de codage LPAS, le signal de parole à transmettre est codé en recherchant la meilleure excitation possible d'un filtre de synthèse, dont les coefficients sont déterminés par une analyse de Prédiction Linéaire. Le plus connu de ces codeurs paramétriques LPAS est le codeur de type CELP (**C**ode-**E**xcited **L**inear **P**rediction) développé en 1984 par les laboratoires Bell et conçu pour le codage de la parole à des débits inférieurs à 16 kbps. Pour le même model de production, une nouvelle technique de modélisation de l'excitation à été développée par McCree au milieu des années 90 dans son algorithme de codage connu sous le

nom MELP (**M**ixed-**E**xcitation **L**inear **P**rediction). C'est l'excitation mixte qui résulte d'un mixage fréquentiel d'une composante périodique et d'un bruit blanc. Ce mixage est contrôlé par une analyse de voisement multi-bandes.

Dans ce travail, on présente la mise au point sous MATLAB d'un codeur fonctionnant à très bas débit (1.2 kbps). Comme la plage de débit visée ne peut être atteinte qu'avec un codeur paramétrique, nous avons choisi l'algorithme de codage paramétrique MELP (**M**ixed **E**xcitation **L**inear **P**rediction), un algorithme procurant un rapport débit/qualité prometteur (dépassant celui de la LPC10). Ce codeur va être utilisé par l'équipe de codage du laboratoire LCPTS pour initier d'autres travaux de recherche, portant particulièrement la multi-description utilisée dans la voix IP pour remédier aux problèmes de pertes des paquets dans le réseau IP.

Le plan que nous avons adopté pour arriver à la réalisation de notre objectif est d'utiliser la prédiction linéaire à excitation mixte telle que décrite par le standard FS-MELP, tout en gardant la même segmentation (des trames d'analyse de 22.5ms) pour ne pas avoir des problèmes de non stationnarité. Afin de baisser d'avantage le débit nous avons opté pour l'utilisation d'une quantification par super-trames : empilement de trois trames et utilisation consécutive de l'interpolation et la quantification vectorielle multi-dictionnaires. Pour cela, nous avons tiré profit de certaines caractéristiques des paramètres représentant le signal de parole. La redondance qui caractérise l'enveloppe spectrale du conduit vocal surtout, dans les zones voisées, nous a motivé à utiliser une quantification conjointe par interpolation des coefficients du filtre représentant le conduit vocal. La procédure que nous avons mise en place consiste à quantifier les paramètres d'une trame et d'utiliser l'interpolation pour avoir les paramètres des deux autres trames. Pour quantifier le timbre vocal ou le pitch, notre choix est porté sur la quantification vectorielle multi-dictionnaires. Ce choix est justifié par la dynamique relativement faible de ce dernier, dans des plages bien définies. Chaque dictionnaire couvrira une certaine plage de variation du pitch.

Ce mémoire est organisé de la manière suivante :

Chapitre I : ce chapitre est une introduction au codage de la parole. Dans ce chapitre, on rappelle les objectifs et la classification des algorithmes de codage, on présente les techniques de quantification et d'évaluation utilisées et enfin on résumera l'état de l'art actuel sur le codage de la parole à bas et à très bas débit.

Chapitre II : dans ce chapitre on présentera trois standards de codage prédictif à bas débit : le FS1015, le FS1016 et le FS1017. Ces standards partagent le même model de production mais avec des modélisations du signal d'excitation différentes. Cette présentation montrera l'évolution des standards de codage et justifiera notre choix porté sur le codage MELP.

Chapitre III : On y en détaillera l'implémentation sous MATLAB d'un codeur à très bas débit fonctionnant à 1.2 kbps en utilisant le codage MELP.

Chapitre IV : ce chapitre présentera les résultats obtenus ainsi une discussion de ces résultats.

Enfin, nous terminerons notre mémoire par une conclusion générale sur ce travail et nous présenterons ses perspectives.

CHAPITRE

1

GENERALITES SUR LE CODAGE DE LA PAROLE

I.1. Introduction

Les limites imposées par les canaux de transmission, les exigences de la qualité, l'incompatibilité des systèmes de codage haut débit avec ces contraintes et en parallèle, le développement théorique et hardware des systèmes de traitement numérique des signaux, toutes ces conditions ont poussé les travaux de recherche à aller plus loin dans le codage de la parole à bas débit. Ces travaux cherchent à bien modéliser l'évolution du contenu spectral du signal de parole (au lieu de représenter la forme d'onde temporelle) dont l'objectif principal est soit d'améliorer la qualité du signal synthétique pour un débit particulier ou bien de minimiser le débit pour avoir une qualité particulière. Le débit approprié pour stocker ou transmette un signal de parole dépend du coût de stockage ou de transmission et de la qualité exigée [1].

Ce premier chapitre est une introduction au codage de la parole à bas et très bas débit. Il donne des généralités sur les systèmes de codage (structure, classification, quantification, évaluation). Il rappelle certaines caractéristiques du signal de parole (production, perception, modélisation) et se termine par un bref état de l'art sur le codage à bas et très bas débit.

I.2. Production et perception de la parole

Pour atteindre une qualité acceptable du signal synthétisé à un débit aussi bas que 4 kbps, il est nécessaire de prendre en considération à la fois le mode de production du signal de parole et les limites du système auditif humain. Ceci permet de construire des modèles qui diminuent la redondance présente dans le signal de parole. C'est-à-dire, qui ne considèrent que l'information perceptuellement importante [2].

I.2.1. Production de la parole

La production de la parole est assurée, chez l'homme, par la contribution de plusieurs organes : poumons, larynx, pharynx et le conduit vocal [3]. L'air sous pression expulsé des poumons, traverse les cordes vocales. Celles-ci se comportent de deux manières:

- Soit elles entrent en action pour donner une excitation quasi-périodique avec une structure spectrale harmonique et produire des sons voisés. Ces sons sont caractérisés par une fréquence de vibration appelée fréquence fondamentale qui varie entre 80Hz et 600Hz (80 à 200 Hz pour une voix masculine, 150 à 450 Hz pour une voix féminine et 200 à 600 Hz pour une voix d'enfant) [4].
- Ou bien elles restent écartées; l'excitation dans ce cas sera assimilée à un bruit blanc et son passage dans le conduit vocal donne des sons non-voisés.

La figure I.1 montre les représentations spectrales d'un son voisé et un autre non voisé.

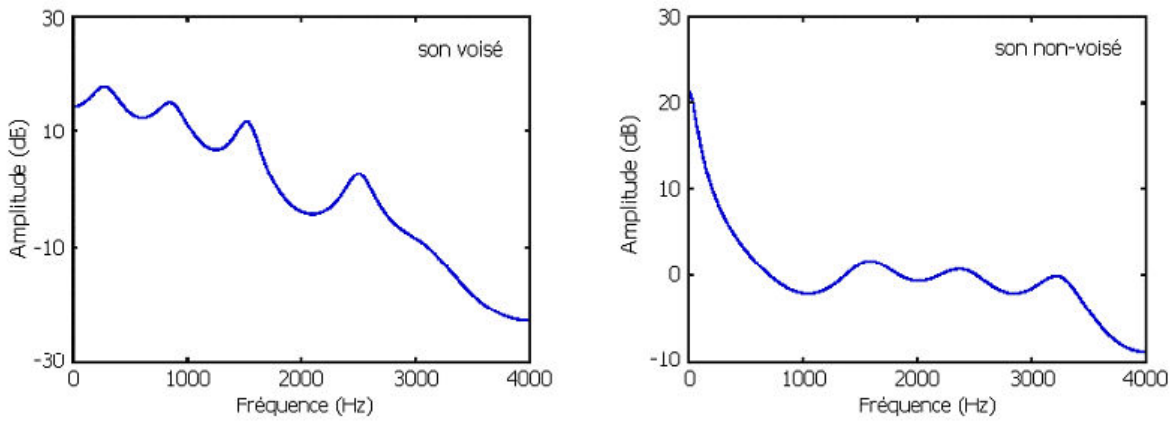


Fig I.1. Spectres d'un son voisé et d'un son non-voisé

La fréquence de vibration des cordes vocales (ou timbre vocal) est une composante fondamentale en traitement de la parole. Elle est souvent utilisée en codage à bas débit et en reconnaissance automatique de la parole. A cet effet plusieurs algorithmes de détection de la fréquence fondamentale ont été proposés dans la littérature [5, 6].

Le conduit vocal joue le rôle d'une caisse de résonance. Il ajoute une enveloppe, caractéristique d'un son, au spectre de l'excitation. Les maxima de cette enveloppe sont

appelés formants. Un modèle simple de production de la parole consiste alors en un signal d'excitation passant au travers d'un système qui représente le conduit vocal [3].

I.2.2. perception de la parole

Pour la perception de la parole, l'oreille réagit à des sons de diverses fréquences qui peuvent être regroupées sur des échelles linéaires ou non linéaires. De plus, notre oreille est limitée dans ses capacités de perception. Il est par exemple impossible de distinguer des sons de plus de 20KHz, plage des ultrasons. De même, l'homme ne peut pas distinguer des sons d'une fréquence inférieure à 20Hz, plage des infrasons. A l'intérieur de cet espace fréquentiel existe un sous-espace délimité par les niveaux d'énergie des sons. Il existe une limite d'énergie en deçà de laquelle l'homme ne percevra pas un son ayant pourtant une fréquence appartenant au spectre de l'audition. Cette limite d'énergie est appelée seuil d'audition. Ce seuil est variable en fonction de la fréquence. Inversement, il existe une limite d'énergie maximale. Cette limite ne doit pas être franchie car la cochlée, et plus particulièrement les cellules ciliées, peuvent être irrémédiablement endommagées. Cette limite s'appelle le seuil de douleur. Elle est aussi variable en fonction de la fréquence [3].

Une autre propriété de la perception humaine souvent exploitée dans le codage de la parole, c'est le masquage (figure I.2).

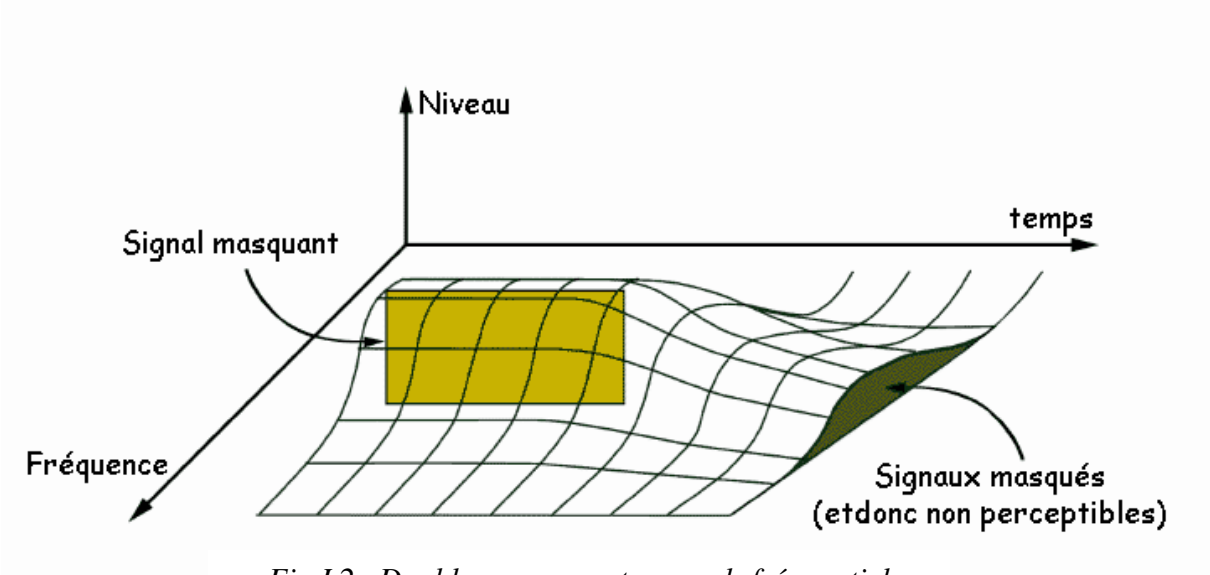


Fig I.2. Double masquage temporel- fréquentiel

On appelle masquage, le phénomène qui apparaît quand certains sons deviennent inaudibles par la présence d'autres sons. Ce phénomène dépend principalement de l'intensité dont les

sons les plus intenses masquent les sons les moins intenses sous d'autres contraintes de fréquence (masquage fréquentiel) et de temps (masquage temporel) [5, 7].

I.3. Caractéristiques du signal de parole

Le signal de parole peut être considéré comme un élément porteur d'information. Il peut être vu aussi comme une fonction du temps au sens de la théorie du signal [4].

Le signal de parole est quasi-stationnaire. Il existe un intervalle du temps sur lequel ses caractéristiques statistiques ne varient pas avec le temps [1, 7]. Sur 50 ms le signal de parole peut être considéré comme stationnaire. C'est la durée moyenne d'un phonème [7]. Dans le codage paramétrique, les paramètres représentant le signal de parole sont pratiquement invariables pour un phonème et le choix de la longueur des trames d'analyse est effectué d'une manière à respecter la quasi-stationnarité de ces paramètres. Dans le tableau I.1, on présente les longueurs des segments d'analyse pour certains vocodeurs (standards de codage). Le MELP (*Mixed Excitation Linear Prediction*), le CELP (*Code Exited Linear Prediction*)et le LPC 10 (*Linear Predictive Coding 10*) sont des standards DoD (département de la défense américaine) que l'on présentera dans le chapitre suivant. Le GSM AMR (adaptive multi-rate) est un standard de codage GSM à débit variable qui utilise le ACELP (Algebric CELP) comme algorithme de codage [8].

Tableau I.1. Longueurs de trames d'analyse pour différents vocodeurs

Vocodeur	Trame d'analyse (ms)
MELP (FS-MELP)	22.5
CELP (FS-1016)	30
LPC10 (FS-1015)	22.5
GSM AMR	50

Une autre caractéristique importante du signal de parole, concerne sa distribution spectrale d'énergie. Le spectre à long terme de ce dernier n'est pas plat. Il est plus énergétique dans les basses fréquences comparativement aux hautes fréquences où on observe une

atténuation d'énergie d'environ 10db/octave (pour les fréquences supérieures à 500 Hz) [9]. Des explications à cette répartition spectrale d'énergie sont données par WC. CHU [1].

I.4. Objectifs d'un Système de codage

Un codeur de parole idéal doit avoir une bonne qualité, un débit faible, une complexité réduite, une robustesse élevée aux erreurs de transmission et un algorithme de calcul très rapide [1, 10]. De nombreux travaux relatifs aux algorithmes de codage tendent à optimiser le compromis entre l'efficacité, le coût et la qualité de transmission des systèmes de communication en fonction des débits disponibles. Les techniques de codage numérique adaptent donc leur débit de transmission aux capacités du canal.

I.5. Structure d'un système de codage/décodage

Un système de codage est constitué de deux parties principales : un encodeur et un décodeur. Le rôle de l'encodeur est de transformer la parole originale en une combinaison numérique (ensemble de bits), cette combinaison alimentera par la suite l'entrée du décodeur pour reconstituer un signal de parole synthétique [11]. L'entrée de l'encodeur et la sortie du décodeur ne sont pas identiques, et le degré de similitude (temporel ou perceptuel) entre les deux définit la qualité du codeur.

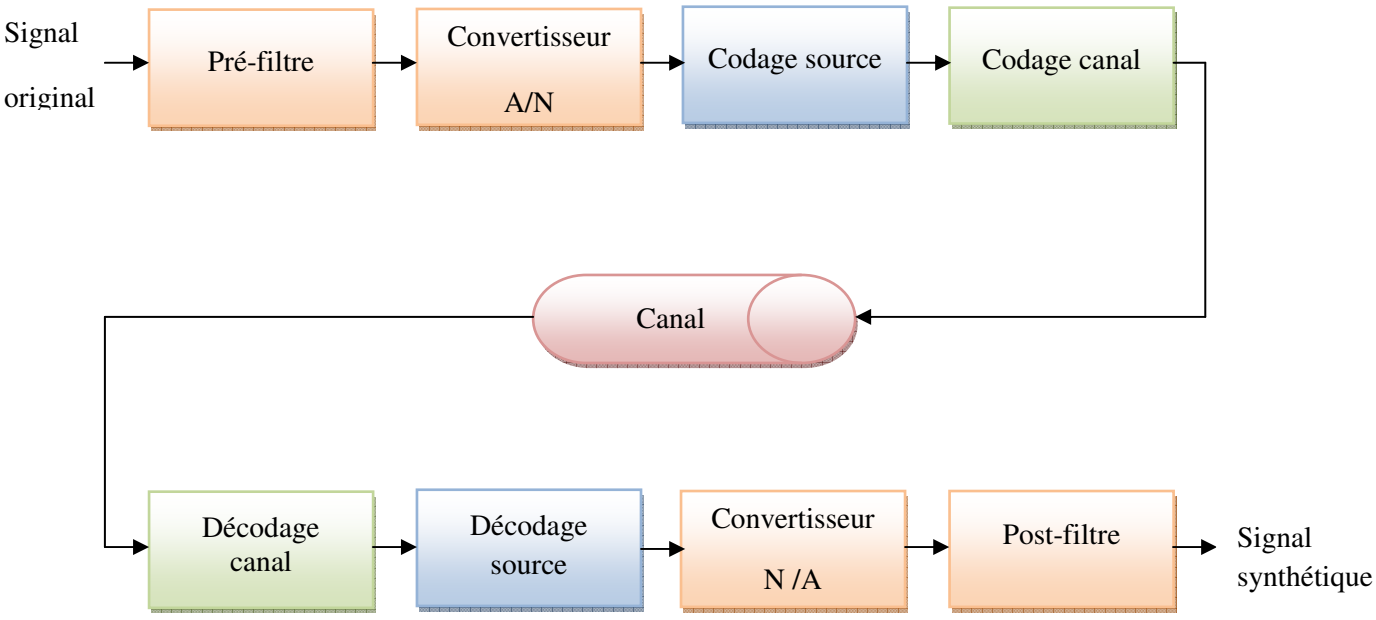


Fig I.3. Structure d'un système de codage/décodqge de parole [1]

La figure I.3 représente le schéma de base d'un système de codage. Comme les algorithmes de codage sont implémentés dans des calculateurs qui n'acceptent que des signaux numériques, un système d'acquisition numérique doit être mis en place. Un tel système est constitué d'un récepteur du signal analogique, d'un pré-filtre, d'un échantillonneur et d'un convertisseur analogique numérique [12]. Pour l'échantillonnage on utilise une fréquence de 8KHz, c'est la fréquence de Nyquist appropriée à la parole (la fréquence maximale en téléphonie est égale à 3400 Hz) [5, 13].

I.6. Classification des codeurs de parole

Plusieurs critères ont été choisis pour la classification des algorithmes de codage dont les deux critères les plus utilisés dans la littérature sont basés sur la technique de codage utilisée et sur la plage de débit où le codeur opère.

I.6.1 Classification des codeurs selon la technique de codage utilisée

Les algorithmes de codage de parole basiques peuvent être divisés en deux classes distinctes: les codeurs temporels ou de forme d'onde et les codeurs paramétriques [4, 14].

A. Codeurs de la forme d'onde (waveform coders)

Plus simples à mettre en œuvre, ils cherchent avant tout à préserver l'allure temporelle du signal de parole, ce qui les rend robustes aux différents types de bruits [15]. La qualité d'un codage de forme d'onde est jugée bonne pour un débit supérieur à 32 kbps [1]. Certains codeurs de forme d'onde peuvent aller jusqu'à 16 kbps [4]. En dessous de ce seuil, leur détérioration est assez nette. On préfère alors utiliser des codeurs paramétriques ou des codeurs hybrides.

B. Codeurs paramétriques

Les codeurs paramétriques utilisent un modèle de production et exploitent certaines caractéristiques de la perception humaine pour décrire le signal de parole par un jeu de paramètres [1]. Ainsi, en augmentant le débit, le signal synthétisé ne converge pas vers la forme du signal original et sa qualité est limitée par la précision du modèle [14]. Comme ces codeurs reposent sur un modèle de production de la parole, leur performance est d'habitude faible pour d'autres signaux [1].

Les codeurs paramétriques peuvent être classés en trois classes : les vocodeurs, les codeurs sinusoïdaux et les codeurs par interpolation de la forme d'onde [2].

Les vocodeurs

C'est la classe la plus importante des codeurs paramétriques. Les vocodeurs sont souvent basés sur la prédiction linéaire et diffèrent principalement dans la manière de construire le signal d'excitation. Le plus simple de ces codeurs utilise une excitation générée par un train d'impulsions espacées de périodes de pitch pour les sons voisés et par un bruit aléatoire pour les sons non-voisés. D'autres vocodeurs utilisent une excitation mixte. Le signal d'excitation de ce modèle proposé initialement par Makhoul [16], est composé d'un train d'impulsions dans les fréquences basses et d'un bruit dans les fréquences hautes. Cette idée a été développée par la suite par McCree et Barnwell dans le coder MELP (Mixed Excitation Linear Prediction) où le mélange de la composante harmonique et de la composante de bruit se fait à l'aide de deux filtres FIR (*Finate Impulse Response)* variables dans le temps [17].

Les codeurs sinusoïdaux

Les codeurs sinusoïdaux synthétisent la parole par une somme de sinusoïdes dont les amplitudes décrivent le spectre à court terme du signal de parole [18]. Pour les bas débits, uniquement les amplitudes sont quantifiées et les fréquences des sinusoïdes sont habituellement harmoniques. Pour les débits plus élevés, l'information concernant les phases et les fréquences des sinusoïdes peut être transmise. Ce modèle convient particulièrement au codage de la parole voisée où le signal synthétisé est reconstruit par une combinaison linéaire des sinusoïdes de fréquences harmoniques de la fréquence fondamentale. Les sons non-voisés peuvent être synthétisés avec le même modèle en utilisant des phases aléatoires. Les représentants les plus importants des codeurs sinusoïdaux sont le codeur STC (Sinusoidal Transform Coder) [19] et le codeur MBE (Multi-Band Excitation) [20].

Les codeurs par interpolation de la forme d'onde

Le dernier groupe de codeurs paramétriques comprend les codeurs par interpolation de forme d'onde (WI pour Waveform Interpolation). Dans ces codeurs, une forme d'onde caractéristique est extraite du signal régulièrement et ses paramètres sont interpolés d'une trame à l'autre [2]. Plus récemment, une décomposition de la forme d'onde

caractéristique a été proposée. Deux formes d'onde sont alors extraites et transmises. La première représente la composante périodique et la seconde la composante aléatoire du signal de parole. Un autre algorithme basé sur l'interpolation de forme d'onde, appelé TFI (Time-Frequency Interpolation), a été proposé par Shoham [2].

Bien que la technique WI soit généralement appliquée dans le domaine du signal d'excitation du filtre LP, elle a été aussi employée dans le domaine du signal de parole.

I.6.2. Classification des codeurs selon la plage des débits

On distingue en général 3 plages de débits [1]:

- Les hauts débits, supérieurs à 15 Kbps, correspondant à des algorithmes de codage de la forme d'onde non spécifiques à la parole,
- Les débits moyens, de 5 Kbps à 15 Kbps, correspondant à des techniques de codage hybrides utilisant des méthodes de codage de la forme d'onde et prenant en compte certaines propriétés de la parole ou de la perception auditive
- Les bas et très bas débits, de quelques centaines de bits par seconde à 5 Kbps, correspondant aux vocodeurs (VOice CODER) spécifiques au codage de la parole.

I.7. La prédiction linéaire

I.7.1. Principe

Le codage de la parole à bas débit exige des représentations compactes et précises du filtre du conduit vocal et de la source d'excitation. C'est pourquoi Atal et Schroeder ont pensé en 1968 à exploiter les redondances du signal de parole en le modélisant par un nombre restreint de paramètres à base d'un filtre linéaire. C'était l'idée de ce qu'on appelle aujourd'hui la prédiction Linéaire (LP : Linear Prediction) [21]. Elle est l'un des outils les plus importants de l'analyse de la parole [16].

Le modèle auto régressif (AR), est souvent utilisé pour modéliser le filtre LP. En plus de sa simplicité et de son efficacité dans les systèmes temps réel, ce modèle représente parfaitement les résonances spectrales des sons voisés [9]. La fonction de transfert du conduit vocal est alors approximée par le modèle tout-pôle $H_{AR}(z)$ suivant [9] :

$$H_{AR}(z) = \frac{G}{1+\sum_{i=1}^{p} a_i z^{-i}} \quad \text{(I.1)}$$

Le facteur de gain G est généralement égal à 1, conduisant à l'expression :

$$H_{AR}(z) = \frac{1}{A(z)} \quad \text{(I.2)}$$

où

$$A(z) = \sum_{i=1}^{p} a_i z^{-i} \quad \text{(I.3)}$$

Les paramètres $\{a_k\}$ sont appelés coefficients LP. Différentes techniques existent pour estimer ces coefficients telles que la méthode d'auto-corrélation et la méthode de covariance ou de Burg [9]. Mais la méthode d'auto-corrélation reste la plus populaire et la plus efficace grâce à l'algorithme de résolution de Levinson-Durbin [16], qui détermine les coefficients $\{a_k\}$ k=1...p, de manière à minimiser l'énergie du signal résiduel par la méthode classique des moindres carrés.

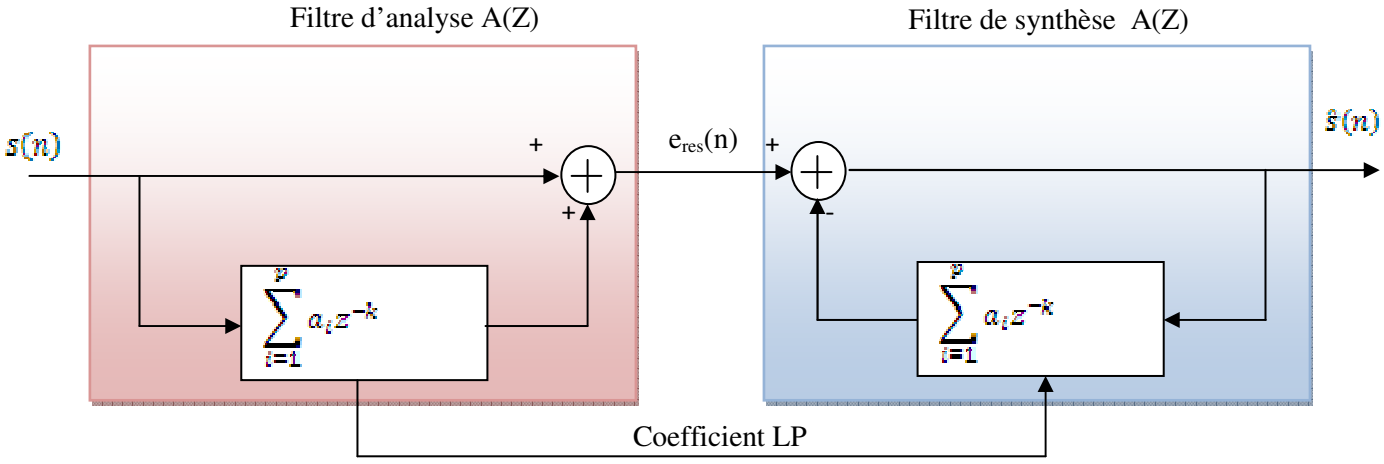


Fig I.4. Analyse et synthèse LP

La modélisation complète peut maintenant être décomposée en deux parties (figure I.4) :

- une partie analyse qui filtre le signal d'entrée avec la fonction de transfert A(z).Ce filtre tout-zéro est défini comme le filtre d'analyse LP et permet d'extraire l'information prédictible du signal et de définir un signal d'erreur résiduelle entre le signal de parole d'entrée $s(n)$ et son estimation : $\hat{s}(n)$

$$e_{res}(n) = s(n) - \hat{s}(n) = s(n) + \sum_{i=1}^{p} a_i s(n-i) \quad \text{(I.4)}$$

- une partie synthèse qui effectue un filtrage par fonction de transfert H_{AR} (z).Ce filtre tout-pôle, connu sous le nom de filtre de synthèse LP, permet de reconstruire, à l'aide d'un signal d'excitation approprié, un signal de parole artificiel.

Le signal résiduel e_{res}(n) est l'excitation idéale du modèle H_{AR} (z). Une modélisation précise de ce signal permet d'obtenir un signal reconstruit naturel. Or l'estimation des paramètres LP du modèle vocal entraîne une approximation du signal d'excitation. A mesure que l'ordre du modèle LP augmente, un meilleur ajustement au spectre de la voix est obtenu. Cependant, plus l'ordre p sera élevé, plus le nombre de paramètres à transmettre augmentera. L'ordre doit ainsi résulter d'un compromis entre une bonne représentation de la structure formantique du signal de parole, de la complexité de calcul et du débit de transmission. En général, en codage de la parole un ordre de dix est souvent utilisé [22].

I.7.2. Gain de prédiction

On appelle gain de prédiction la grandeur suivante [1] :

$$G_p = 10\log\left(\frac{\sum_{n=0}^{+\infty} s(n)^2}{\sum_{n=0}^{+\infty} e_{res}(n)^2}\right) \quad \text{(I.5)}$$

Cette grandeur est le rapport entre l'énergie du signal de parole et l'énergie du résiduel de prédiction. Elle mesure la précision de la prédiction.

La figure I.5 illustre les variations du gain de prédiction en fonction de l'ordre du filtre d'analyse. On constate que la précision augmente avec l'ordre du filtre.

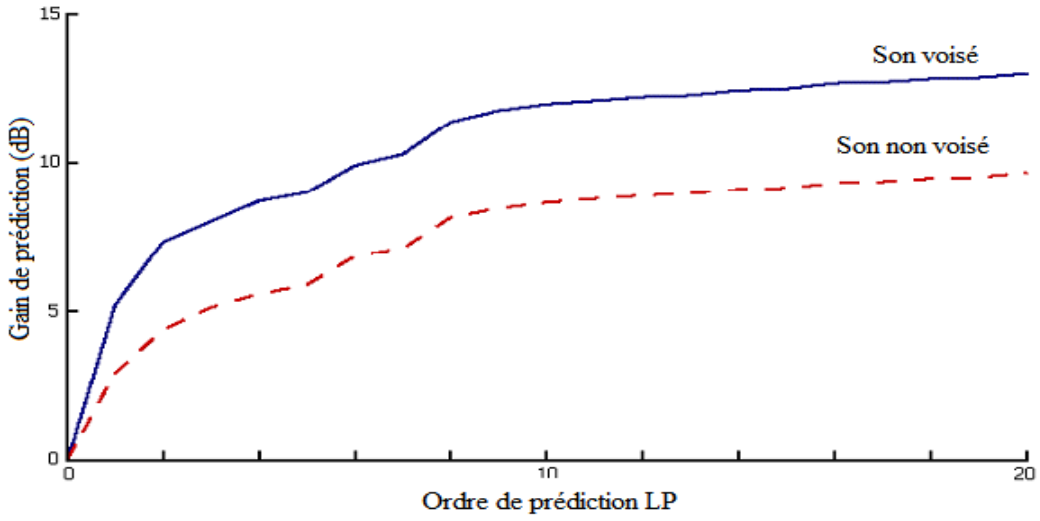


Fig I.5. Variation du gain de prédiction en fonction de l'ordre du filtre LP [22]

I.7.3. Représentation des coefficients LP

Dans tout codage de parole à bas débit, il est nécessaire, pour transmettre les coefficients de prédiction linéaire, de les quantifier en utilisant un nombre restreint de bits. Cependant, les coefficients LP ne sont pas appropriés à une quantification directe à cause de leur intervalle de définition (dynamique) trop important. Leur codage nécessiterait en effet 8 à 10 bits par coefficient [23]. Plusieurs représentations alternatives aux coefficients LP ont été proposées dont les plus connues sont :

- LSF ou line spectral frequency
- RC ou coefficients de réflexion (reflection coefficients)
- AC ou coefficients d'auto-corrélation (auto-corrélations coefficients)

La première représentation est la plus utilisée. Elle modélise le spectre LPC en déterminant ses lignes de fréquences spectrales (LSF) ou ses lignes de raies spectrales (LSP) [14]. Les coefficients LSF sont les racines des polynômes P(z) et Q(z) suivants [14]:

$$P(z) = \left(1 + z^{-(p+1)}\right)A(z) \tag{I.6}$$

$$Q(z) = \left(1 - z^{-(p+1)}\right)A(z) \tag{I.7}$$

Pour un filtre A(z*)* stable, il suffit que tous les zéros de P(z) et Q(z) soient situés sur le cercle unité et qu'ils alternent. Les LSF ont l'avantage d'avoir une relation simple avec les formants et ils sont plus faciles à trouver que les pôles du filtre A(z) parce que la recherche d'une racine LSF est une opération unidimensionnelle. De façon générale, plus petite est la distance entre deux fréquences LSF voisines, plus grande est l'amplitude de l'enveloppe spectrale entre ces fréquences. Cette propriété permet de quantifier les LSF en exploitant certaines caractéristiques de la perception humaine.

Pour la quantification des coefficients LP, la plupart des codeurs standardisés après 1994 utilisent la quantification vectorielle pour les coefficients LSF (représentants des coefficients LP). Dans la majorité des cas, ce choix est justifié par les performances fournies par la quantification vectorielle dans le codage à bas débit en comparaison avec la quantification scalaire [1]. Les techniques de quantification vectorielle souvent utilisées sont :

la quantification vectorielle multi étage, la quantification vectorielle divisée et la quantification vectorielle prédictive [1, 14]. Quelquefois on utilise une quantification adaptative [24].

I.8. Quantification

Au cours du traitement numérique du signal de parole, toutes les données sont représentées par un certain nombre d'éléments binaires avec une précision finie. Cette opération consiste à représenter les amplitudes du signal analogique, à des instants discrets du temps, par une valeur choisie parmi un ensemble fini [1]. Outre la nécessité de la quantification pour numériser les données, elle est aussi un moyen de compression.

I.8.1. Quantification scalaire (Scalar Quantization (SQ))

Soit s(n) le paramètre à quantifier. Notons $f_S(s)$ la densité de probabilité de la variable s(n) et b sa résolution ou le nombre de bits nécessaires pour la représenter. L'opération de quantification scalaire consiste à discrétiser le paramètre s pour obtenir une représentation numérique i(n) de l'information qu'il représente. Son domaine de définition est alors partitionné en $L = 2^b$ intervalles distincts et un représentant est défini pour chacun de ces intervalles. Si la longueur de l'intervalle est fixe la quantification scalaire est dite uniforme sinon elle est dite non uniforme.

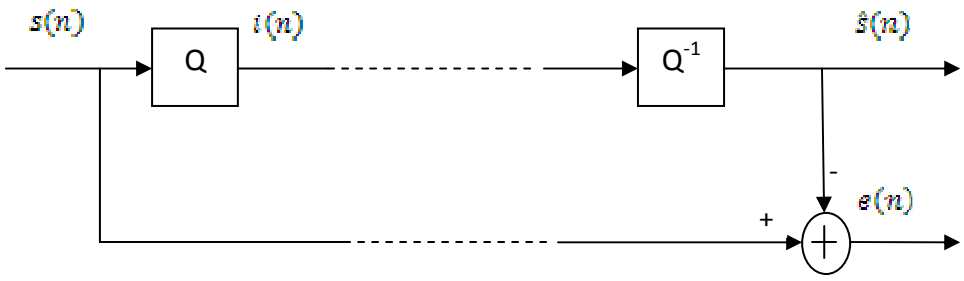


Fig I.6. Quantification scalaire

La procédure d'encodage Q décide à quel intervalle appartient le paramètre s(n) et lui associe le numéro $i(n) \in \{1...L\}$ correspondant (figure I.6). C'est ce numéro d'intervalle qu'il faudra transmettre au processus de décodage Q^{-1} qui effectuera la procédure inverse en associant à i(n) son représentant $\hat{s}(n)$. L'opération de quantification apportera toujours des

dégradations irréversibles par rapport au signal d'origine qui se traduisent par une erreur, ou bruit, de quantification e (n) :

$$e(n) = s(n) - \hat{s}(n) \tag{I.8}$$

Ou $\{ \hat{s}(n) \}_{1 \leq i \leq L}$ est l'ensemble des représentants, appelé dictionnaire.

I.8.2. Quantification vectorielle (Vector Quantization (VQ))

Contrairement à la quantification scalaire qui traite chaque échantillon indépendamment des précédents, le quantificateur vectoriel permet de prendre en compte la dépendance entre les différentes composantes du signal [21]. La propriété fondamentale de ce quantificateur est la recherche de la corrélation qui peut exister entre les échantillons successifs du signal [5].

Lorsque le débit devient inférieur ou égal à 16 kbps, soit 2 bits par échantillon, il devient nécessaire de les regrouper pour former un vecteur et de quantifier l'ensemble. Le schéma de principe est le même que celui donné en figure I.6. Les scalaires $s(n)$ et $\hat{s}(n)$ apparaissant à l'instant n sont simplement remplacés par des vecteurs. La quantification vectorielle (VQ) choisit les représentants des vecteurs de données à quantifier dans un ensemble de vecteurs prédéterminés, de même dimension, stockés dans une table ou un dictionnaire.

La quantification vectorielle permet d'atteindre de meilleures performances que la quantification scalaire. La VQ peut être considérée comme une généralisation de la SQ. Les techniques de quantification vectorielle procurent des algorithmes permettant la détermination de dictionnaires quasi-optimaux, tels que l'algorithme de Lloyd-Max et montrent que l'on peut obtenir, pour une résolution donnée (nombre de bits disponibles par échantillon), une distorsion très proche des limites théoriques entre les signaux original et quantifié [25].

I.8.3. Quantification vectorielle multi-étages (MSVQ : Multi Stage Vector Quantization)

Le principe de fonctionnement de la quantification vectorielle de type multi-étages est représenté à la figure I.7. Il procéde par approximations successives. Une quantification vectorielle grossière est réalisée en utilisant le premier dictionnaire (première étage). Ensuite, une deuxième VQ porte sur l'erreur de quantification commse par la première et ainsi de suite [5].

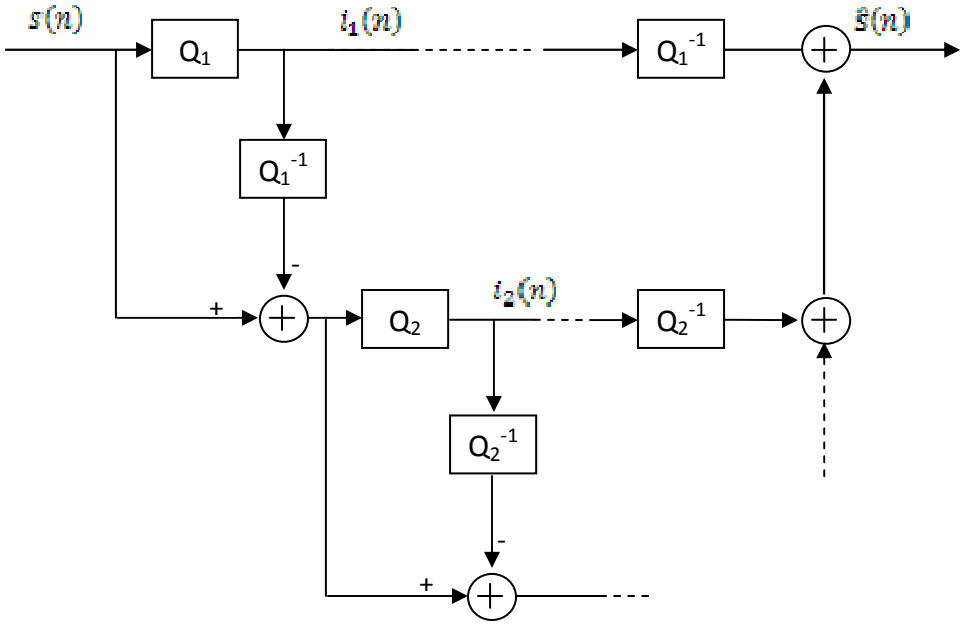


Fig I.7. Quantification vectorielle multi-étage

I.9. Evaluation de la qualité d'un codeur de parole

Du point de vue qualité, un codeur de parole idéal doit assurer une conservation parfaite des informations portées par le signal de parole. Ces informations peuvent être divisées en trois classes [1] :

- **Information phonétique :** c'est le message textuel porté par le signal de parole.
- **Information prosodique :** dans ce type d'information on trouve les émotions, les inflexions, ….
- **Information du locuteur :** elle concerne les composants du signal de parole permettant de caractériser et d'identifier le locuteur.

Devant l'impossibilité d'adopter les méthodes d'évaluation objectives comme le SNR (*Signal to Noise Ratio*) et le SNR segmental au codage de parole à bas et très bas débit (paramétrique) et le coût considérablement élevé de l'implémentation des méthodes subjectives : DRT (*Diagnostic Rhyme Test*), MRT (*Modified Rhyme Test*), MOS (*Mean Opinion Score*), …, des travaux de recherche sont orientés vers le développement des méthodes d'évaluation objectives ayant la même efficacité que les méthodes subjectives [1].

Parmi les méthodes développées, on trouve les méthodes basées sur la distance spectrale entre le spectre du signal original et celui du signal synthétique [26]. On trouve aussi le PSQM (*Perceptual Speech Quality Mesure*) ou mesure perceptuelle de la qualité de parole qui consiste à effectuer des transformations perceptuelles sur le signal synthétique et sur le signal original retardé (pour simuler le retard dû au codage). Cette transformation contient un modèle d'audition simple. En 1996, La PSQM à été choisie par l'Union Internationale des Télécommunications (ITU) pour être un standard d'évaluation recommandé sous la référence P.861 [27].

Une autre méthode proposée par Voran et nommée Measuring Normalizing Block ou MNB a été mise au point dont le but est de rectifier certaines limitations du standard P.861 surtout pour les très bas débits et lorsqu'on utilise des bits de protection. La MNB a été choisie comme un standard d'évaluation par l'ITU en 1998 [28].

Cependant, ni la PSQM ni la MNB ne convient à l'évaluation de bout en bout dans les réseaux de télécommunication à cause de la présence simultanée de plusieurs problèmes : filtrage linéaire, retard variable, bruit de fond,… Ces problèmes ont conduit à la naissance d'un nouveau algorithme d'évaluation dit le PESQ (Perceptuel Evaluation Speech Quality) ou évaluation perceptuelle de la qualité de parole standardisé en 2001 par le ITU-T sous la référence P.862 [29]. Le principe de cette méthode est illustré dans la figure I.8.

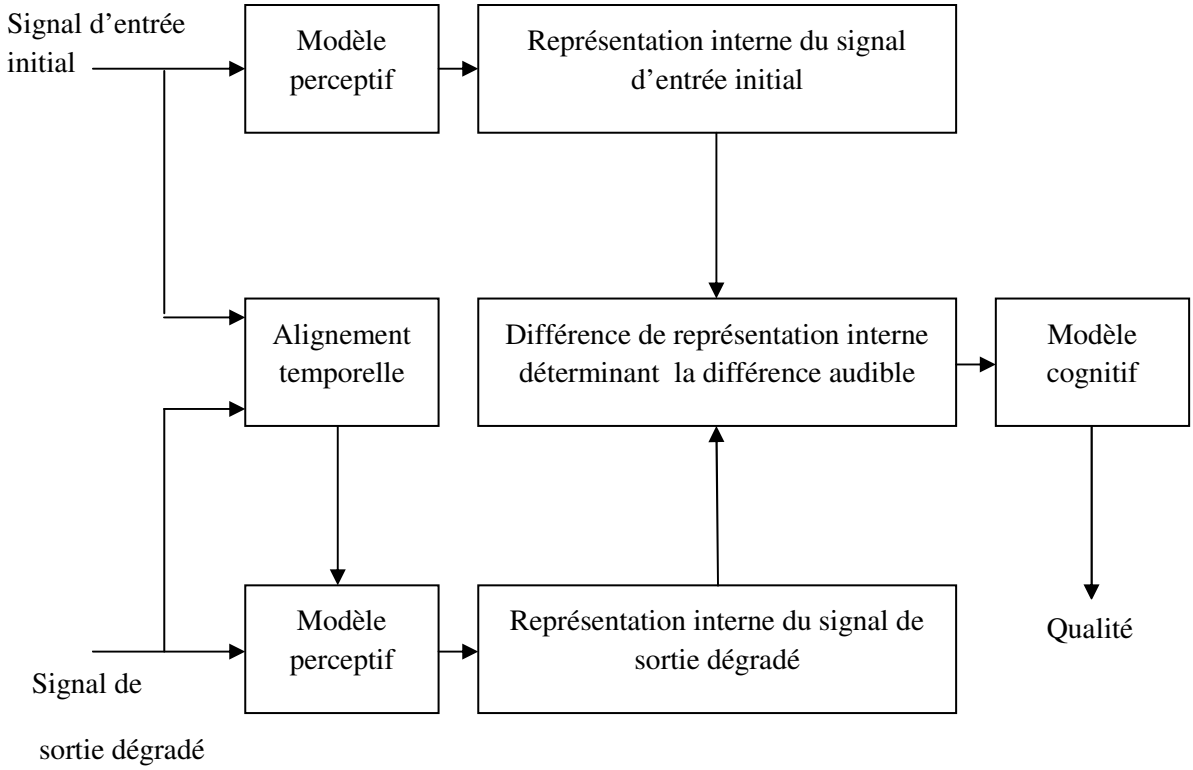


Fig I.8. Schéma bloc de la méthode PESQ

La méthode d'évaluation PESQ prend en compte ces effets au moyen de l'égalisation de la fonction de transfert, de l'alignement temporel et d'un nouvel algorithme de calcul des valeurs moyennes de distorsion dans le temps.

I.10. Evolution du codage de la parole à bas et à très bas débit

En 1972 Atal et Hanauer ont développé le premier codeur à bas débit à base de prédiction linéaire [30]. Malgré les avantages offerts par ce codeur en termes de débit, il n'a pas connu une large utilisation au début de son apparition pour les raisons suivantes [31] :

- La qualité métallique du signal synthétique due au modèle d'excitation trop simplifié. Mais, l'utilisation d'une excitation par code (CELP) ou excitation mixte (MELP), améliore significativement la qualité perceptuelle

- La capacité de calcul demandée incompatible avec les calculateurs existants à cette époque. Ce problème a été résolu avec l'apparition des DSP dans les années 80.

C'est à partir des années 80 que la prédiction linéaire a connu une large utilisation en codage de la parole. Elle a été la base de plusieurs algorithmes devenus par la suite des standards de codage pour différentes applications.

Pour des communications militaires, on accepte une dégradation de qualité encore plus importante et on peut alors atteindre des débits encore plus bas. En 1984, le département de la défense des États Unis a accepté le standard LPC-10 (FS1015) [32]. Il s'agit d'un vocodeur LPC opérant à 2.4 kbps. Pour augmenter la performance, un nouveau standard a été accepté en 1991. Ce standard (FS1016) utilise l'algorithme CELP et opère à un débit de 4.8 kbps [33]. Finalement, un codeur MELP à 2.4 kbps de qualité comparable au FS1016 a été standardisé en 1997 [34] pour remplacer les deux codeurs précédents, FS1015 et FS1016.

Le tableau I.2 résume les standards de codage basés sur la prédiction linéaire avec les applications concernées et leurs années de standardisation [1].

Tableau 1.2. Standards de codage de la parole

Standard	**Année d'apparition**	**Débit (KBS)**	**Applications**
FS1015 LPC	1984	2.4	Militaire
ETSI GSM 6.10 RPE-LTP	1987	13	Communication mobile
TIA IS54 VSELP	1990	7.95	Réseau cellulaire (USA)
ETSI GSM6.20 VSELP	1990	5.6	Réseau GSM
RCR STD-27B VSELP	1990	6.7	Réseau cellulaire (JAPON)
FS1016CELP	1991	4.8	Militaire
ITU-T G.728 LD-CELP	1992	16	Utilisations diverses (téléphonie, multimédia, …)
TIA IS96 VBR-CELP	1993	8.5, 4, 2, 0.8	Réseau cellulaire (USA)
ITU-T G.723.1 MP-MLQ/ ACELP	1995	5.3, 6.3	Multimédia, vidéophone, …
ITU-T G.729 CS-ACELP	1995	8	Utilisations diverses (téléphonie, multimédia, …)
ETSI GSM EFR ACELP	1996	12.2	Utilisations diverses (téléphonie, multimédia, …)
TIA IS641 ACELP	1996	7.4	Réseau cellulaire (USA)
FS MELP	1997	2.4	Militaire
ETSI AMR-ACELP	1999	12.2, 10.2, 7.95, 7.4, 6.7, 5.9 ,5.15, 4.75	Utilisations diverses(téléphonie, multimédia, …)

I.11. Conclusion

On a vu, dans ce chapitre que seuls les codeurs paramétriques peuvent descendre à la plage des bas et très bas débits. Ces codeurs sont basés sur un modèle de production. Une modélisation AR permet d'avoir un signal synthétique de qualité. Dans le cas de la parole, cette modélisation comprend la modélisation du conduit vocal et la modélisation du signal d'excitation. Dans le chapitre suivant on étudiera trois standards de codage à bas débit utilisant le même modèle de production mais avec des modèles du signal d'excitation différents.

CHAPITRE

2

CODAGE PREDICTIF LINEAIRE

II.1. Introduction

Les techniques de codage à bas débit cherchent à bien modéliser l'enveloppe spectrale du signal de parole au lieu de représenter la forme d'onde temporelle [25]. Ils doivent éliminer les informations sans pertinence pour la perception. Rappelons que ces codeurs utilisent certaines caractéristiques de la perception et de la production de la parole, aussi sont-ils généralement très peu efficaces pour les signaux autres que la parole, comme les signaux musicaux par exemple.

Ce chapitre donne des généralités sur le codage de la parole à bas débit à base de la prédiction linéaire. Il présente trois algorithmes de codage prédictif : le codage prédictif à deux états d'excitation représenté par l'algorithme LPC, le codage prédictif à excitation par code ou CELP et le codage prédictif à excitation mixte ou MELP. Ces algorithme partagent le même modèle de production de la parole (un modèle autorégressif ou AR d'ordre 10) et ils diffèrent essentiellement dans la modélisation de l'excitation (simple, par code ou mixte). Ces algorithmes de codage seront présentés tels qu'ils sont décrits dans les standards DoD FS1015 pour la LPC10, le FS1016 pour le CELP et le FS1017 pour le MELP.

II.2. Codage prédictif à deux états d'excitation (LPC)

Le premier codeur prédictif à deux états d'excitation a été développé par Atal et Schroeder en 1967 [35]. Dans cette technique de codage, les trames de parole sont classées en trames voisées et en trames non voisées et par conséquence, deux types d'excitation sont utilisées. Une excitation périodique formée d'un train d'impulsions pour les trames voisées et

d'un bruit blanc pour les trames non voisées. Donc, à l'entrée du codeur, une analyse doit être effectuée sur des segments (des trames) du signal de parole pour :

- Calculer les coefficients du filtre de prédiction linéaire
- Etudier la périodicité (voisement) avec estimation de la fréquence fondamentale (le pitch)
- Calculer l'énergie du signal. La figure II.1 illustre le schéma d'analyse LPC [1].

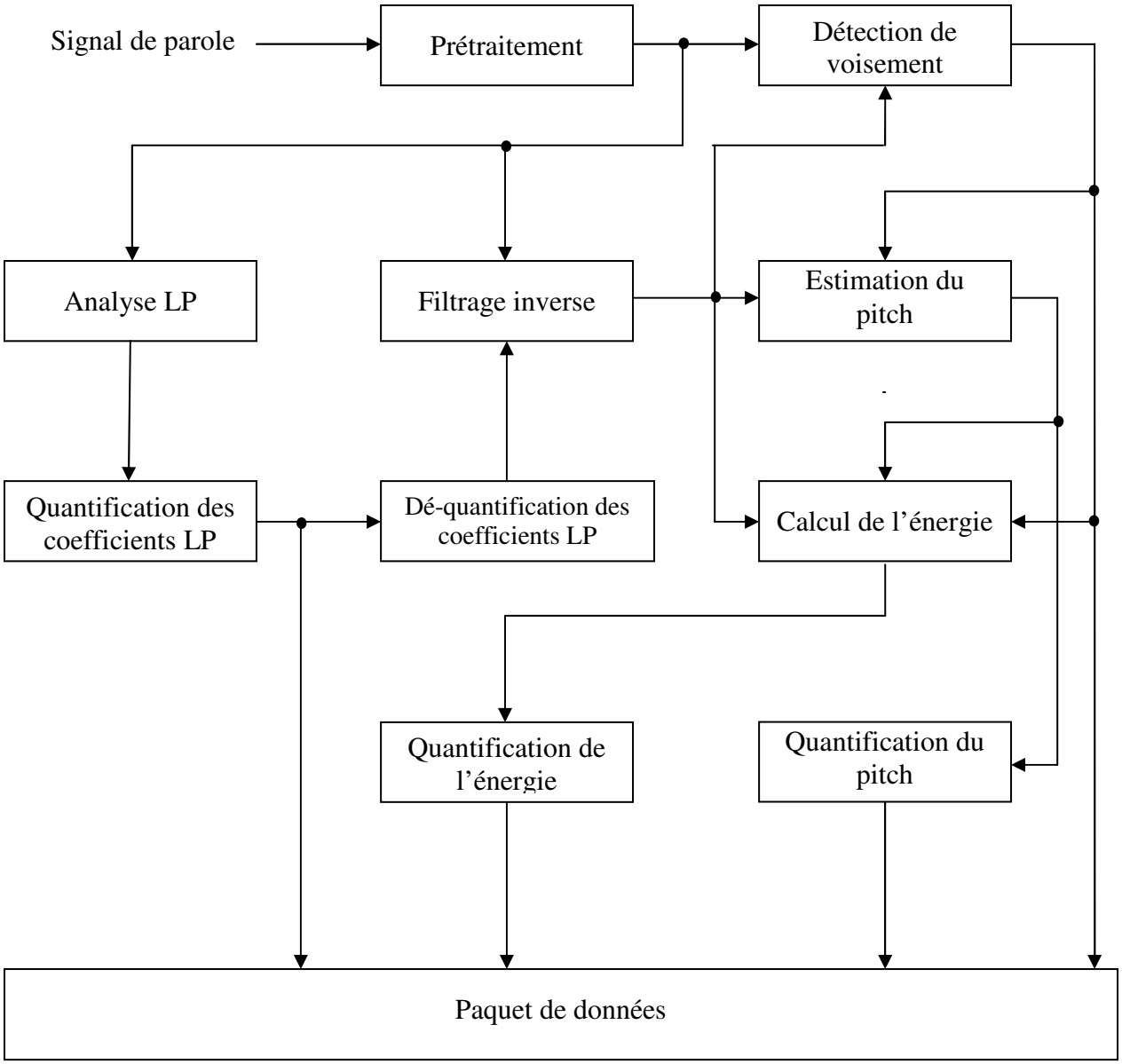


Fig II.1. Schéma bloc de l'analyse LPC [1]

II.2.1. L'analyse LP

Le vocodeur LPC utilise la méthode de covariance pour déterminer les coefficients LP [36] mais des implémentations récentes de ce dernier utilisent la méthode d'auto-corrélation [1]. L'analyse LP est synchronisée avec le pitch dont les positions des fenêtres d'analyse sont ajustées en fonction de la fréquence fondamentale de la trame concernée.

II.2.2. Détection de voisement et estimation du pitch

Pour la détection de voisement, un classifieur linéaire est utilisé à base des paramètres suivants [1, 37] :

- L'énergie de la bande basse du signal.
- la fonction des différences des amplitudes.
- Le taux de passage par zéro.

En ce qui concerne le pitch, le codeur LPC cherche à trouver la fréquence fondamentale dans la bande [50 400]Hz (en nombre d'échantillons [20 160]) parmi 60 candidats : 20, 21, …,39, 40, 42, …, 78, 80, 84, …, 156 [1]. Cette recherche est basée sur la fonction des différences des amplitudes ou MDF (magnitude difference function) [14], appliquée au résiduel de prédiction. La méthode MDF est moins précise que la méthode d'auto-corrélation mais elle offre plus de souplesse en implémentation car les additions sont plus simples à réaliser que les multiplications [1].

II.2.3. Calcul du gain

Le gain à transmettre dans le codeur LPC est la puissance du résiduel de prédiction. Le calcul de cette dernière pour les trames voisées doit prendre en considération la valeur du pitch [1].

Soient e(n) le résiduel de prédiction, N la longueur de la trame et T le pitch. La formule de calcul du gain G sera donnée par :

$$G = \begin{cases} \frac{1}{N}\sum_{n=1}^{N} e(n)^2 & si\ la\ trame\ est\ non-voisée \\ \frac{1}{\lfloor N/T \rfloor T}\sum_{n=1}^{\lfloor N/T \rfloor T} e(n)^2 & si\ la\ trame\ est\ voisée \end{cases} \quad \text{(II.1)}$$

Où $\lfloor x \rfloor$ est la partie entière de x

II.2.4. synthèse

La figure II.2 montre le Schéma du décodeur LPC.

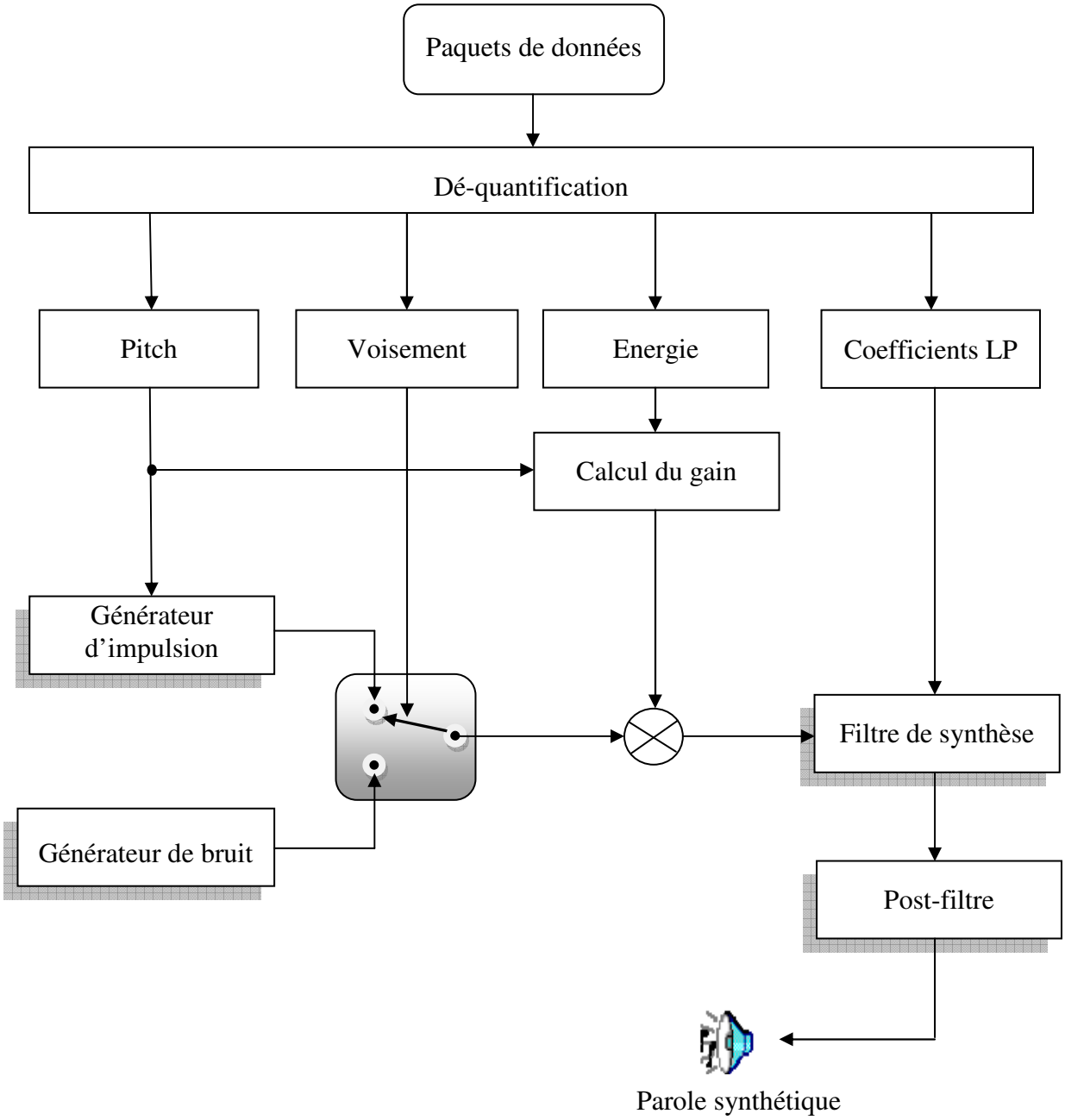


Fig II.2. Décodeur LPC

La décision de voisement et le pitch sont utilisés pour générer un signal d'excitation. Sur ce dernier, on va appliquer un filtrage inverse en utilisant les coefficients LP reçus de l'encodeur et enfin la sortie de ce filtre est mise à l'échelle pour donner au signal synthétique

le même niveau d'énergie que le signal original en utilisant le gain. Ce modèle de production trop simplifié est à l'origine de la qualité métallique perçue dans le signal synthétique [1].

II.2.5. Format de transmission

Dans sa version standardisée par le DoD sous la référence FS1015, le LPC fonctionne à 2.4 kbps, avec des trames d'analyse de 22.5 ms, ce qui donne 54 bits par trame [1]. Ce codeur utilise la quantification scalaire (SQ) pour l'ensemble des paramètres extraits de l'analyse. Le tableau II.1 montre la distribution de ces 54 bits.

Tableau II.1. Allocation des bits dans le FS1015 [1]

Paramètres	**Résolution**	
	Trame voisée	**Trame Non voisée**
Pitch et décision de voisement	7	7
Energie	5	5
LPC	41 (5, 5, 5, 5, 4, 4, 4, 4, 3, 2)	20 (5, 5, 5, 5, 0, 0, 0, 0, 0, 0)
Synchronisation	1	1
Protection	-	21
Total	**54**	**54**

II.3. Le vocodeur CELP

Atal et Schroeder ont joué sur l'excitation du codeur LPC pour élaborer le codeur CELP (**C**ode **E**xcitated **L**inear **P**rediction) [31]. Ils ont utilisé un dictionnaire regroupant dans ses lignes des séquences d'excitation pour le filtre de synthèse dont à chaque fois une ligne est choisie et validée par la technique d'analyse par synthèse.

En plus de l'excitation par code et l'analyse en boucle fermée, le CELP utilise un post-filtre pour améliorer la qualité perçue du signal synthétique.

II.3.1. L'analyse par synthèse

Dans la méthode d'analyse en boucle ouverte utilisée par la plupart des codeurs paramétriques, le signal de parole est analysé et les résultats sont transmis après leur quantification directement au décodeur pour la synthèse.

La technique d'analyse par synthèse (figure II.3) travaille en boucle fermée où toutes les valeurs possibles d'un paramètre sont passées dans le filtre de synthèse dont la qualité de la parole synthétique produite, détermine la valeur à retenir. D'où le nom analyse par synthèse [38].

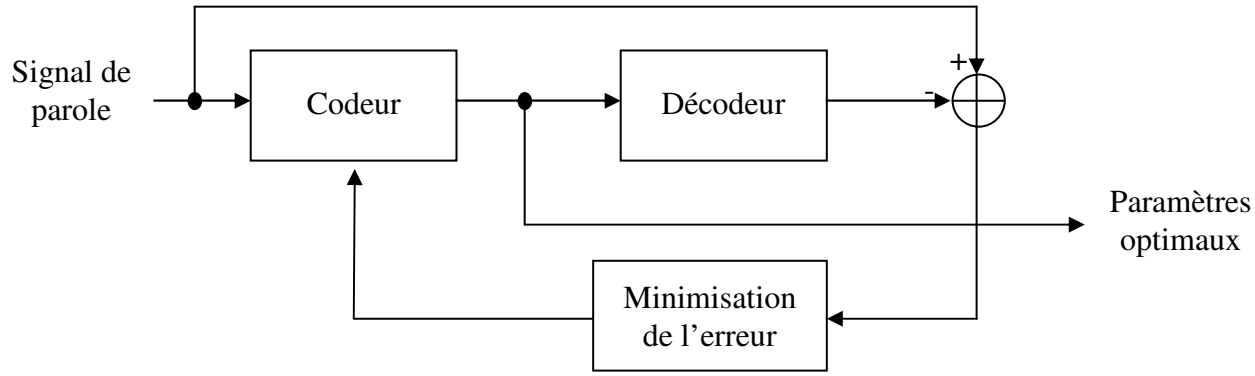


Fig II.3. Principe de l'encodage par la méthode d'analyse par synthèse [1]

Théoriquement, tous les paramètres peuvent être optimisés par cette technique mais, en pratique, et à cause de la capacité mémoire et du temps de calcul demandés, incompatibles avec les contraintes d'une application en temps réel comme le codage de la parole , seulement certains paramètres sont optimisés par l'analyse par synthèse. Les autres sont calculés par la méthode d'analyse en boucle ouverte [1]. Pour le CELP, la boucle fermée est utilisée pour choisir une séquence d'excitation et le reste des paramètres sont calculés en boucle ouverte (figure II.4).

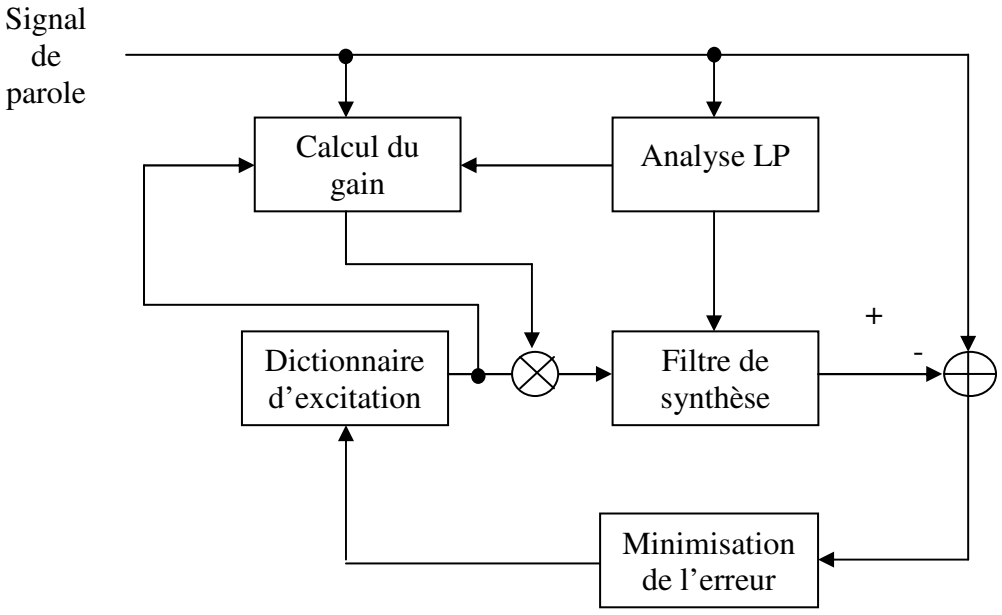


Fig II.4. Calcul de l'excitation CELP [1]

II.3.2.Filtrage perceptuel

Le phénomène de masquage (chapitre I) dans notre système d'audition peut être exploité pour mesurer l'erreur dans l'analyse par synthèse CELP [1]. On sait que dans les régions de fréquence où la parole est plus énergétique, les bruits sont pratiquement inaudibles (ils sont masqués par la parole). Le principe du filtrage perceptuel (figure II.5) consiste à donner plus d'importance à la mesure de l'erreur dans les zones les moins énergétiques, la où elle est intolérable car les bruits engendrés ne seront pas masqués par la parole.

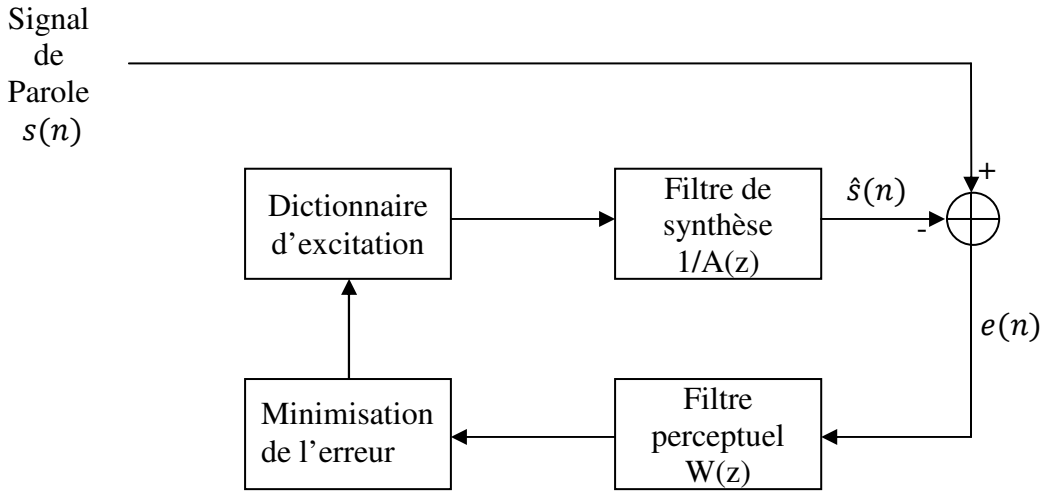


Fig II.5. Analyse par synthèse avec filtrage perceptuel

La fonction de transfert du filtre perceptuel est donnée par [1, 5, 38]:

$$W(z) = \frac{A(Z)}{A(Z/\gamma)} = \frac{1+\sum_{i=1}^{M} a_i Z^{-i}}{1+\sum_{i=1}^{M} a_i \gamma^i Z^{-i}} \qquad \text{(II.2)}$$

La constante γ est choisie entre 0 et 1. Elle permet le contrôle du filtrage.

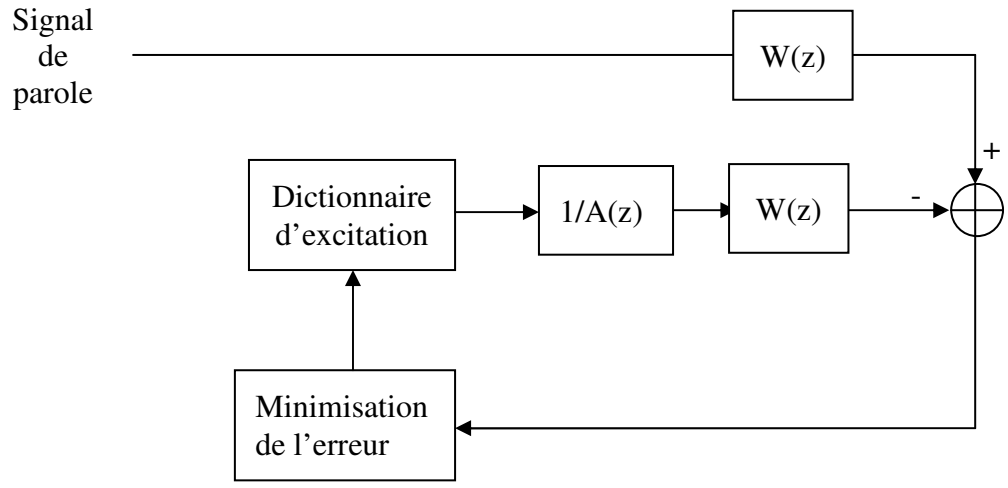
- si $\gamma \to 1 \ alors \ W(z) \to 1$: aucune modification ne sera apportée au signal d'erreur.
- si $\gamma \to 0 \ alors \ W(z) \to A(z)$: on retrouve ainsi le filtre d'analyse.

La constante γ introduit un effet d'élargissement de la bande sur le signal d'erreur filtré. Il n'y a pas de méthode analytique pour calculer la valeur de cette constante, mais en général on fait recours à des tests d'écoute (subjectifs) pour la déterminer. Pour le CELP, elle est choisie entre 0.8 et 0.9 [1, 5].

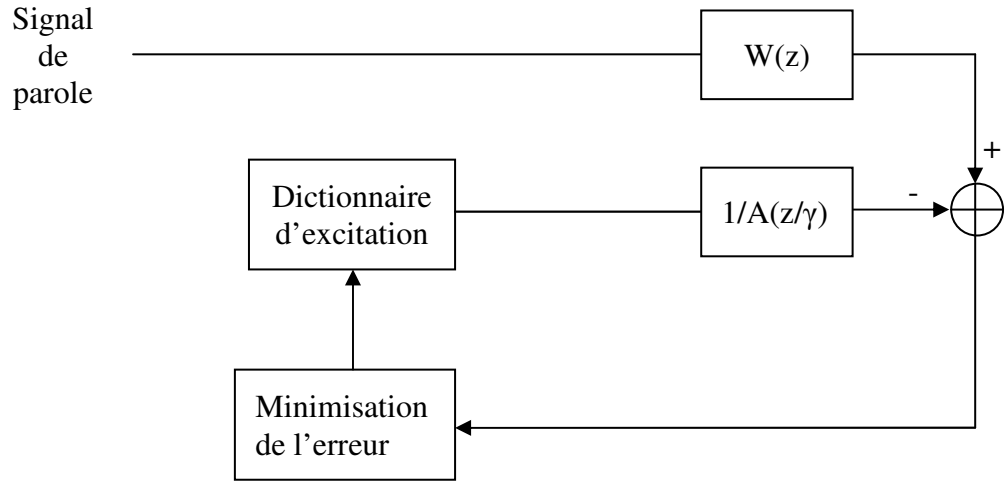
Le filtre perceptuel amplifie le spectre du signal d'erreur dans les zones non formantiques du signal de parole. Il l'atténue dans les zones formantiques.

En profitant de la linéarité du filtrage, des transformations peuvent être faites (figure II.6 (a)) pour avoir le système donné par la figure (figure II.6 (b)). La mise en cascade du filtre de synthèse et du filtre perceptuel donne ce qu'on appelle filtre de synthèse modifié dont la fonction de transfert est [1]:

$$H_f\left(\frac{Z}{\gamma}\right) = \frac{1}{A\left(\frac{Z}{\gamma}\right)} = \frac{1}{1+\sum_{i=1}^{M} a_i \gamma^i Z^{-i}} \qquad \text{(II.3)}$$



(a)



(b)

Fig II.6 filtrage perceptuel a) modifié
b) modifié et simplifié

II.3. 3. Post-filtrage

Nous avons vu dans la section précédente que la qualité subjective du codeur CELP peut être améliorée par l'utilisation d'un filtre perceptuel dans la boucle de sélection de la séquence d'excitation. Ce filtre va réduire la composante bruit au niveau des anti-formants (les vallées de la fonction de transfert) et l'accroitre au niveau des formants (les pics). Pour l'atténuer dans ces zones un post-filtrage est utilisé [38].

La réponse fréquentielle d'un post-filtre idéal devrait suivre les pics et les vallées de l'enveloppe spectrale du signal de parole sans le modifier [1]. On sait que la réponse fréquentielle du filtre de synthèse dans les codeurs utilisant la prédiction linéaire suit parfaitement l'enveloppe spectrale de la parole. Aussi, le post-filtre sera un dérivé de la prédiction linéaire. Selon la complexité d'implémentation tolérée et la qualité recherchée, plusieurs modèles ont été proposés pour ce dernier [1]. Pour le codeur CELP, on utilise un post-filtrage dit à court terme. Donc on considère un filtre de synthèse LP à bande élargie comme post-filtre [1] :

$$H_1(Z) = \frac{1}{1+\sum_{i=1}^{M} a_i\, \alpha^i Z^{-i}} \tag{II.4}$$

où $0 < \alpha < 1$ est une constante. La figure II.7 montre la réponse fréquentielle de ce filtre pour différentes valeurs de α . Ce post-filtre peut réduire le niveau de bruit dans le spectre du signal synthétique mais cette réduction est accompagnée par l'étouffement du signal à cause du filtrage passe bas apporté par ce post-filtre [1, 38].

Pour réduire l'effet spectral de ce post-filtre tout-pôle, on lui ajoute des zéros. La fonction de transfert du nouveau post-filtre devient [1]:

$$H_2(z) = \frac{1+\sum_{i=1}^{M} a_i\, \beta^i\, Z^{-i}}{1+\sum_{i=1}^{M} a_i\, \alpha^i Z^{-i}} \tag{II.5}$$

Avec $0 < \beta < \alpha < 1$. L'amplitude de la réponse fréquentielle en échelle logarithmique est donnée par :

$$Amp = 20\log\left(\frac{1}{\left|1+\sum_{i=1}^{M} a_i\, \alpha^i Z^{-i}\right|}\right) - 20\log\left(\frac{1}{\left|1+\sum_{i=1}^{M} a_i\, \beta^i Z^{-i}\right|}\right) \tag{II.6}$$

C'est la différence entre les réponses fréquentielles de deux filtres de synthèse LP à bande élargie. Les valeurs de α et β sont déterminées à l'aide de tests d'écoute. Si on choisit $\alpha = 0.8$ et $\beta = 0.5$, on réduit considérablement l'étouffement du signal synthétique (figure II.8).

La réponse fréquentielle du post-filtre sera davantage améliorée et l'effet du filtrage passe bas sera encore mieux réduit si on ajoute un filtre du premier ordre dont la fonction de transfert est : $(1 - \mu Z^{-1})$ [1, 38], en cascade avec notre post-filtre, ce qui nous donne le filtre suivant :

$$H_3(z) = (1 - \mu Z^{-1}) \frac{1+\sum_{i=1}^{M} a_i \beta^i Z^{-i}}{1+\sum_{i=1}^{M} a_i \alpha^i Z^{-i}} \qquad \text{(II.7)}$$

Une valeur de μ=0.5 donne des bons résultats [3] (figure II.8).

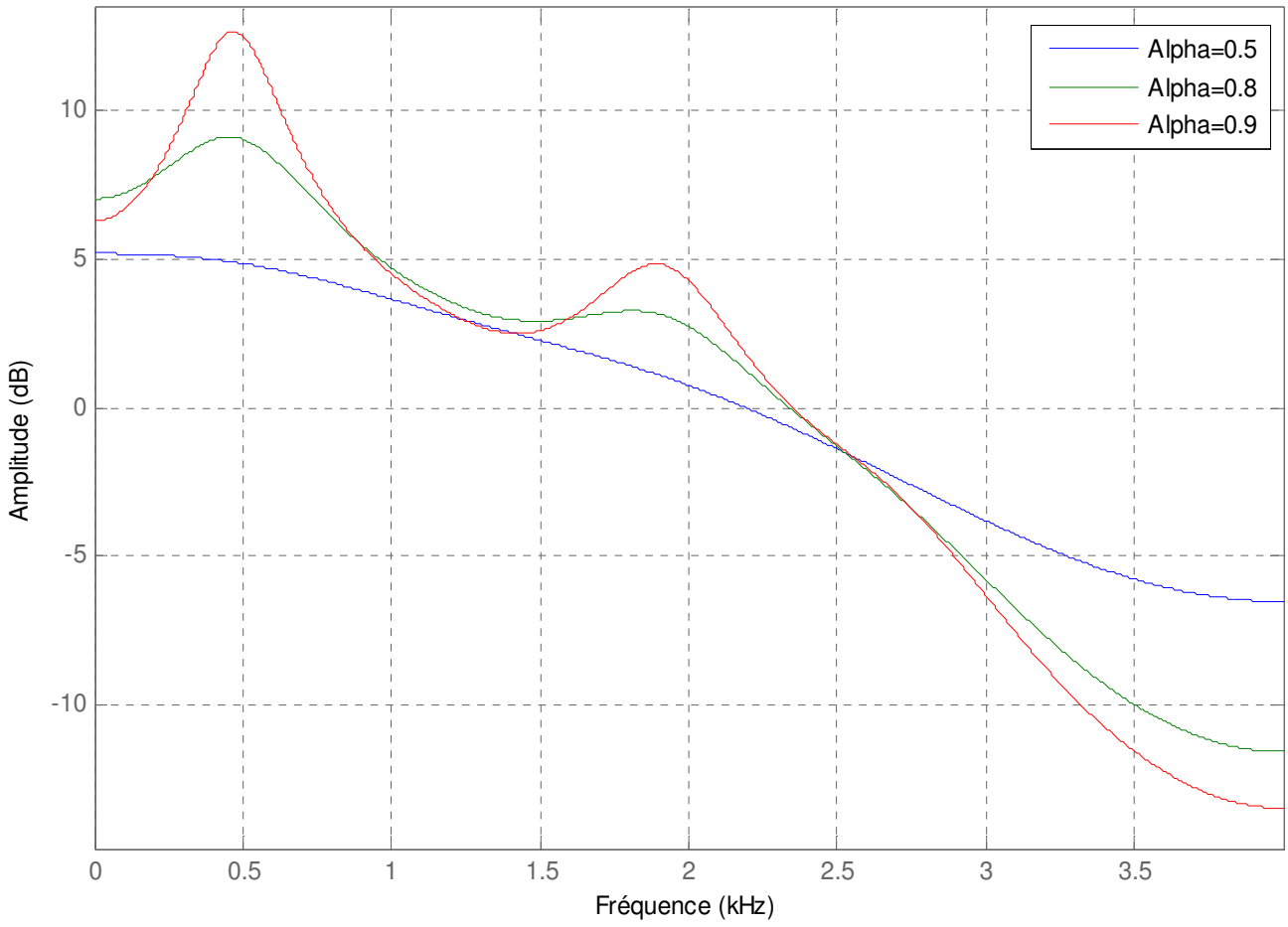


Fig II.7. Réponse fréquentielle du post-filtre pour différentes valeurs de α

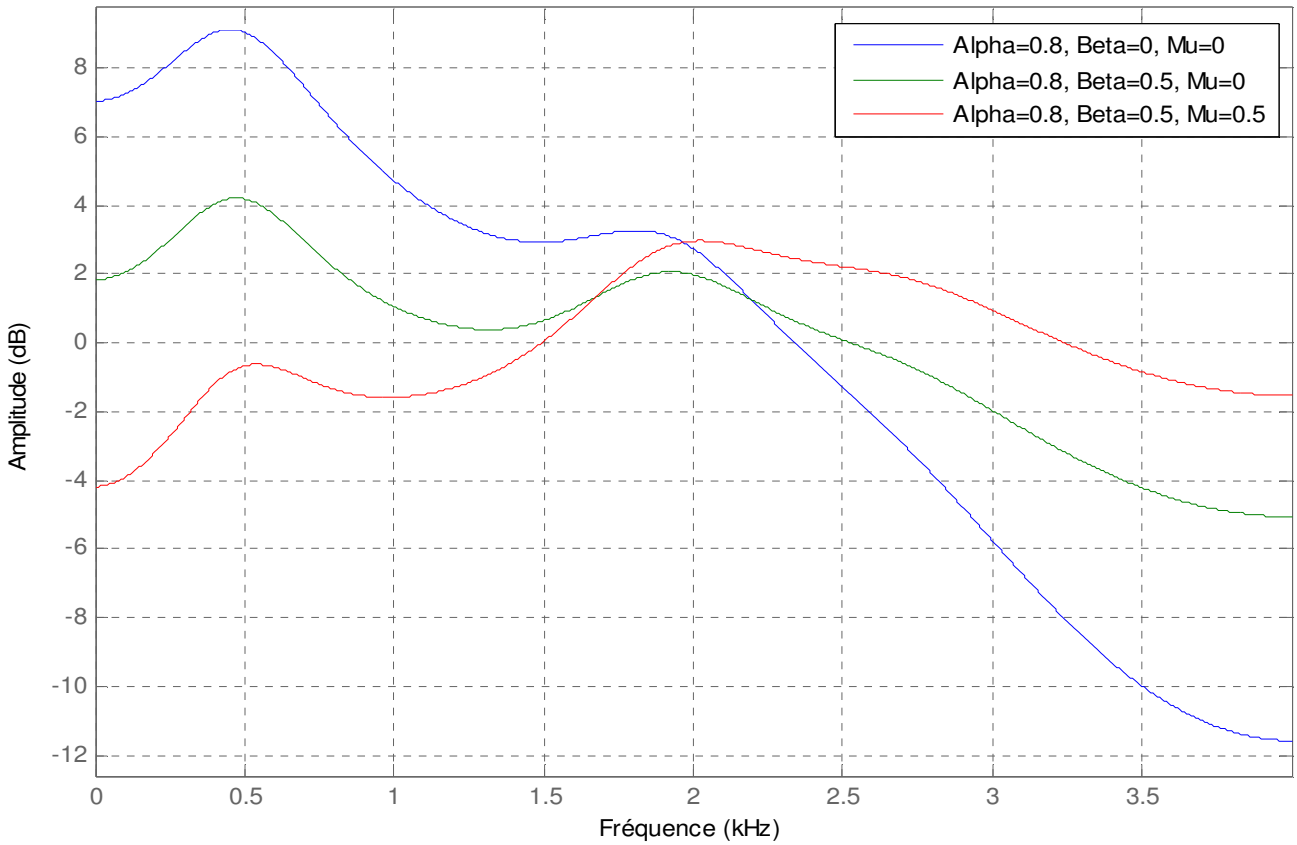


Fig II.8. Réponse fréquentielle pour les trois post-filtres

Enfin on obtient le diagramme du post-filtrage donné par la figure II.9. Pour neutraliser l'effet du post-filtrage sur l'énergie du signal synthétique, on multiplie la sortie du post-filtre par un gain. Ce gain est le rapport d'énergie du signal bruité et celle du signal filtré [1].

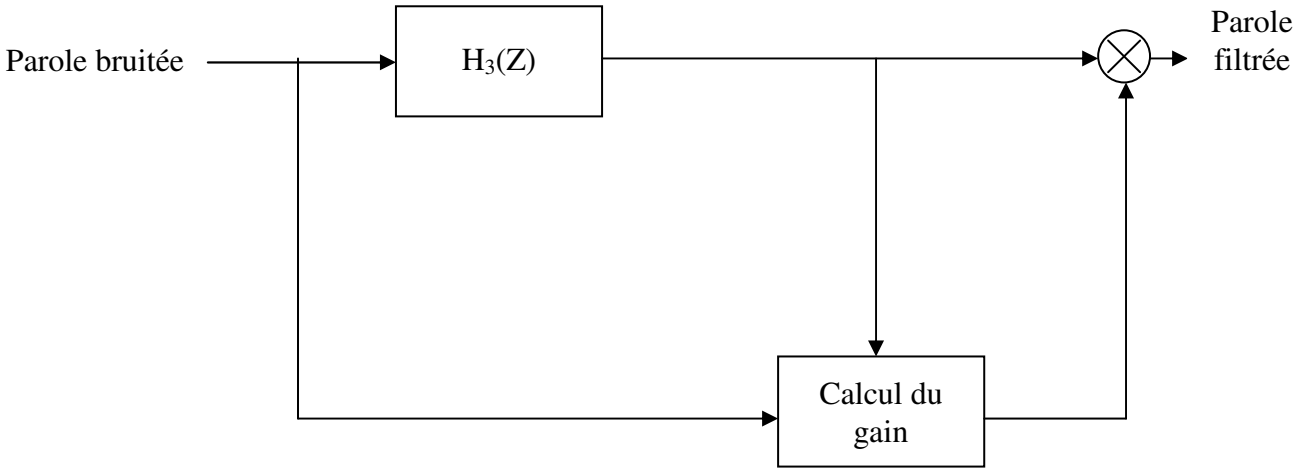


Fig II.9. Diagramme du post-filtre [1]

II.3.3. Format de transmission

En plus de la quantification vectorielle utilisée pour quantifier le signal d'excitation, une quantification scalaire non uniforme est utilisée pour le reste des paramètres. Le standard FS1016 développé à base du CELP travaille à 4.8kbps et utilise des trames de 30 ms de longueur. L'encodeur doit transmettre au décodeur 144 bits pour chaque trame [1]. La distribution de ces 144 bits est donnée par le tableau II.2.

Tableau II.2. Allocation des bits pour le FS1016 [1]

Paramètre	**Nombre de paramètres par trame**	**Résolution**	**Nombre de bits par trame**
LPC	10	3,4,4,4,4,3,3,3,3,3	34
Pitch	4	8, 6, 8, 6	28
Gain adaptative	4	5	20
Indice d'excitation	4	9	36
Gain stochastique	4	5	20
synchronisation	1	1	1
protection	4	1	4
Expansion	1	1	1
		Total	**144**

II.4. Le vocodeur MELP

Le vocodeur MELP (***M**ixed **E**xcitation **L**inear **P**rediction*) 2.4kbps, développé initialement par McCree et Barnwell en 1995 [39] est une version améliorée du codeur LPC. Les améliorations apportées par ce dernier au codage prédictif peuvent être résumées par les trois points suivant [1]:

- Une excitation mixte.
- Une analyse de voisement plus précise.
- Un filtre de dispersion utilisé en position de post-filtre.

II.4.1. L'excitation mixte

L'idée d'utiliser une excitation mixte pour le filtre de synthèse est apparue dans les travaux de Makhoul et al [16], avant même l'apparition du vocodeur MELP. Ces auteurs ont conclu que l'addition d'un bruit de haute fréquence à la composante périodique réduit la métallisation dans le signal synthétique. McCree et son équipe ont aussi prouvé ce résultat avec un model d'excitation mixte différent : un model multi-bande avec une intensité de voisement définie pour chaque bande de fréquence [17].

Une analyse de voisement multi-bande (figure II.10) doit être effectuée au niveau de l'encodeur où le signal de parole passe par cinq filtres passe-bande adjacents. Cette étape va définir une intensité de voisement pour chaque sous-bande de fréquence et permet de la classer comme sous-bande voisée ou non-voisée. Les cinq bandes passantes des filtres d'analyse sont : [0 500] Hz, [500 1000] Hz, [1000 2000] Hz, [2000 3000] Hz et [3000 4000] Hz (figure II.11).

Au niveau du décodeur, la composante périodique et la séquence du bruit vont être filtrées par des filtres de mise en forme avant d'effectuer leur somme. Ces filtres sont une somme pondérée des filtres passe-bande dont la pondération dépend de l'intensité de voisement transmise par l'encodeur pour chaque sous-bande.

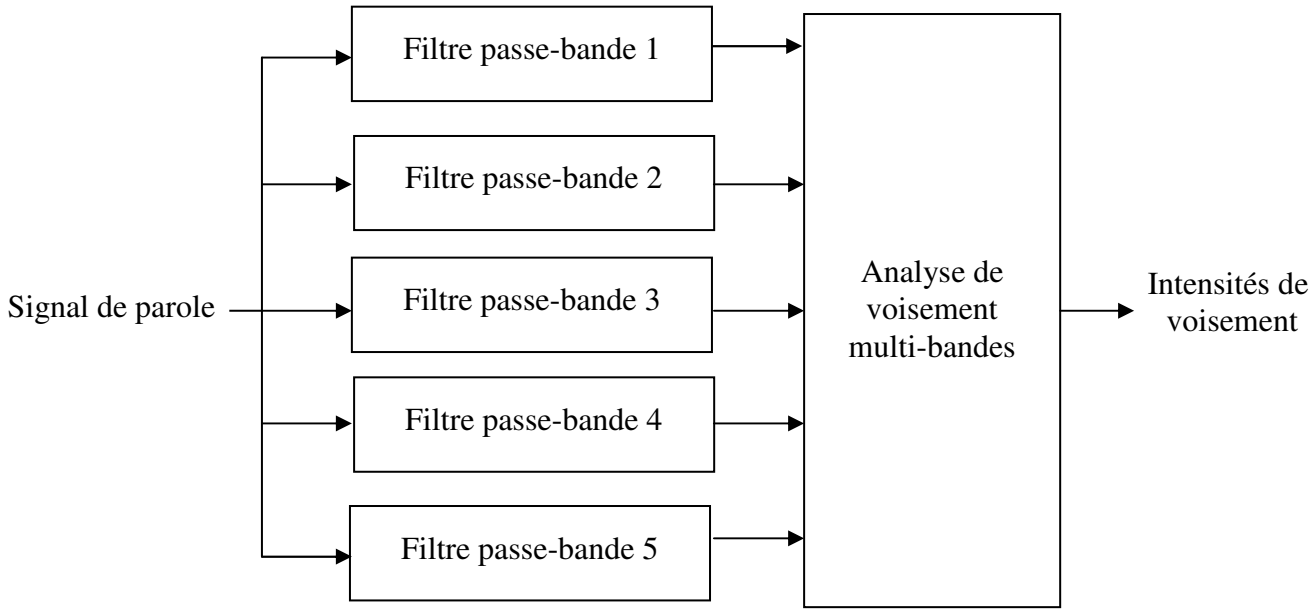


Fig II.10. Analyse de voisement MELP

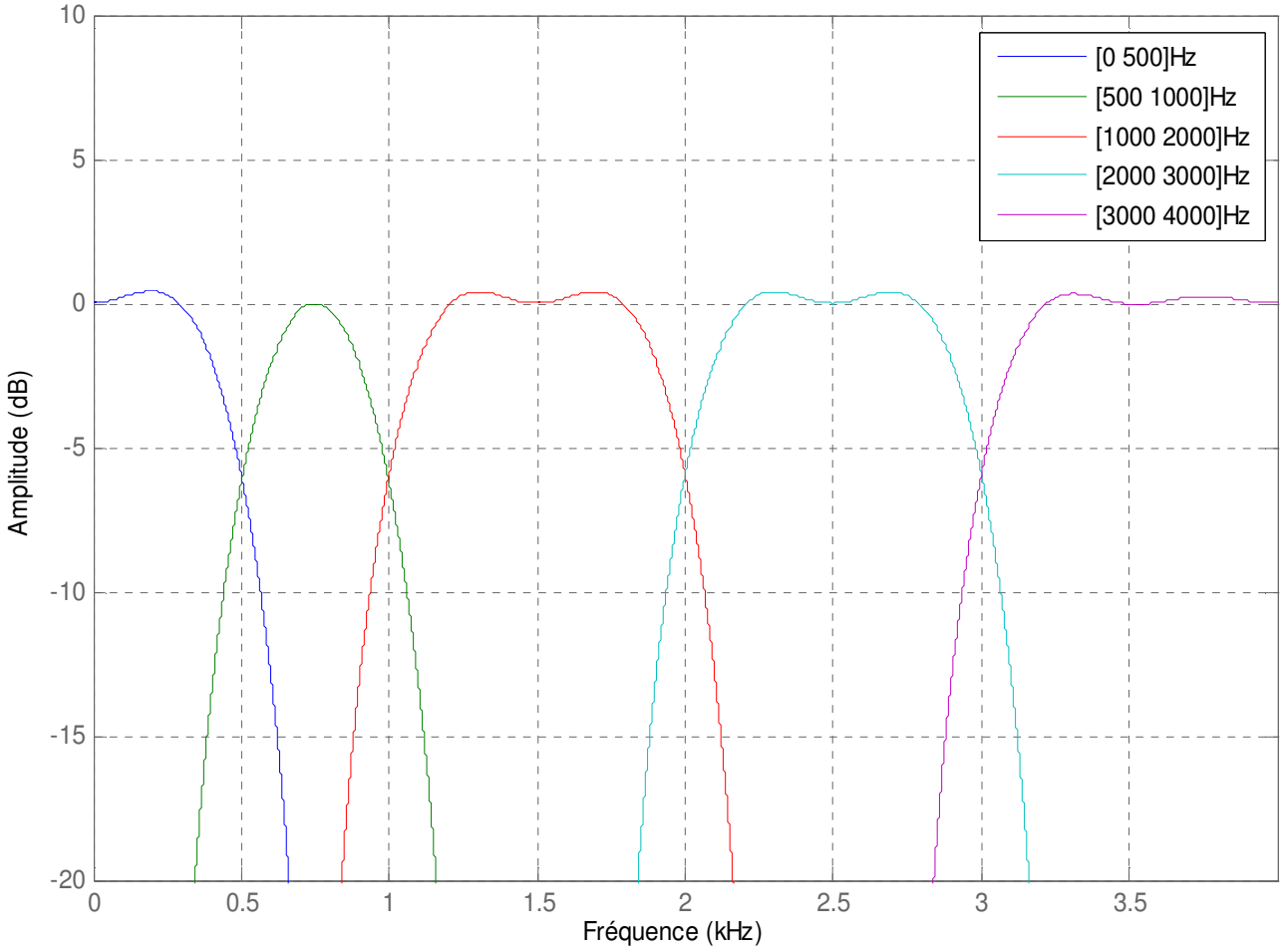


Fig II.11. Filtre passe bande de l'analyse de voisement

La première version du MELP utilise un train d'impulsions pour générer la composante périodique du signal d'excitation [39], mais des versions ultérieures utilisent des amplitudes, correspondant au pitch et ses harmoniques, dans le spectre de l'erreur de prédiction pour générer un signal d'excitation périodique plus naturel [34].

II.4.2. Analyse de voisement

Les trames du signal de parole dans le vocodeur LPC sont considérées soit voisées soit non voisées. Cette classification étroite est à l' origine des tonalités courtes et isolées entendues souvent dans le signal synthétique en particulier des sons aigus (femmes et enfant surtout) [1]. Certains chercheurs ont proposé d'ajouter un bruit de basse fréquence au signal d'excitation pour éliminer ces tonalités. Mais avec l'ajout de ce bruit (en plus du bruit de

l'excitation), la sortie devient trop bruitée [17]. Une autre solution plus efficace consiste à perturber la périodicité de certaines trames voisées (excitation apériodique) par la variation non uniforme de la période (pitch) jusqu' à 25% pour simuler les mouvements glottiques dans les zones de transition [17]. Cette solution nécessite la définition d'une classe intermédiaire entre les classes voisées et non-voisées. C'est la classe des trames dites apériodiques et qui représente les zones de transition. Elle peut être considérée comme une sous-classe dans la classe des sons voisés caractérisée par auto-corrélation moins importante [1].

L'analyse de voisement du MELP est basée sur l'auto-corrélation mais aussi sur une grandeur appelée ' peakiness' définie par [1, 17]:

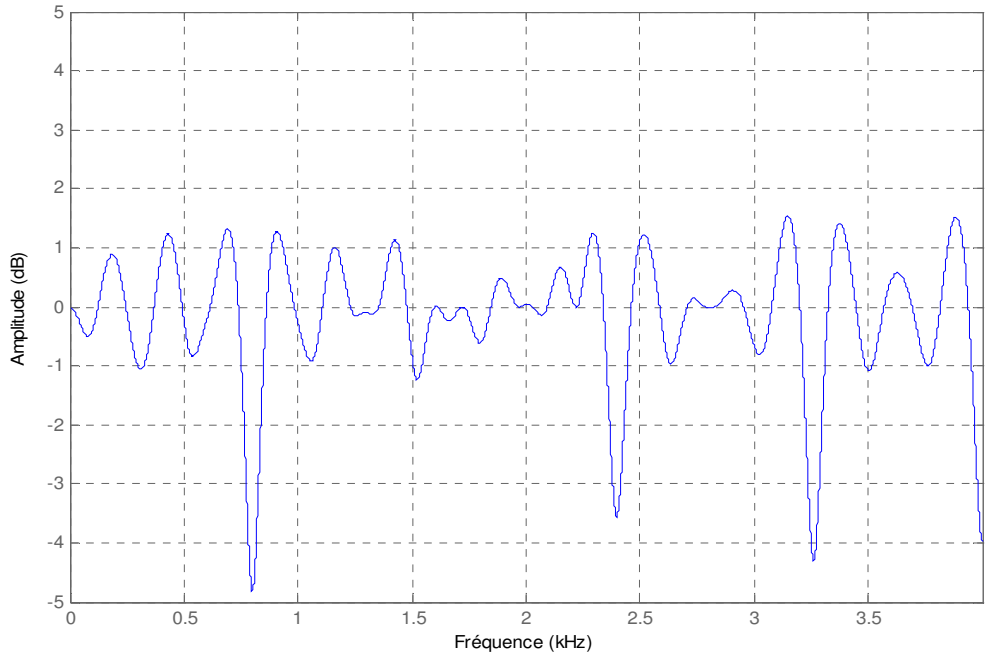
$$peakiness = \frac{\sqrt{\sum x_n^2}}{\sum \|x_n\|} \tag{II.9}$$

Cette grandeur représente le rapport de la norme L2 sur la norme L1 du signal d'erreur. Plus les échantillons sont distants de la moyenne du signal, plus cette grandeur est importante.

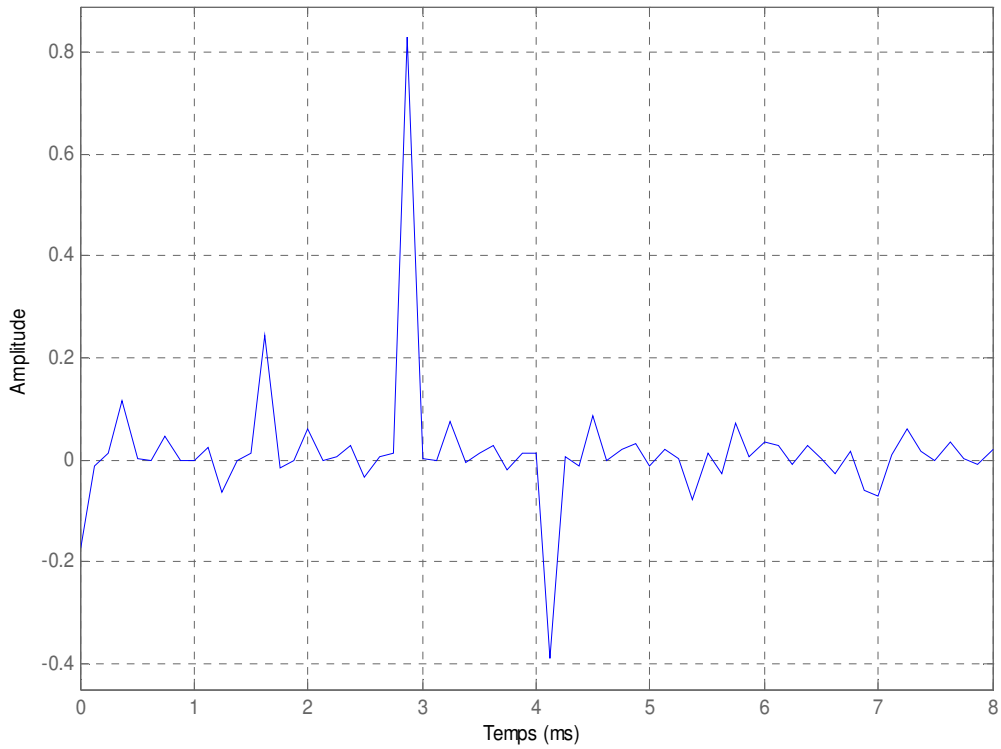
II.4.3. Filtre de dispersion

C'est le dernier filtre dans le décodeur MELP. Son rôle est d'étaler l'énergie des impulsions sur une période de pitch. Ceci améliore la similitude entre le signal synthétique et la parole naturelle dans les zones non formantiques où la sortie du filtre de synthèse LP prend des valeurs très faibles entre deux impulsions de pitch, alors que dans le cas réel, la fermeture de la glotte est incomplète, ce qui permet le passage d'air. Il y'a aussi les bruits de fond qui deviennent visibles entre les pics de pitch [17].

Le filtre de dispersion est un filtre à réponse impulsionnelle finie (FIR) d'ordre 65, basé sur une impulsion glottique synthétique à spectre plat. Ses coefficients sont générés par la transformée de Fourier d'une impulsion triangulaire unitaire [1, 17]. La figure II.12 illustre sa réponse impultionnelle et sa réponse fréquentielle.



(a)



(b)

Fig II.12. Filtre de dispersion : a) Réponse fréquentielle
b) Réponse impulsionnelle

II.4.4. Format de transmission

La version standardisée du MELP fonctionne à un débit de 2.4 kbps. Elle utilise des trames d'analyse de 22.5 ms, ce qui donne 54 bits pour chaque trame [34]. Les paramètres fournis par l'analyse MELP ainsi que l'allocation des 54 bits et le type de quantification utilisé pour chaque paramètre sont donnés par le tableau II.3

Tableau II.3. Allocation des bits pour le FS 1017 [34]

Paramètres	**Trame voisée**	**Trame non voisée**	**quantification**
Coefficients LSF	25	25	MSVQ (4 étages 7 6 6 6)
Coefficients de Fourier	8	-	VQ
Gains (deux par trame)	8	8	VQ
Fréquence fondamentale	7	7	SQ
Bande de voisement	4	-	SQ
Indicateur de voisement	1	-	-
Protection	-	13	-
Bit de synchronisation	1	1	-
Total	**54**	**54**	**-**

II.5. Comparaison de la qualité des standards FS1015, FS1016 et FS1017

Le tableau II.4 présente la qualité en MOS des trois standards. La qualité du codeur MELP est proche du celle du CELP avec un débit de moitié et une complexité plus réduite où on utilise une analyse en boucle ouverte. Avec le même débit et presque la même complexité, le MELP présente une qualité meilleure que le FS1015.

Tableau II.4. Qualité MOS des standards DoD

Standard	**Année**	**Algorithme**	**Débit (Kbs)**	**Qualité MOS**
FS1015	1984	LPC10	2.4	2.3
FS1016	1991	CELP	4.8	3.5
FS1017	1997	MELP	2.4	3.2

II.6. Conclusion

Dans ce chapitre nous avons étudié trois algorithmes de codage prédictif. Le premier et le plus ancien est le LPC. Il est le plus simple de point de vue débit et complexité, mais il est aussi très limité dans sa qualité. Avec le CELP, on a amélioré considérablement la qualité mais au profit d'un débit plus important et un algorithme plus complexe et plus lourd en temps de calcul à cause de l'analyse en boucle fermée utilisée. Le MELP, le plus récent des trois, est moins complexe que le CELP. Son débit est plus réduit et sa qualité est légèrement différente. Elle est cependant meilleure que celle du LPC.

Ces avantages justifient notre choix porté sur l'algorithme d'analyse MELP pour mettre au point un codeur à très bas débit fonctionnant à 1.2 kbps. Nous l'avons mis au point pour l'utiliser dans différents travaux de recherche liés au codage de la parole à très bas débit dans notre laboratoire. La mise en œuvre de ce codeur basé sur l'algorithme MELP sera détaillée dans le chapitre suivant.

CHAPITRE

3

MISE EN ŒUVRE D'UN CODEUR A TRES BAS DEBIT

III.1. Introduction

Depuis son apparition au milieu des années 90 [39], le MELP a fait l'objet de plusieurs travaux de recherche dans le but d'aller plus loin dans la réduction de débit en jouant essentiellement sur la quantification des coefficients LP. A ce jour, on trouve plusieurs versions de ce codeur. Citons le MELP 2.4 kbps [40], le MELP 1.7 kbps [41], le MELP 1.6 kbps [42], le MELP 1.2 kbps [43] et même le MELP 0.6 kbps [44, 45]. Malheureusement, toutes ces versions n'ont pas pu préserver la même qualité que le FS-MELP 2.4 kbps. C'est la raison pour laquelle ce dernier n'a pas été remplacé par une autre version depuis sa standardisation par le département de la défense américaine (DoD FS1017) en 1997.

Ce chapitre présente notre contribution à la mise en place sous MATLAB d'un codeur MELP multi-trames travaillant à 1.2 kbps et utilisant le même algorithme d'analyse que le FS-MELP 2.4 kbps mais avec des algorithmes de quantification différents.

III.2. Structure du codeur mis au point

Pour faciliter l'implémentation de notre codeur, son évaluation ou sa modification ultérieur, nous avons divisé son algorithme en deux modules : encodeur et décodeur, chacun contenant deux sous-modules comme indiqué à la figure III.1.

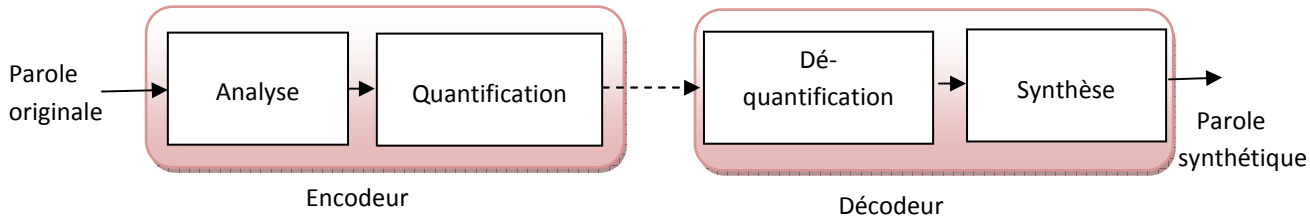


Fig.III.1 : structure de l'algorithme MELP mis au point

Le sous-module d'analyse : son rôle est d'analyser le signal de parole et d'extraire les paramètres nécessaires à la reconstitution du signal synthétique.

Le sous-module de quantification : ce module convertit les sorties décimales du sous-module d'analyse en une combinaison de bits (binaires) selon le débit souhaité.

Le sous-module de dé-quantification : il reçoit les paramètres venant de l'encodeur sous format binaire et les convertit en nombres décimaux.

Le sous- module de synthèse : c'est le dernier bloc du vocodeur mis au point. Il utilise les paramètres dé-quantifiés pour produire un signal de parole synthétique.

III.3. Implémentation du module d'encodage

La figure III.2 illustre le module d'encodage MELP implémenté. Dans ce module il y'a deux tâches à réaliser : l'analyse du signal pour l'extraction des paramètres et la quantification de ces paramètres.

III.3.1. Sous-module d'analyse mis en place

Nous avons mis en œuvre ce sous-module pour l'extraction des paramètres : les coefficients LP, le pitch, les intensités de voisement des cinq bandes de fréquence, l'indice d'apériodicité, les deux gains et les amplitudes de Fourier.

A. Prétraitement effectué sur le signal de parole

Avant tout traitement, le signal de parole doit être filtré afin d'enlever la composante continue du signal. Pour ce faire, nous avons utilisé un filtre passe-haut de Chebyshev de quatrième ordre. Ce filtre de fréquence de coupure 60Hz (figure III.3) va éliminer la

composante basse-fréquence du signal de parole. Cette dernière est inaudible mais elle est très gênante pour l'analyse [39].

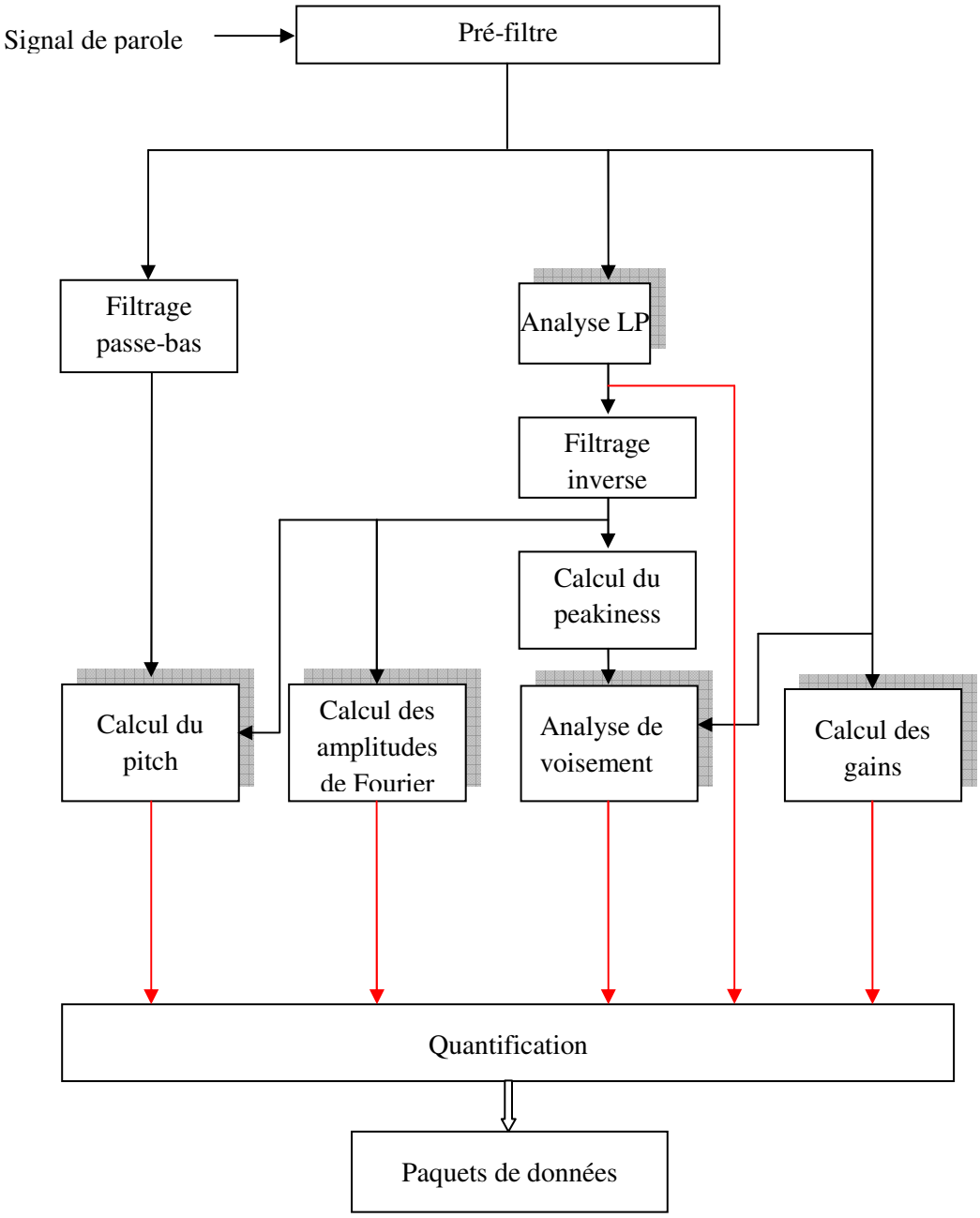


Fig III.2. Encodeur MELP mis au point

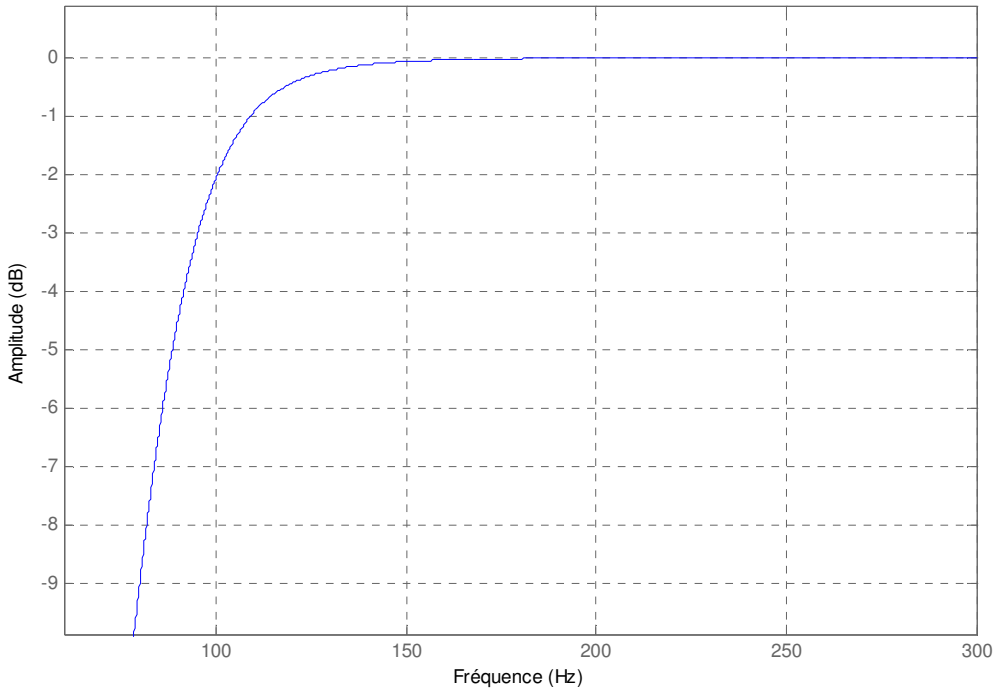


Fig III.3 : réponse fréquentielle du pré-filtre

B. Positions des fenêtres d'analyse utilisées

Le MELP traite des signaux de parole échantillonnés à 8Khz. D'abord il les divise en segments de 22.5 ms. Donc chaque segment (ou trame) doit avoir 180 échantillons. Pour analyser une trame, on doit tenir compte des deux trames voisines : la précédente et la suivante [1]. La figure III.4 montre les positions et les longueurs des fenêtres d'analyse pour chaque paramètre. Les positions de ces fenêtres sont choisies d'une manière à faciliter l'interpolation des paramètres par la suite. Certains fenêtres sont recouvrées d'autres non et la plupart de ces fenêtres (sauf pour le gain) sont centrées sur la fin de la trame en cours d'analyse.

C. L'analyse LP effectuée

Pour calculer les coefficients LP, nous avons implémenté un algorithme d'analyse par prédiction linéaire d'ordre 10. Notre programme utilise des fenêtres de 200 échantillons centrées sur la fin de la trame. On utilise 100 échantillons de la trame courante et 100 autres échantillons de la trame suivante, soit un recouvrement entre les fenêtres de 20 échantillons ou 10% (voir figure III.4). Avant de procéder au calcul des coefficients LPC une fenêtre de Hamming est appliquée sur les 200 échantillons.

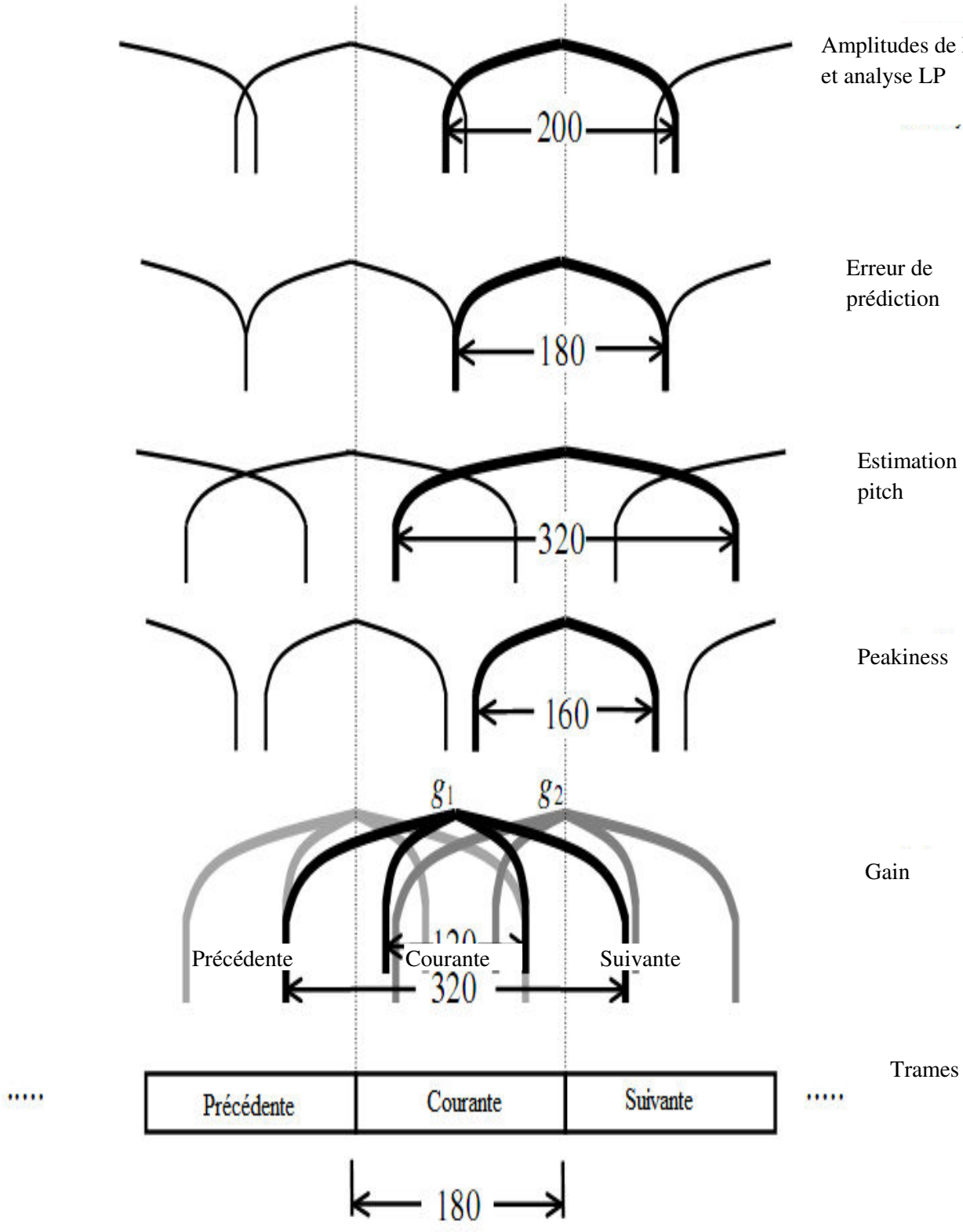


Fig III.4 : les fenêtres d'analyse MELP [1]

D. Estimation du pitch

L'algorithme que nous avons choisi pour l'estimation du pitch est basé sur l'auto-corrélation du signal de parole. Cette estimation se fait en deux étapes : la première pour calculer la valeur entière du pitch et la seconde pour raffiner cette valeur et avoir un pitch fractionnel. Avant l'estimation, le signal de parole est filtré avec un filtre passe bas de Butterworth de sixième ordre et de fréquence de coupure fc=1 KHz (figure III.5). Nous avons utilisé ce filtrage pour limiter la bande de recherche du pitch en éliminant les hautes fréquences (cela nous évitera de faire des recherches inutiles et nous permettra de simplifier le programme de la détection).

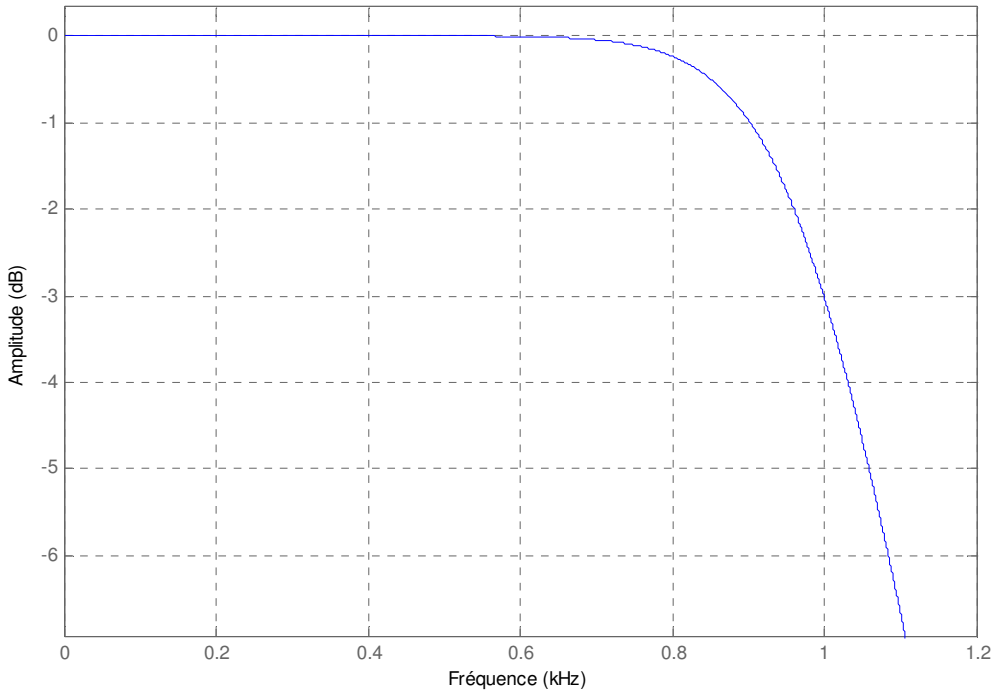


Fig III.5 : Filtre de Buterworht de 6eme ordre à fc= 1Khz

La valeur entière du pitch correspond au maximum de la fonction d'auto-corrélation normalisée du signal filtré [5] :

$$r(\tau) = \frac{c_\tau(0,\tau)}{\sqrt{c_\tau(0,0)c_\tau(\tau,\tau)}} \qquad \text{(III.1)}$$

Avec :

$$c_\tau(m,n) = \sum_{k=-\lfloor \tau/2 \rfloor -80}^{\lfloor \tau/2 \rfloor +79} s_{k+m} s_{k+n} \qquad \text{(III.2)}$$

C_τ est la corrélation croisée et τ est le retard en nombre d'échantillons.

Le pitch est une grandeur fractionnelle. Pour affiner les valeurs du pitch et les rendre fractionnelles, nous avons utilisé l'algorithme développé par Medan [46] pour calculer la correction Δ à apporter à la fonction de corrélation normalisée :

$$\Delta = \frac{c_T(0,T+1)c_T(T,T) + c_T(0,T)c_T(T,T+1)}{c_T(0,T+1)\left[c_T(T,T) - c_T(T,T+1)\right] + c_T(0,T)\left[c_T(T+1,T+1) - c_T(T,T+1)\right]} \qquad \text{(III.3)}$$

Où T est la valeur entière de pitch.

La corrélation normalisée correspondant à cette valeur interpolée du pitch est donnée par :

$$r(T+\Delta) = \frac{(1-\Delta)c_T(0,T) + \Delta c_T(0,T+1)}{\sqrt{c_T(0,0)\left[(1-\Delta)^2 c_T(T,T) + 2\Delta(1-\Delta)c_T(T,T+1) + \Delta^2 c_T(T+1,T+1)\right]}} \qquad \text{(III.4)}$$

Cette valeur de corrélation va aussi être utilisée dans l'analyse de voisement.

Notre choix a porté sur la méthode d'auto-corrélation normalisée à cause des problèmes posés par la dynamique de l'amplitude de l'auto-corrélation ordinaire liée à l'énergie du signal. En effet, la normalisation permet de simplifier le seuillage.

E. Analyse de voisement effectuée

Le MELP utilise une analyse multi-bandes où l'intensité de voisement est mesurée dans cinq bandes adjacentes. Pour séparer ces bandes, nous avons utilisé cinq filtres passe-bande de Butterworth d'ordre six : [0,500] Hz, [500,1000] Hz, [1000,2000] Hz, [2000,3000] Hz, [3000,4000] Hz). L'intensité de voisement de chaque bande se détermine par l'auto-corrélation normalisée correspondant au pitch. La valeur entière du pitch est la même dans toutes les bandes contrairement à la valeur fractionnelle qui change d'une bande à l'autre et, par conséquence, l'intensité de voisement ou l'auto-corrélation normalisée varie également [5]. Notons que ce calcul ne s'effectue que dans le cas des trames voisées.

Nous rappelons que le 'peakiness' (chapitre II) est un fort indice pour la détection de voisement. Une grande valeur du peakiness signifie que le signal est fortement voisé. La procédure que nous avons mise en place pour le calculer consiste à :

- prendre une fenêtre du signal de parole de 160 échantillons toujours centrée sur la fin de la trame (les 80 derniers échantillons de la trame courante et les 80 premiers échantillons de la trame suivante)
- appliquer un filtrage inverse sur ce signal pour calculer le résiduel de prédiction en utilisant les coefficients LP calculés précédemment.
- appliquer la formule de calcul de peakiness (déjà vue dans le chapitre précédent) :

$$Peakiness = \frac{\sqrt{\frac{1}{160}\sum_{n=-80}^{79} e^2[n]}}{\frac{1}{160}\sum_{n=-80}^{79} |e[n]|} \qquad \text{(III.5)}$$

Les intensités de voisement (Iv) des cinq bandes sont converties en binaire car ce qui nous intéresse c'est la décision voisée ou non-voisée et non pas la valeur de l'intensité elle-même. Pour cette conversion, nous avons effectué un premier seuillage sur les intensités [47]:

Soit Iv_i l'intensité de voisement de la i^{eme} sous bande et Ivb_i sa version binaire :

- Si Iv_i>0.6 alors Ivb_i=1 si non Ivb_i=0.

Nous effectuons en suite un deuxième seuillage sur la valeur du peakiness [47]:

- Si peakiness >1.34 alors Ivb_1=1.
- Et si peakiness >1.64 alors Ivb_2=1 et Ivb_3=1.

Rappelons que le MELP classe les trames de parole en trois catégories : trames voisées, trames non voisées et trames apériodiques pour désigner les zones de transition, selon l'intensité de voisement. Si cette dernière est supérieure à 0.6, la trame est voisée, et si elle est inferieure à 0.5, la trame est non voisée. Hors de ces deux cas, la trame est considérée comme apériodique.

F. Correction du pitch estimé

Pour améliorer davantage la détermination du pitch, nous avons mis en place un algorithme de correction qui consiste à recalculer le pitch mais en utilisant le signal d'erreur de prédiction. Nous avons ensuite comparé les résultats obtenus sur les deux signaux : parole et erreur de prédiction. En effet, l'utilisation de l'erreur de prédiction peut offrir plus de précision pour le calcul de pitch car le filtrage LP inverse élimine la structure formantique du signal de parole [1]. Ainsi, l'intensité de voisement la plus grande déterminera le pitch qu'il faut retenir.

G. Calcul du gain

Le gain (ou l'énergie du signal) doit être mis à jour deux fois par trame. Pour cela, nous avons utilisé deux fenêtres adaptatives identiques dont la largeur dépend du pitch estimé. La première fenêtre est centrée sur le milieu de la trame et la deuxième sur sa fin. La procédure que nous avons mise en place pour l'adaptation des largeurs est la suivante [1]:

- Si la trame est voisée (proprement voisée), la largeur sera le plus petit multiple supérieur à 120 échantillons de la valeur entière du pitch. Si ce multiple dépasse les 320 échantillons, la largeur sera sa moitié.

- Si la trame est apériodique ou non voisée, on prend 120 échantillons comme largeur de fenêtre.

L'équation de calcul du gain est :

$$g = 10\log_{10}\left(0.01 + \frac{1}{N}\sum_{n=1}^{N} s^2[n]\right) \qquad \text{(III.6)}$$

N étant la largeur de la fenêtre.

H. Calcul des amplitudes de Fourier

Nous avons besoin de dix amplitudes de Fourier pour générer l'excitation périodique [1, 47]. Comme ces amplitudes sont celles correspondant au pitch et ses harmoniques dans le spectre de l'erreur de prédiction, nous avons utilisé la procédure suivante pour les calculer :

- appliquer une FFT de taille 512 sur 200 échantillons du signal de l'erreur observée par une fenêtre de Hamming, pour avoir le spectre.

- déterminer les dix premiers pics dans le spectre du résiduel. La recherche doit être effectuée autour du pitch et ses neufs premiers harmoniques.

- normaliser ces amplitudes en utilisant la norme L_2 du vecteur qui les regroupe.

Notons que lorsque la période du pitch est faible, donc la fréquence est grande, il est possible qu'on ne puisse pas avoir dix harmoniques dans la bande de Nyquist. Dans ce cas, on retient les amplitudes qui sont dans la bande et on les complète à dix avec des 1.

III.3.2. Quantification JVQ (Joint Vector Quantization) proposée

Rappelons que notre MELP 1.2 kbps utilise le même algorithme d'analyse que FS-MELP 2.4 kbps où un ensemble de paramètres doit être calculé à chaque trame de 22.5 ms. Mais, pour avoir un débit plus réduit, nous avons utilisé des super-trames de 67.5 ms. Donc, chaque super-trame regroupe trois trames de 22.5 ms.

La technique de quantification proposée est basée sur la quantification vectorielle dite conjointe (ou JVQ). Cette dernière exploite la redondance existant dans le signal de parole pour ne quantifier que les informations utiles. Lors de la création des dictionnaires, il a été tenu compte des propriétés statistiques des sons voisés et des sons non voisés (fréquence d'occurrence). Chaque super-trame est caractérisée par un mode de voisement regroupant les décisions de voisement des trois trames qui la constituent. Les 81 bits correspondant à un débit de 1.2 kbps seront distribués selon ce mode. Cette allocation des bits sera précisée par la suite.

A. Quantification du pitch

La procédure que nous avons adopté pour la quantification du pitch consiste à :

- Regrouper les trois pitch correspondant à une super trame dans un seul vecteur et mettre les pitch des trames non voisées à zéro.

- Pour les super trames avec au plus une trame voisée, quantifier ces pitch séparément en utilisant une quantification scalaire (SQ) uniforme à 99 niveaux soit à 7 bits

- Pour les super-trames avec au moins deux trame voisées, utiliser une quantification vectorielle (VQ) multi-dictionnaires pour quantifier le vecteur des pitch

Pour la VQ multi-dictionnaires nous avons utilisé cinq dictionnaires de dimension 3 et de taille 512 avec la mise en place d'un nouveau critère de sélection des vecteurs-codes qui tient compte des variations de pitch d'une trame à l'autre (il prend en considération l'évolution temporelle du pitch) en plus de la distorsion.

Au total, nous avons utilisé 12 bits pour la quantification du pitch et des décisions de voisement des trames.

Tableau III.1. Allocation des bits pour la quantification du pitch

<table>
<tr><th>Mode de voisement</th><th>3-bit</th><th>9-bit</th></tr>
<tr><td>UUU</td><td rowspan="4">000</td><td rowspan="4">• QS uniforme de 99 niveaux (7 bits) pour la trame voisée
• 2 bits pour les deux autres trames.</td></tr>
<tr><td>UUV</td></tr>
<tr><td>UVU</td></tr>
<tr><td>VUU</td></tr>
<tr><td>VVU</td><td>001</td><td rowspan="3">QV avec le même dictionnaire de dimension 512.</td></tr>
<tr><td>VUV</td><td>010</td></tr>
<tr><td>UVV</td><td>100</td></tr>
<tr><td rowspan="4">VVV</td><td>011</td><td>QV dictionnaire A 512 niveaux</td></tr>
<tr><td>101</td><td>QV dictionnaire B 512 niveaux</td></tr>
<tr><td>110</td><td>QV dictionnaire C 512 niveaux</td></tr>
<tr><td>111</td><td>QV dictionnaire D 512 niveaux</td></tr>
</table>

La JVQ du pitch utilisée est détaillée dans le tableau III.1. Sur les 12 bits retenus, nous avons utilisé 3 bits pour quantifier le mode de voisement (représentant les 8 cas possibles). Les 9 bits restants sont utilisés pour quantifier les valeurs du pitch.

Pour les cas où au moins deux trames sont voisées, la JVQ du pitch passe par trois étapes :

1ere étape :

On calcule la distance entre le vecteur de pitch et les mots codes du dictionnaire. De ces mots codes, on choisi les M candidats qui minimisent cette distance donnée par :

$$d_j = \sum_{i=1}^{3} w_i \left| p_i - \hat{p}_{j,i} \right|^2 \qquad pour \quad j = 1,...,512 \tag{III.7}$$

Avec :

$$w_i = \begin{cases} 1 \;\; si\, la \;\; trame \; est \; voisée \\ 0 \;\; si \; non \end{cases} \tag{III.8}$$

p_i et $\hat{p}_i$ sont respectivement le pitch à quantifier et celui quantifié se trouvant dans le dictionnaire.

2eme étape :

Pour les M candidats choisis dans la première étape, on calcule d'abord les variations des pitch non quantifiés sur une super-trame par l'expression :

$$\Delta p_i = \begin{cases} p_i - p_{i-1} \quad si \quad la \; i^{ème} \quad et \quad la \quad (i-1)^{ème} \quad sont \quad non-voisées \\ 0 \quad sinon \end{cases} \tag{III.9}$$

i=1, 2 ou 3.

p_0 sera le dernier pitch de la super-trame précédente.

Pour les mêmes candidats que précédemment, on calcule les variations du pitch quantifié en remplaçant Δp_i et p_i respectivement par $\Delta \hat{p}_i$ et $\hat{p}_i$. De même, $\hat{p}_0$ sera la version quantifiée de p_0.

3eme étape :

Parmi les M candidats déterminés lors la première étape, on choisit celui qui minimise le critère suivante :

$$d' = \sum_{i=1}^{3} w_i \left| p_i - \hat{p}_i \right|^2 + \delta \sum_{i=1}^{3} \left| \Delta p_i - \Delta \hat{p}_i \right|^2 = d + \delta \sum_{i=1}^{3} \left| \Delta p_i - \Delta \hat{p}_i \right|^2 \qquad \text{(III.10)}$$

Où δ est un paramètre qui contrôle le degré de la contribution de la distorsion de la dynamique (variation) du pitch dans la quantification. Dans notre cas nous l'avons pris égal à 1 (contribution maximale).

B. Quantification des coefficients LP

Vues leur importance et leur influence directe sur la qualité du signal synthétique, nous devons donner beaucoup d'importance à la quantification de ces coefficients. C'est pour cette raison que nous avons alloué 43 sur 81 bits à la quantification des coefficients LP, soit plus de 50%. Nous avons exploité la redondance qui caractérise le spectre du conduit vocal dans les zones voisées pour faire une interpolation. Cette approche nous a permis de gagner plus de 32 bits par super-trame en comparaison avec le FS MELP à 2.4 kbps qui utilise 75 bits pour la quantification des coefficients LP.

En pratique, notre approche a consisté d'abord à convertir les coefficients LP en coefficients LSF avant leur quantification. Puis, nous avons choisi une stratégie de quantification qui tient compte du mode de voisement de chaque super-trame dont nous distinguons les cas suivants:

- *Une super-trame avec au plus une trame voisée :*

  Nous quantifions alors les trames séparément, chacune avec une VQ simple et en utilisant un dictionnaire de 9 bits lorsque la trame est non-voisée, sinon nous utilisons une MSVQ avec un dictionnaire 4 étages. 25 bits sont alors distribués sur les 4 étages selon l'ordre suivant : 7, 6, 6 et 6 bits respectivement (il s'agit du même dictionnaire que celui dictionnaire utilisé dans le standard FS1017).

- *Une supe- trame avec plus d'une trame voisée* :

  Dans ce cas nous quantifions seulement la troisième trame. Les deux autres trames seront déduites par interpolation entre la dernière trame de la super trame précédente et celle de la super trame courante. Notons que pour la trame à quantifier, nous utilisons une MSVQ de 25 bits lorsque la trame est voisée, sinon, on utilise une VQ à 9 bits comme décrit dans le cas précédent.

Pour le calcul des coefficients d'interpolation, nous avons utilisé la procédure suivante [5]:

Soient l_1, l_2, et l_3 les vecteurs des coefficients LSF non quantifiés représentant les trois trames et $\hat{l}_1$, $\hat{l}_2$ et $\hat{l}_3$ leurs valeurs correspondantes quantifiées. Les coefficients d'interpolation pour les deux premières trames seront comme suit :

$$\tilde{l}_1(j) = a_1(j)\hat{l}_p(j) + [1 - a_1(j)]\hat{l}_3(j) \quad \text{(III.11)}$$

$$\tilde{l}_2(j) = a_2(j)\hat{l}_p(j) + [1 - a_2(j)]\hat{l}_3(j) \quad \text{(III.12)}$$

Pour $j = 1,...,10$

Où $\hat{l}_p$ est le vecteur des coefficients quantifiés de la super-trame précédente. $a_1(j)$ et $a_2(j)$ sont les coefficients d'interpolation. Ces coefficients sont stockés dans un dictionnaire de dimension 20 et de taille 16 (4 bits). La sélection de la bonne combinaison se fait en minimisant le critère suivant:

$$E = \sum_{j=1}^{10} w_1(j)\left|l_1(j) - \tilde{l}_1(j)\right|^2 + \sum_{j=1}^{10} w_2(j)\left|l_2(j) - \tilde{l}_2(j)\right|^2 \quad \text{(III.13)}$$

$w_i(j)$ sont les coefficients de pondération obtenus avec la même procédure que celle utilisée dans le FS-MELP [34]. Par la suite, on calcule les résiduels (erreurs) d'interpolation :

$$r_1(j) = l_1(j) - \tilde{l}_1(j) \quad \text{(III.14)}$$

$$r_2(j) = l_2(j) - \tilde{l}_2(j) \quad \text{(III.15)}$$

Pour $j = 1,...,10$

Ces résiduels doivent être aussi quantifiés et transmis pour la correction des résultats. Pour cela, on met les deux résiduels dans un seul vecteur de 20 composantes que l'on quantifie par une MSVQ de quatre étages lorsque la trame est non-voisée et de deux étages lorsque celle-ci est voisée.

Finalement le tableau III.2 donne l'allocation des bits retenus pour la quantification des LSF.

Tableau III.2. Allocation des bits pour la quantification des LSF

Mode de voisement	LSF l_1	LSF l_2	LSF l_3	Coefficients d'interpolation	Résiduels	Totale
UUU	9	9	9	0	0	27
VUU	7.6.6.6	9	9	0	0	43
UVU	9	7.6.6.6	9	0	0	43
UUV	9	9	7.6.6.6	0	0	43
UVV	0	0	7.6.6.6	4	8.6	43
VUV						
VVV						
VVU	0	0	9	4	8.6.6.6	39

C. Quantification du gain

Le gain est calculé deux fois par trame ce qui nous donne six gains pour chaque super-trame. Pour les quantifier, nous avons utilisé une VQ simple avec un dictionnaire de dimension 6 et de taille 1024 (10 bits).

D. Quantification des intensités de voisement

La bande la plus basse représente la décision de voisement de la trame. Elle est déjà quantifiée avec le pitch. La décision de voisement des quatre bandes restantes, est transmise uniquement pour les trames voisées. Nous avons quantifié les décisions de voisement de chaque trame séparément en utilisant un dictionnaire de dimension 4 et aussi de taille 4 (2 bits). Ce dictionnaire ne contient que les combinaisons les plus probables des intensités de voisement.

E. Quantification des amplitudes de Fourier

Les amplitudes de la dernière trame voisée dans la super-trame courante sont quantifiées de la même façon que dans le FS-MELP (VQ avec un dictionnaire de 8 bits) [39].

Les amplitudes de Fourier, des autres trames voisées, seront reconstituées par interpolation à l'aide du vecteur quantifié de la super-trame courante et celui de la super-trame précédente.

F. Quantification de l'indice d'apériodicité

L'indice d'apériodicité n'est obtenu que pour les trames voisées. Nous avons utilisé un bit (taille 2) pour le quantifier.

G. Création des dictionnaires

Pour la création des dictionnaires, nous avons utilisé l'algorithme LBG et une base de données arabe phonétiquement équilibrée [48]. Sous MATLAB et à partir de la version 7.5, il existe un « toolbox » dédié à la quantification vectorielle appelé **vqdtool** (ou **v**ector **q**uantization **d**esing **t**ool). Celui-ci nous a permis la création ainsi que l'évaluation des dictionnaires [49]. Avant la génération des dictionnaires, un travail préparatoire est effectué pour la création des bases d'apprentissage des différents paramètres en utilisant la base de données arabe et le sous-module d'analyse décrit précédemment.

Le dictionnaire des coefficients d'interpolation est généré en utilisant l'algorithme de Lloyd généralisé. Dans l'algorithme d'apprentissage nous avons utilisé deux procédures alternatives : la première consiste à quantifier les coefficients LSF à l'aide d'une base d'apprentissage avec la mesure de la distorsion E (formule III.13) précédente et la deuxième optimise le dictionnaire d'une manière à minimiser la somme des distorsions sur toutes les super-trames de la base d'apprentissage. Cette optimisation nous donne :

$$a_1(j) = \frac{\sum_n w_{1,n}(j)[\hat{l}_{3,n}(j) - l_{1,n}(j)].[\hat{l}_{3,n}(j) - \hat{l}_{p,n}(j)]}{\sum_n w_{1,n}(j)[\hat{l}_{3,n}(j) - \hat{l}_{p,n}(j)]^2} \qquad \text{(III.16)}$$

$$a_2(j) = \frac{\sum_n w_{2,n}(j)[\hat{l}_{3,n}(j) - l_{2,n}(j)].[\hat{l}_{3,n}(j) - \hat{l}_{p,n}(j)]}{\sum_n w_{2,n}(j)[\hat{l}_{3,n}(j) - \hat{l}_{p,n}(j)]^2} \qquad \text{(III.17)}$$

pour $j = 1,...,10.$

Le tableau III.3 montre la répartition des 81 bits, utilisés dans notre implémentation, sur les différents paramètres que nous avons choisis pour représenter une super-trame du signal de parole.

Tableau III.3. Allocation des bits pour la quantification des paramètres

Paramètres	Mode de voisement				
	VVV	UVV VUV	VVU	UUV UVU VUU	UUU
Pitch et décisions globales de voisement	12	12	12	12	12
LSFs	43	43	39	43	27
gains	10	10	10	10	10
Voisement des sous-bandes	6	4	4	2	0
Amplitudes de Fourier	8	8	8	8	0
Indice d'apériodicité	1	1	1	1	0
Synchronisation	1	1	1	1	1
Protection contre les erreurs	0	2	6	4	31
Total	**81**	**81**	**81**	**81**	**81**

III.4. module de décodage mis en place

Le rôle de ce module est d'utiliser les paramètres reçus de l'encodeur pour reconstruire un signal de parole intelligible. Il contient un sous-module de dé-quantification (décompression) des paramètres quantifiés et un sous-module pour la synthèse de la parole.

Le schéma donné à la figure III.6 illustre le diagramme du sous-module de synthèse que nous avons implémenté. L'opération de synthèse est synchronisée avec le pitch. Le signal d'excitation et les trois autres paramètres (coefficients LPC, gains et amplitudes de Fourier) sont mis à jour à chaque période de pitch en utilisant une interpolation linéaire des paramètres : LPC, pitch, amplitudes de Fourier et gains.

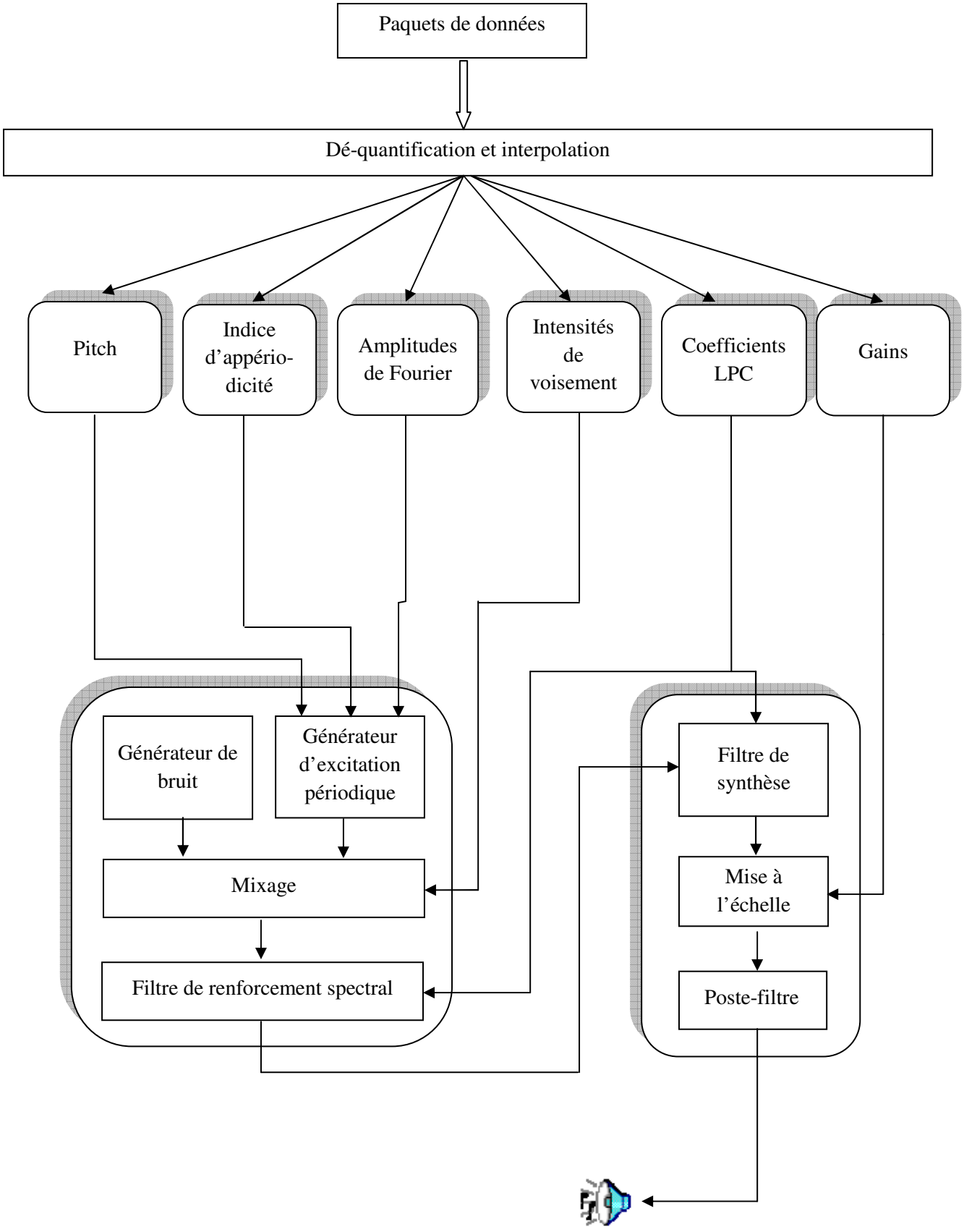


Fig III.6 : bloc de la synthèse MELP

III.4.1. Mise au point du générateur d'excitation

Pour les trames voisées, nous avons calculé l'excitation périodique en appliquant la transformée de Fourier inverse du vecteur des amplitudes de Fourier interpolés. Cela se fait sur une période de pitch. Nous avons procédé de la même manière pour les trames apériodiques, mais avec un pitch « perturbé » (présentant des « jitters ») comme suit [1]:

$$\hat{P} = P(1 + 0.25x) \tag{III.18}$$

Pour $x \in [-1, 1]$

Pour avoir une excitation mixte à bande de fréquences complète, nous avons mixé deux signaux : l'excitation périodique et la composante bruit (générée à partir d'un bruit blanc). Ce mixage a été contrôlé par les filtres de mise en forme.

Cette excitation mixte doit passer par un filtre de renforcement spectral avant d'arriver au filtre de synthèse. Il s'agit du post-filtre utilisé dans le standard FS1016 mais avec des paramètres : α et β adaptatifs [1]. Pour cette adaptation nous avons utilisé deux paramètres : le premier est donnée par :

$$\rho = \frac{g_{\text{int}} - g_n - 12}{18} \tag{III.19}$$

Où g_{int} est le gain interpolé de la période de pitch courante et g_n est une estimation du gain du bruit de fond.

Le deuxième paramètre représente le premier coefficient de réflexion k_1 calculé à partir des coefficients LPC dé-quantifiés est interpolés.

La fonction de transfert du filtre de renforcement spectral est alors:

$$H(z) = \frac{A(\alpha z^{-1})}{A(\beta z^{-1})}(1 + \mu z^{-1}) \tag{III.20}$$

avec

$$\alpha = 0.5\rho \qquad \beta = 0.8\rho \qquad \mu = \max(0.5k_1, 0)$$

III.4.2. Filtre de synthèse appliqué

Nous avons utilisé les coefficients LSF interpolés et convertis en coefficients LP pour le filtrage direct du signal d'excitation. Afin de donner à la trame de parole synthétique le même niveau d'énergie que la trame originale, nous avons multiplié la sortie du filtre de synthèse par un gain (ou facteur de mise à l'échelle) donné par [1]:

$$S_{gain} = \frac{10^{\frac{g\,\mathrm{int}}{20}}}{\sqrt{\frac{1}{T}\sum_{n=1}^{T}\hat{s}_n^2}} \qquad \text{(III.21)}$$

Le numérateur de ce facteur est le gain interpolé en échelle linéaire et le dénominateur est la norme L_2 d'une période de pitch T du signal synthétique utilisé pour la normalisation.

Afin d'améliorer la qualité du signal synthétique dans les zones de fréquence non-formantiques, nous avons utilisé un filtre à réponse impulsionnelle finie (FIR) d'ordre 6.

III.5. Conclusion

Dans ce chapitre, nous avons vu la mise au point d'un codeur à très bas débit basé sur le MELP et fonctionnant à 1.2 kbps. Nous avons utilisé une approche multi-trames pour mettre au point notre codeur. Cette approche a nécessité l'utilisation de plusieurs quantifications scalaires et vectorielles. Ainsi, pour quantifier les coefficients LSF, les amplitudes de Fourier et les gains, nous avons utilisé seulement une quantification vectorielle conjointe. Une technique d'interpolation a été appliquée aux coefficients LSF. En ce qui concerne le pitch, deux quantifications conjointes ont été employées, l'une scalaire et l'autre vectorielle. Les résultats obtenus seront présentés et interprétés dans le chapitre suivant.

CHAPITRE

4

RESULTATS ET EVALUATIONS

IV.1. Introduction

Une considération importante dans tout codage de parole est la qualité perceptuelle du signal reconstruit qui peut être évaluée par des critères de mesure subjective [1, 10] ou objective [1, 26]. L'évaluation de la qualité perceptuelle d'un codeur de parole à très bas débit est une tâche très délicate et les seuls critères capables d'évaluer globalement cette qualité sont les critères subjectifs où la parole synthétisée est écoutée et évaluée par des auditeurs [1]. Cependant une telle évaluation est souvent difficile à mettre au point. Pour cette raison, des nouvelles techniques d'évaluation objective, connue sous le nom mesure perceptuelle objective, basées sur la modélisation de l'audition et qui tiennent compte de certaines caractéristiques de la perception humaine ont été mises en place [5].

Dans le présent chapitre nous présentons et nous interprétons les résultats obtenus par les différents blocs de notre codeur et de notre décodeur dont nous rappelons les schémas aux figures IV.1 et IV.2. Nous allons aussi évaluer la qualité d'implémentation de chaque bloc pour arriver à la fin à l'évaluation de la qualité perceptuelle du signal décodé. Mais avant de chiffrer nos résultats nous allons d'abord définir les outils linguistiques et mathématiques que nous avons utilisés lors de nos tests.

IV.2. Méthodes d'évaluation utilisées

Sur les figures IV.1 et IV.2 nous avons fait apparaitre les blocs essentiels de notre codeur ayant fait l'objet d'évaluation.

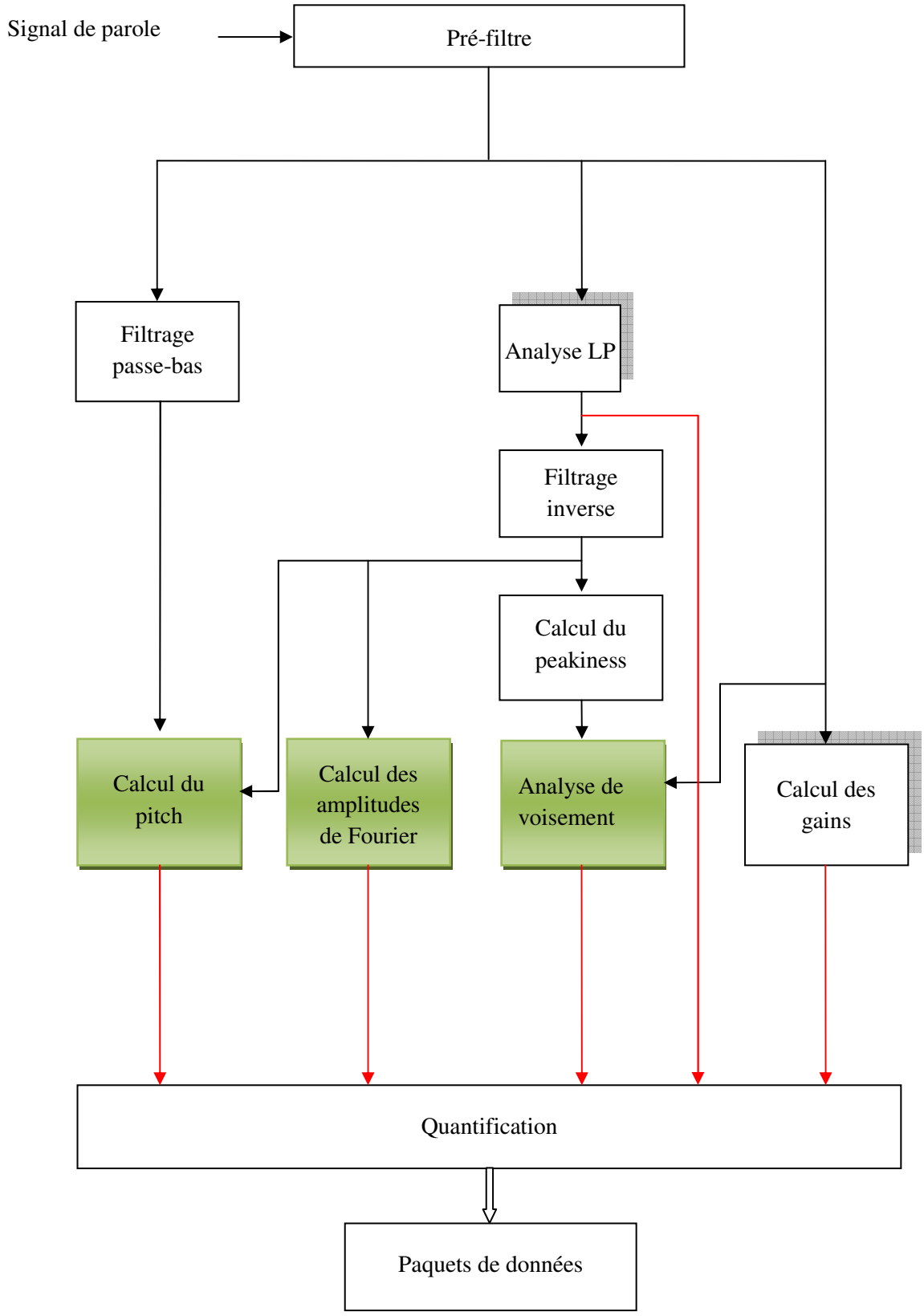


Fig IV.1. Blocs à évaluer dans le codeur MELP mis au point

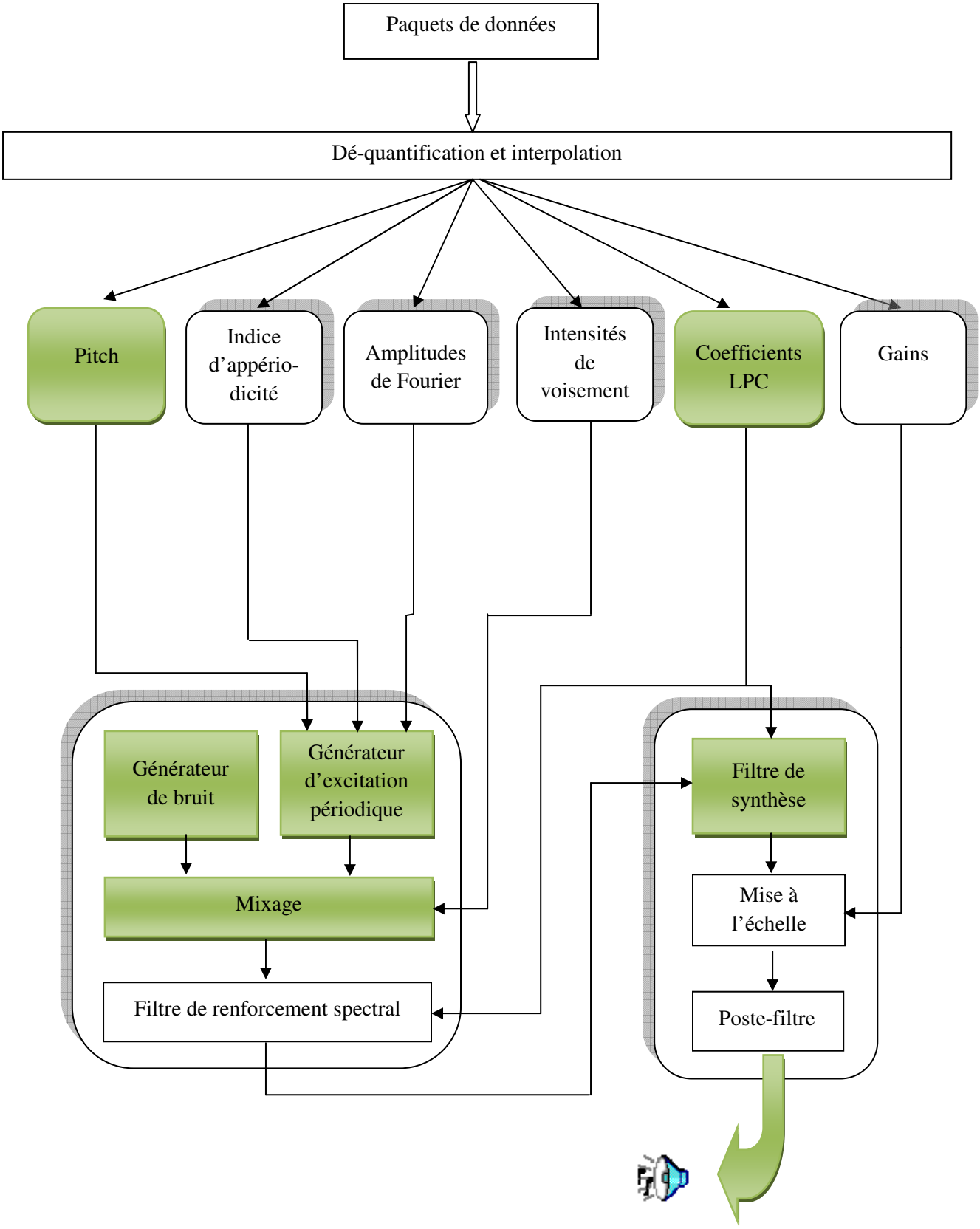


Fig IV.2. Blocs à évaluer dans le décodeur MELP mis au point

Comme outil linguistique employé dans nos tests, nous avons utilisé principalement un corpus arabe phonétiquement équilibré (PAPE) enregistré dans une chambre sourde et échantillonné à 10 KHz [48]. Nous avons dû opérer un sous-échantillonné à 8Khz pour se mettre dans le cas de la téléphonie. D'autres phrases issues de corpus anglais et français ont été également utilisées.

Pour évaluer la précision de la détection et la quantification du pitch, nous avons utilisé l'erreur absolue.

En ce qui concerne la quantification des coefficients LP nous avons utilisé la distorsion spectrale.

Afin de mesurer la qualité perceptuelle objective nous avons utilisé la méthode PESQ (Perceptual evalution speech quality) décrite par le standard P862 de ITU [29].

De plus nous avons utilisé d'autres mesures de la qualité objective comme les LLR (ou log likelihood ratio), les WSS (ou weighted spectral slope) et le SNR segmental.

IV.2.1. Evaluation par mesure de la distorsion spectrale (SD)

La distorsion spectrale est souvent utilisée comme mesure objective de la performance d'encodage des systèmes fonctionnant à bas débit [50]. Dans nos tests, nous avons utilisé la distorsion moyenne développée et calculée sur une largeur de bande limitée [n_0 n_1] et pour un nombre M des trames. L'expression discrète de la distorsion spectrale pour la i^{eme} trame est donnée par [50] :

$$SD_i = \sqrt{\frac{10}{n_1 - n_0} \sum_{n=n_0}^{n_1} \left[\log_{10} \left(\frac{S(f_n)}{\hat{S}(f_n)} \right) \right]^2} \qquad \text{(IV.1)}$$

$S(f_n)$ et $\hat{S}(f_n)$ sont respectivement le spectre de puissance original et quantifié du filtre de synthèse LP dont une FFT de longueur N=256 points a été utilisé pour les calculer. Dans la littérature, on trouve plusieurs propositions de la largeur de bande [n_0 n_1] à considérer. Dans notre cas, nous avons considéré la bande complète [0, 4000] Hz avec une résolution de 31.25 Hz [51].

La SD moyenne sera donnée par :

$$\overline{SD} = \frac{1}{M}\sum_{i=1}^{M} SD_i \qquad \text{(IV.2)}$$

En général, une $\overline{SD}$ d'environs 1dB indique que la distorsion perçue pendant la quantification est négligeable [50]. Dans le passé, cette valeur a été suggérée comme seuil pour la transparence spectrale de la parole (une qualité de quantification est dite transparente si aucune distorsion audible due à cette quantification n'est introduite dans la parole codée) [50]. Paliwal et Atal ont établi que la SD moyenne n'est pas suffisante pour mesurer la transparence d'un quantificateur. Ils ont introduit la notion des trames spectrales externes ou (outliers frames). Par conséquent, nous pouvons obtenir une qualité de codage transparent si les conditions suivantes sont maintenues [52]:

- La SD moyenne est d'environ 1dB
- Le pourcentage des trames « outliers » ayant une SD entre 2 et 4 dB est moins de 2%
- Aucune trame ne doit avoir une SD qui dépasse 4 dB

IV.2.2. Rapport signal sur bruit segmental (SNR segmental)

Pour le SNR segmental nous avons retenu la formule de Hansen et Pellom donnée par [53]:

$$SNRseg = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\left[\frac{\sum_{n=N_m}^{N_m+N-1} s^2(n)}{\sum_{n=N_m}^{N_m+N-1}(s(n)-\hat{s}(n))^2}\right] \qquad \text{(IV.3)}$$

Où $s(n)$ et $\hat{s}(n)$ sont respectivement le signal original et bruité, N est la langueur du trame et M le nombre de trames.

*IV.2.3. Evaluation par la mesure de la distance LLR (**L**og **L**ikelihood **R**atio)*

Le LLR ou rapport de vraisemblance logarithmique est une mesure objective basée sur les coefficients LP. La formule de la distance LLR est donnée par [54] :

$$d_{LLR}(A,\hat{A}) = \log_{10}\left[\frac{A\,R\,A^T}{\hat{A}\,R\,\hat{A}^T}\right] \quad \text{(IV.4)}$$

Où R est la matrice d'auto-corrélation, A et $\hat{A}$ sont respectivement le vecteur des coefficients LP du signal original et synthétique.

IV.2.4. Evaluation par mesure de la distance WSS (Weighted Spectral Slope)

La pente spectrale peut être utilisée comme mesure objective de la qualité perceptuelle. En fait, on calcule la distance pondérée et moyennée entre les pentes spectrales du signal original et synthétique, pour plusieurs bandes de fréquence et plusieurs segments du signal. Cette mesure est donnée par la formule suivante [55] :

$$d_{WSS} = \frac{1}{M}\sum_{m=0}^{M-1}\frac{\sum_{j=1}^{K} W_{WSS}(j,m)(S(j,m)-\hat{S}(j,m))^2}{\sum_{j=1}^{K} W_{WSS}(j,m)} \quad \text{IV.5}$$

W_{WSS} représentent les pondérations, k est le nombre de bandes de fréquence (dans notre cas, nous avons pris k=25) et M le nombre de segments du signal. $S(j,m)$ et $\hat{S}(j,m)$ sont respectivement les pentes spectrales du signal original et synthétique pour la j^{eme} bande de fréquence.

IV.3. Evaluation de l'estimation du pitch

Donnée à titre d'exemple, la figure IV.3-a illustre le mot arabe « آذاه » prononcé par un locuteur masculin (un bon exemple pour valider l'algorithme de calcul de pitch mis en place car ce mot est composé uniquement de phonèmes voisés). Pour la validation, nous avons d'abord estimé manuellement le pitch (courbe de mélodie) de ce mot. Nous avons ensuite comparé les résultats obtenus avec ceux donnés par notre algorithme. La figure IV.3-b montre les résultats de cette comparaison. Afin d'évaluer la précision de cette estimation nous avons tracé l'erreur absolue (figure IV.3-c).

Nous avons procédé de la même manière pour un locuteur féminin (figures : IV.4-a, IV.4-b et IV.4-c), prononçant la même phrase.

Pour les deux locuteurs, on constate que cette erreur est plus importante dans les zones dynamiques de pitch (zones de transition) où la périodicité est quasi perdue. L'influence de

l'erreur d'estimation de pitch dans ces zones est négligeable. On rappelle que la synthèse MELP utilise un pitch perturbé (jiter) pour les zones de transition.

Les résultats des tests d'écoute que nous avons effectué sur des signaux synthétiques en utilisant des valeurs de pitch estimées manuellement et des valeurs données par notre algorithme d'estimation du pitch, sont satisfaisants et confirment l'efficacité de la méthode d'auto-corrélation que nous avons implémentée.

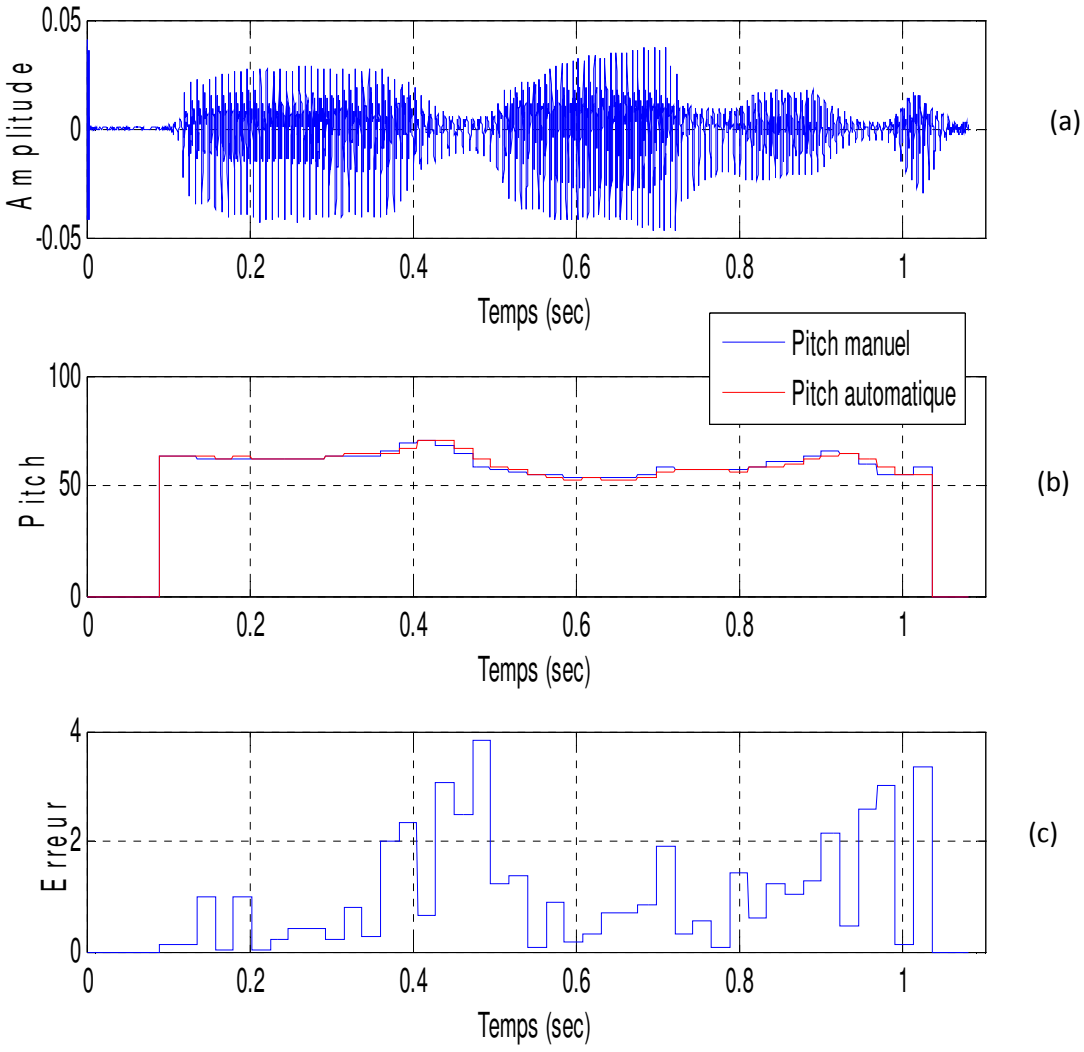


Fig.IV.3. Détection de pitch pour un locuteur masculin a- Signal de parole du mot arabe « آذاه »

b- Courbe de mélodie

c- Erreur d'estimation

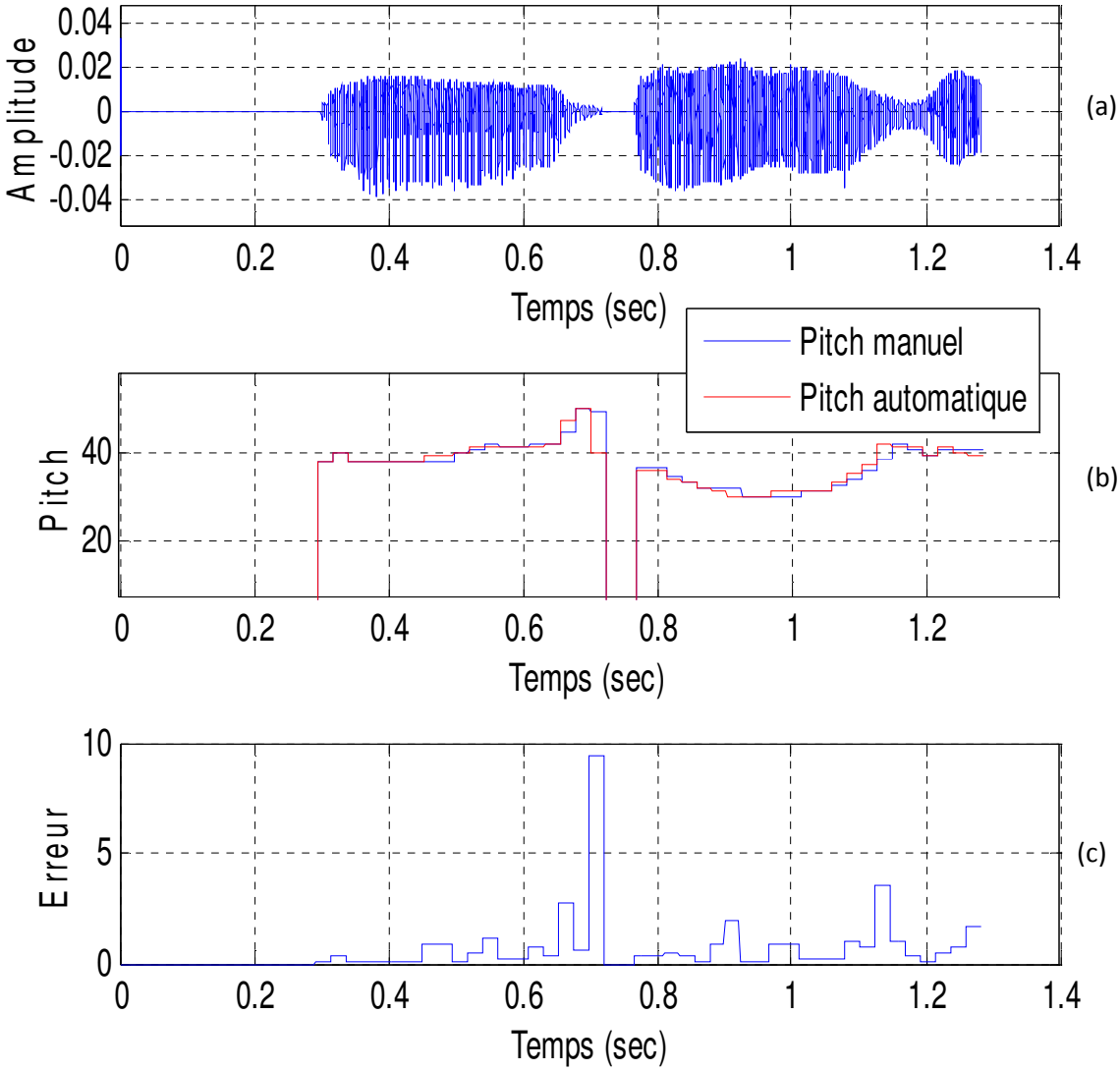


Fig.IV.4. Détection de pitch pour un locuteur féminin a- Signal de parole du mot arabe « آذاه »

b- Courbes de mélodie

c- Erreur d'estimation

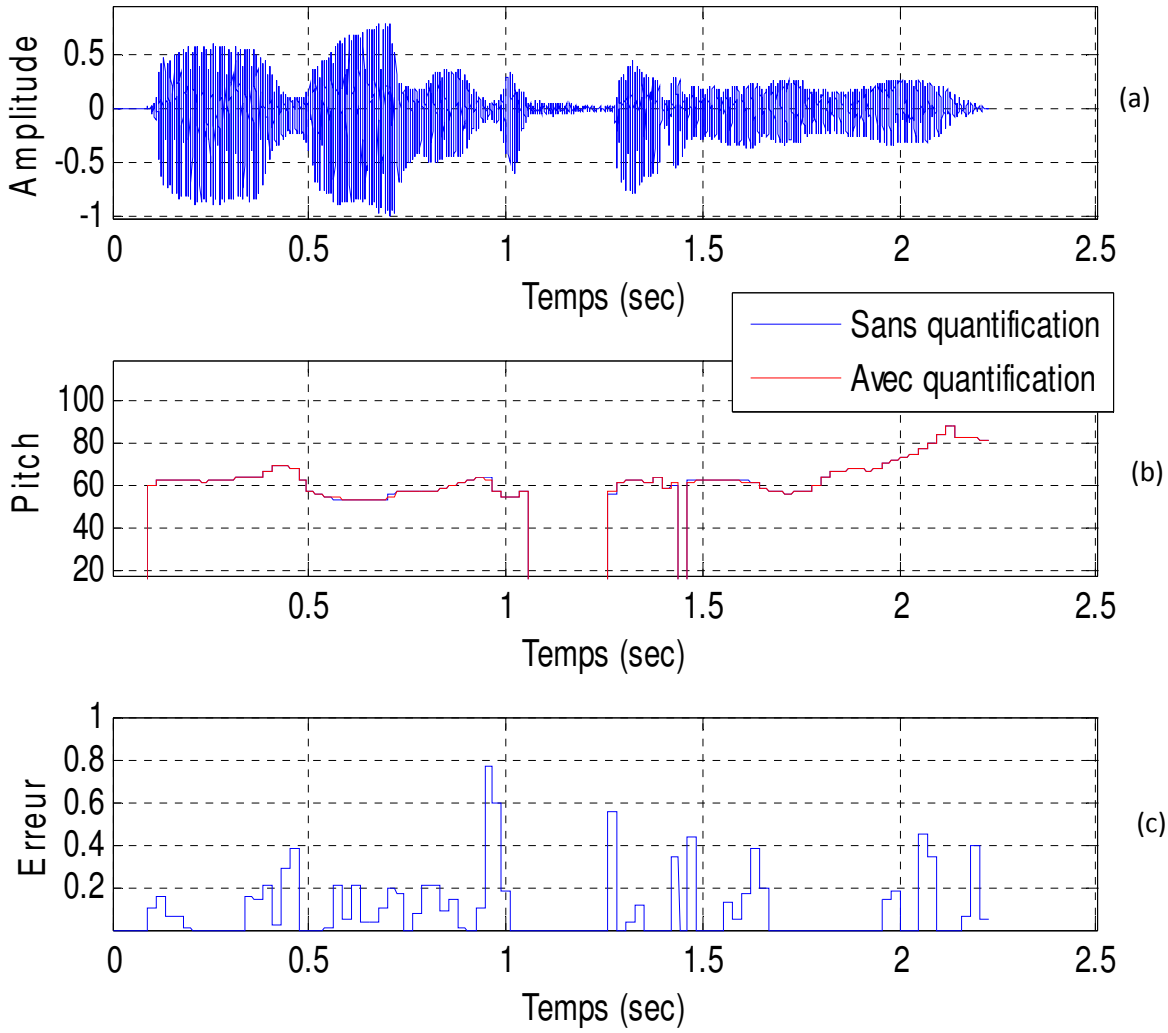


Fig.IV.5. Quantification du pitch pour un locuteur masculin a- Signal de parole du mot arabe «صـعـد الإمام فـوق المنـــبر»

b- Courbes de mélodie

c- Erreur de quantification

En ce qui concerne la quantification du pitch, la figure IV.5 illustre un exemple de quantification du pitch pour la phrase arabe «صـعـد الإمام فـوق لمنـــبر» prononcée par un locuteur masculin (figure IV.5-a). La figure IV.5-b représente les courbes de mélodie avant et après quantification. L'erreur de cette quantification est représentée à la figure IV.5-c.

On constate que l'erreur de quantification est faible (pour le corpus arabe PAPE nous avons trouvé une erreur moyenne de 1.17). Les tests d'écoute effectués ont montré que l'effet de

cette erreur sur la qualité du signal synthétique reste tolérable et en général non perceptible par l'oreille. Ces résultats nous permettent de dire que la quantification vectorielle conjointe multi-dictionnaires que nous avons utilisée dans notre implémentation a donné de bon résultats. Par ailleurs, notons que cette quantification nous a permis d'obtenir un gain en débit considérable (cette technique nous a permis de gagner plus de 133 bits par seconde en comparaison avec le standard MELP 2.4 kbps).

IV.4. Evaluation du signal d'excitation

Nous rappelons que pour générer un signal d'excitation périodique, on utilise les amplitudes de Fourier correspondant aux dix premiers harmoniques du pitch dans le spectre du résiduel de prédiction. Dans le chapitre précédent nous avons décrit la procédure utilisée pour la détection de ces amplitudes. La précision de cette détection influe notablement sur la qualité du signal synthétique.

La figure IV.6 montre un exemple de spectre du résiduel de prédiction correspondant à une trame voisée. Les dix amplitudes de Fourier détectées par notre algorithme sont clairement indiquées. On constate que ces amplitudes correspondent bien aux dix premiers pics du spectre du résiduel.

L'ensemble des figures : IV.7, IV.8 et IV.9 illustrent le signal d'excitation, délivré par le générateur d'excitation mis au point, avec ses deux composantes, périodique et bruit avant et après mise en forme et pour différents cas de voisement.

La figure IV.7 représente le cas d'une trame non voisée. On constate l'absence de la composante périodique, bloquée par le filtre de mise en forme. Le signal d'excitation est formé uniquement par un bruit dont la mise en forme est effectuée par un filtre passe-tout.

Il s'agit de modéliser l'écartement total des cordes vocales pour produire les sons non-voisés.

La figure IV.8 représente le cas d'un signal voisé. L'excitation est alors complètement périodique. Ce cas correspond à une vibration survenant à la fermeture complète des cordes vocales pour produire des sons voisées non fricatifs.

La figure IV.9 illustre le cas d'une trame voisée avec friction (des sous-bandes non voisées). Les filtres de mise en forme ont fait passer les deux composantes : périodique et bruit, avec un effet de filtrage plus visible sur le bruit pour donner un bruit de haute

fréquence, du fait que la trame est voisée. Ce cas correspond à un rétrécissement du chenal expiratoire pour produire des sons voisés mais fricatifs.

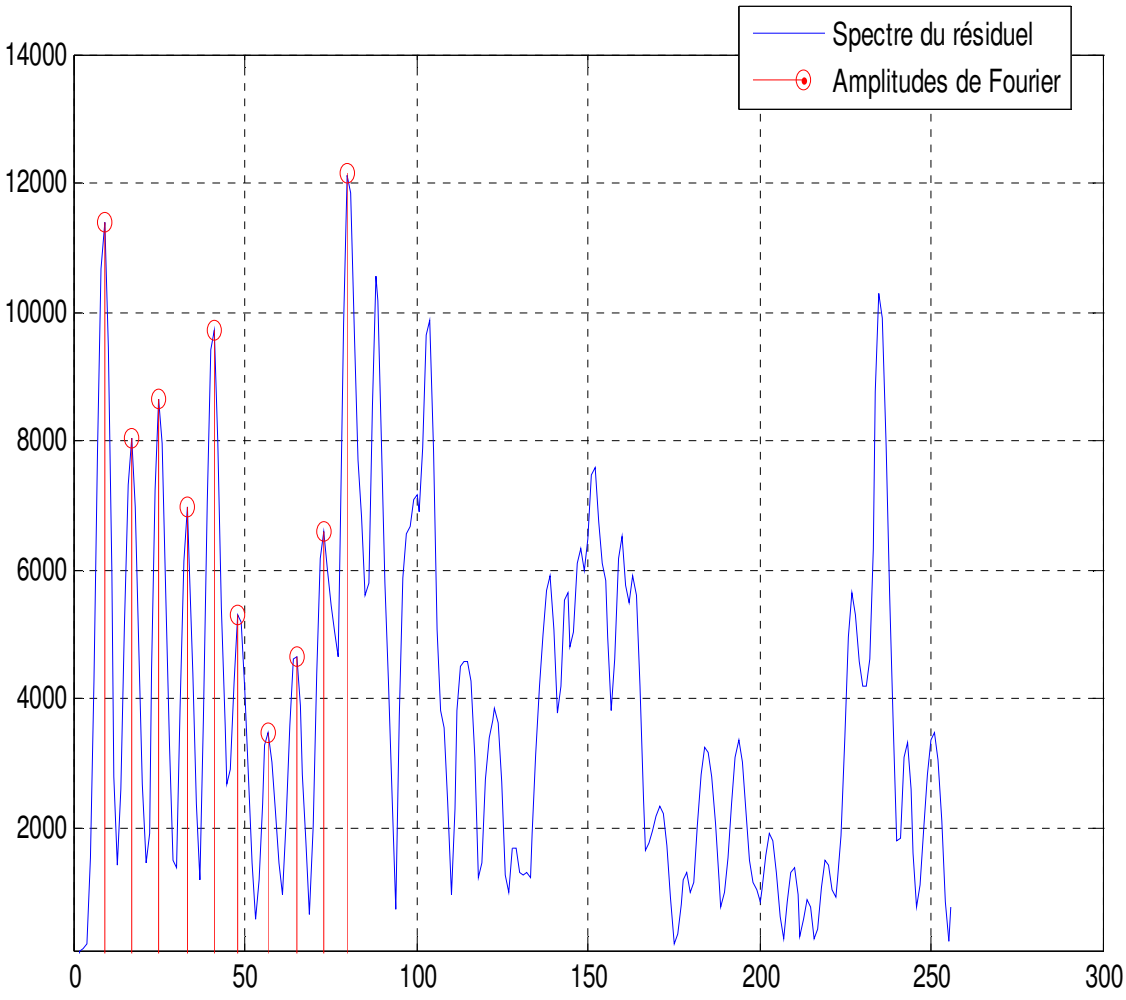


Fig IV.6. Spectre du signal résiduel de prédiction avec les dix amplitudes de Fourier

mise en forme

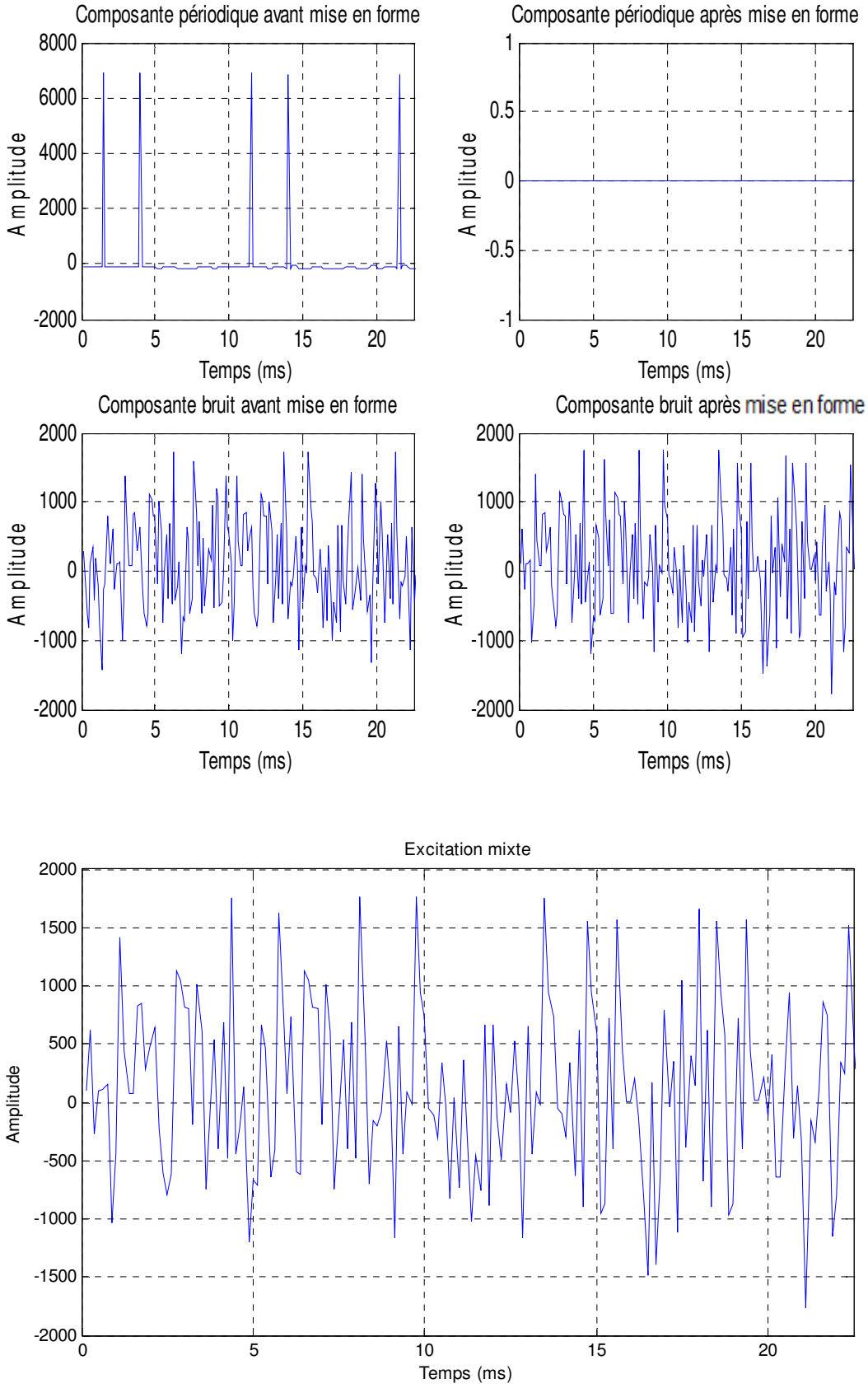
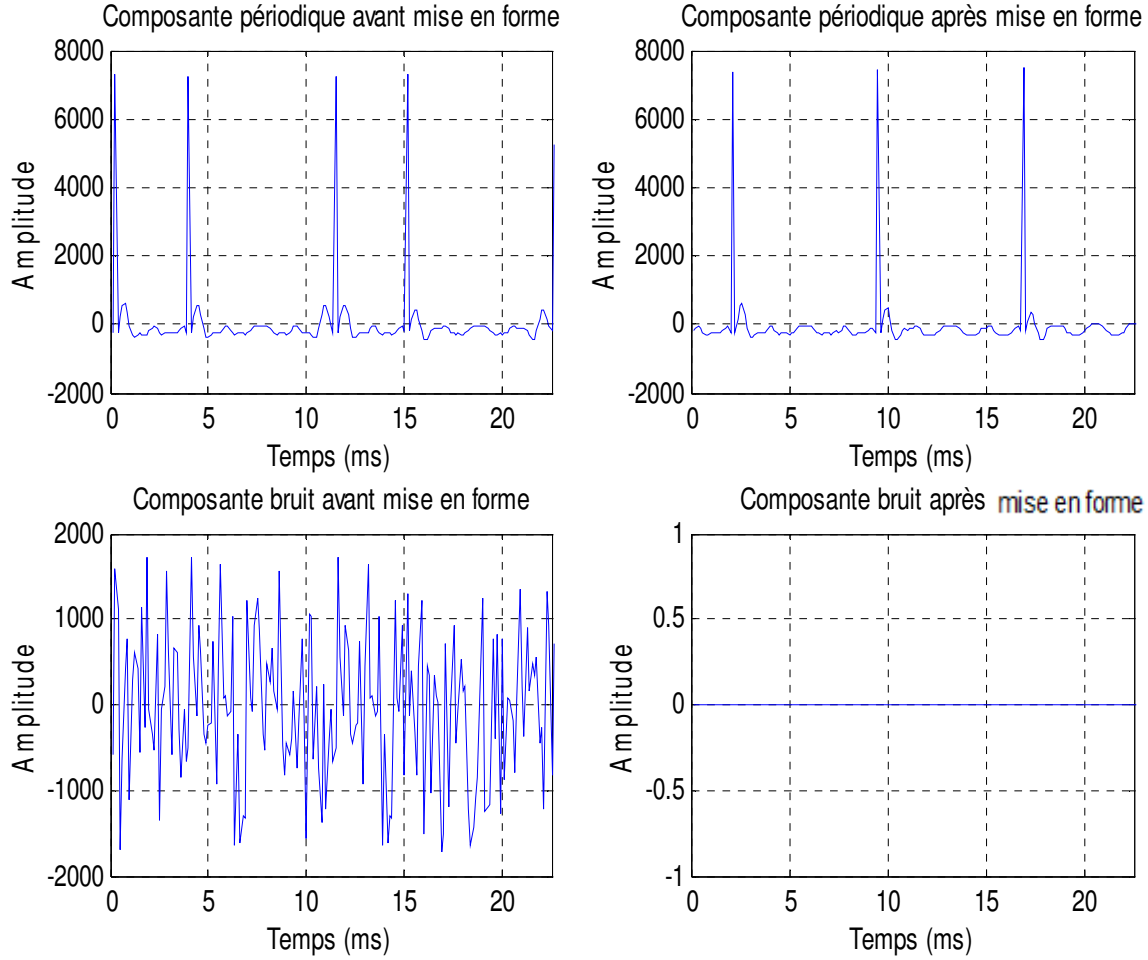


Fig IV.7. Signaux d'excitation mixte pour une trame non-voisée



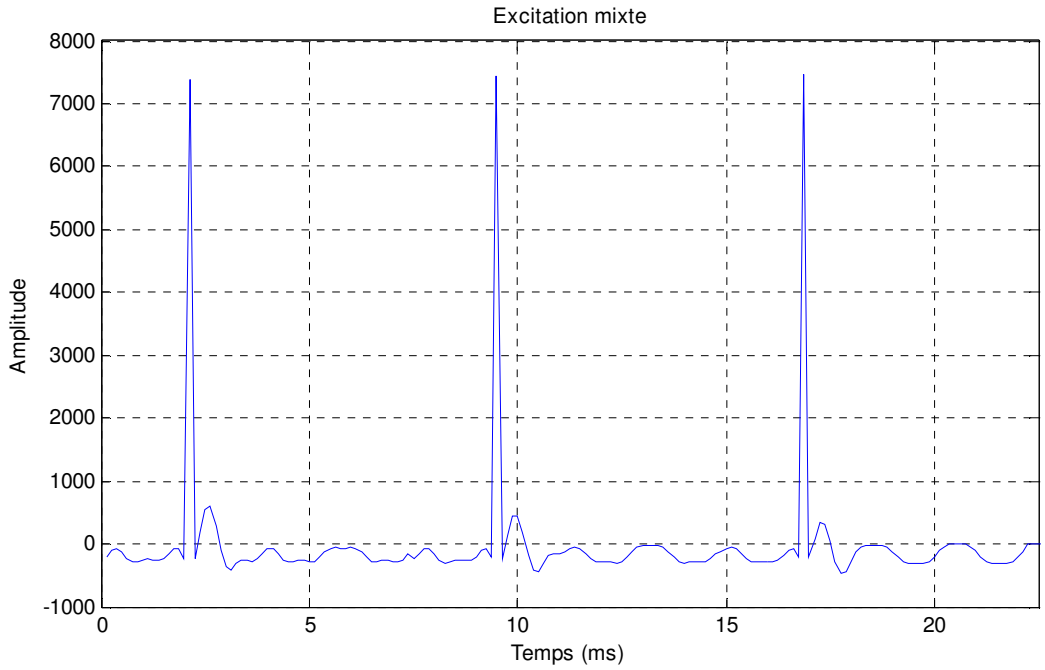


Fig IV.8. Signaux d'excitation mixte pour une trame voisée sans friction

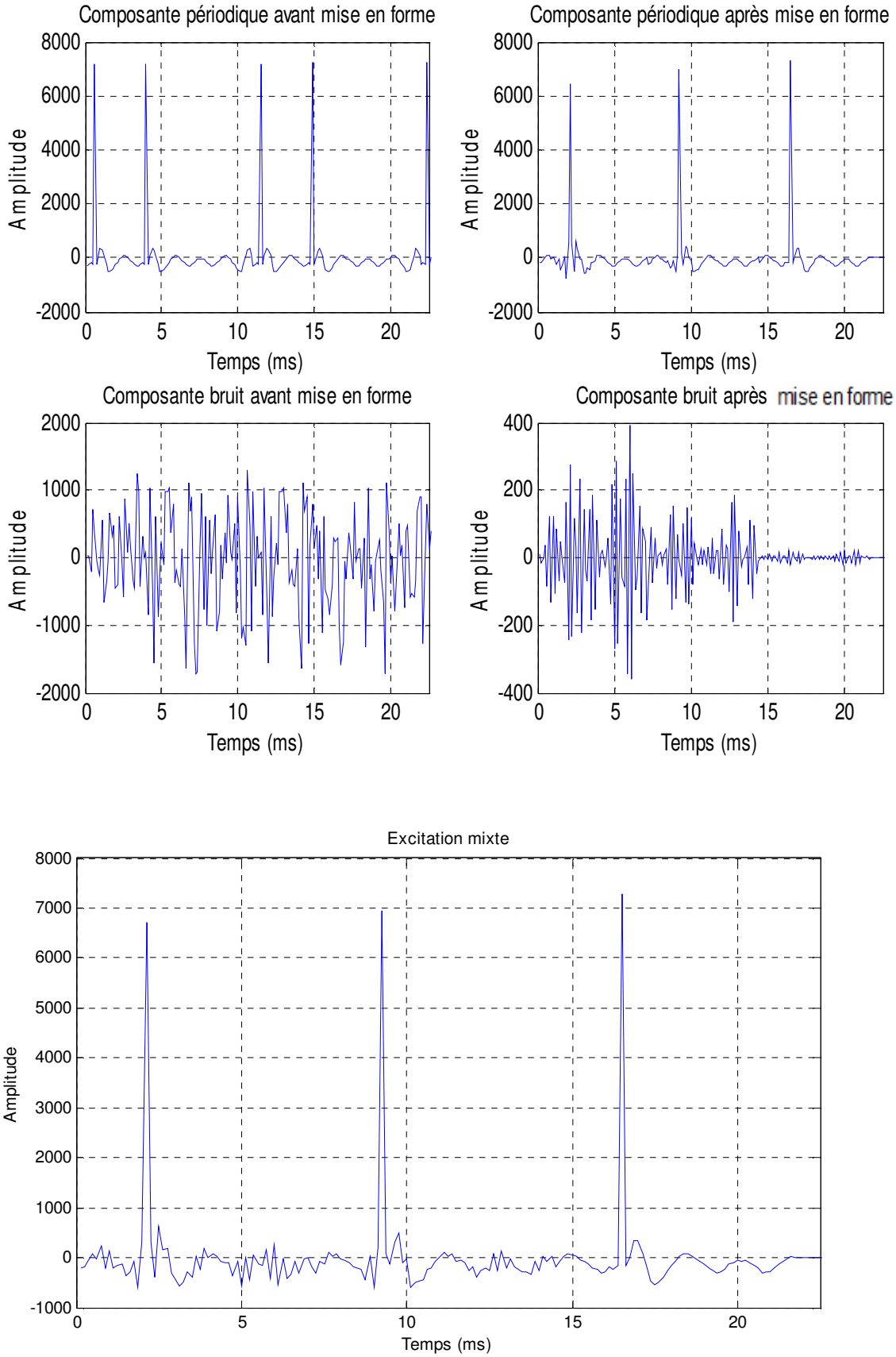


Fig IV.9. Signaux d'excitation mixte pour une trame voisée avec friction

IV.5. Evaluation du filtre de synthèse mis au point

La figure IV.10-a représente la réponse fréquentielle d'un filtre de synthèse pour une trame voisée. Le signal filtré est donné à la figure IV.10-b. La réponse fréquentielle montre une structure formantique très remarquable.

La figure IV.11-a illustre la réponse fréquentielle d'un filtre de synthèse correspondant à une trame non voisée. Le signal d'excitation correspondant et le signal synthétique sont donnés par les figures IV.11-b et IV.11-c. On remarque bien l'absence de la structure formantique dans la réponse fréquentielle du filtre de synthèse.

Afin d'évaluer la technique quantification vectorielle conjointe (JVQ ou *Joint Vector Quantization*) avec interpolation proposée pour quantifier les coefficients LP, nous avons utilisé la mesure de la distorsion spectrale décrite précédemment. Pour 2031 trames analysées (M=2031) nous avons obtenu les résultats suivants :

Tableau IV.1. Résultats du test de transparence

Distorsion moyenne (dB)	**Outlier > 2dB (%)**	**Outlier > 4 dB (nombre de trames)**
1.16	8	3

Nous remarquons que la condition de transparence n'est pas vérifiée. Distorsion spectrale moyenne est supérieure à 1dB, plus de 2% des trames ont une distorsion spectrale supérieure à 2dB. De plus certaines trames ont une distorsion spectrale supérieure à 4dB.

Ces résultats peuvent être justifiés par le cumul des erreurs de quantification vectorielle et des erreurs de l'interpolation car nous avons utilisé seulement les troisièmes trames des deux super-trames consécutives.

Certes, cette erreur de quantification des coefficients LP influencera même la qualité perceptuelle en comparaison avec le FS-MELP. Malheureusement, cette perte est nécessaire pour réduire le débit à 1.2 kbps. Cela nous a permis en fait de gagner 474 bits par seconde par rapport au standard FS1017.Notons cependant que l'absence de cette transparence n'influe pas l'intelligibilité de la parole décodée. Nos résultats restent conformes à ceux donnés dans la littérature dés lors qu'on réalise des codeurs VLBR.

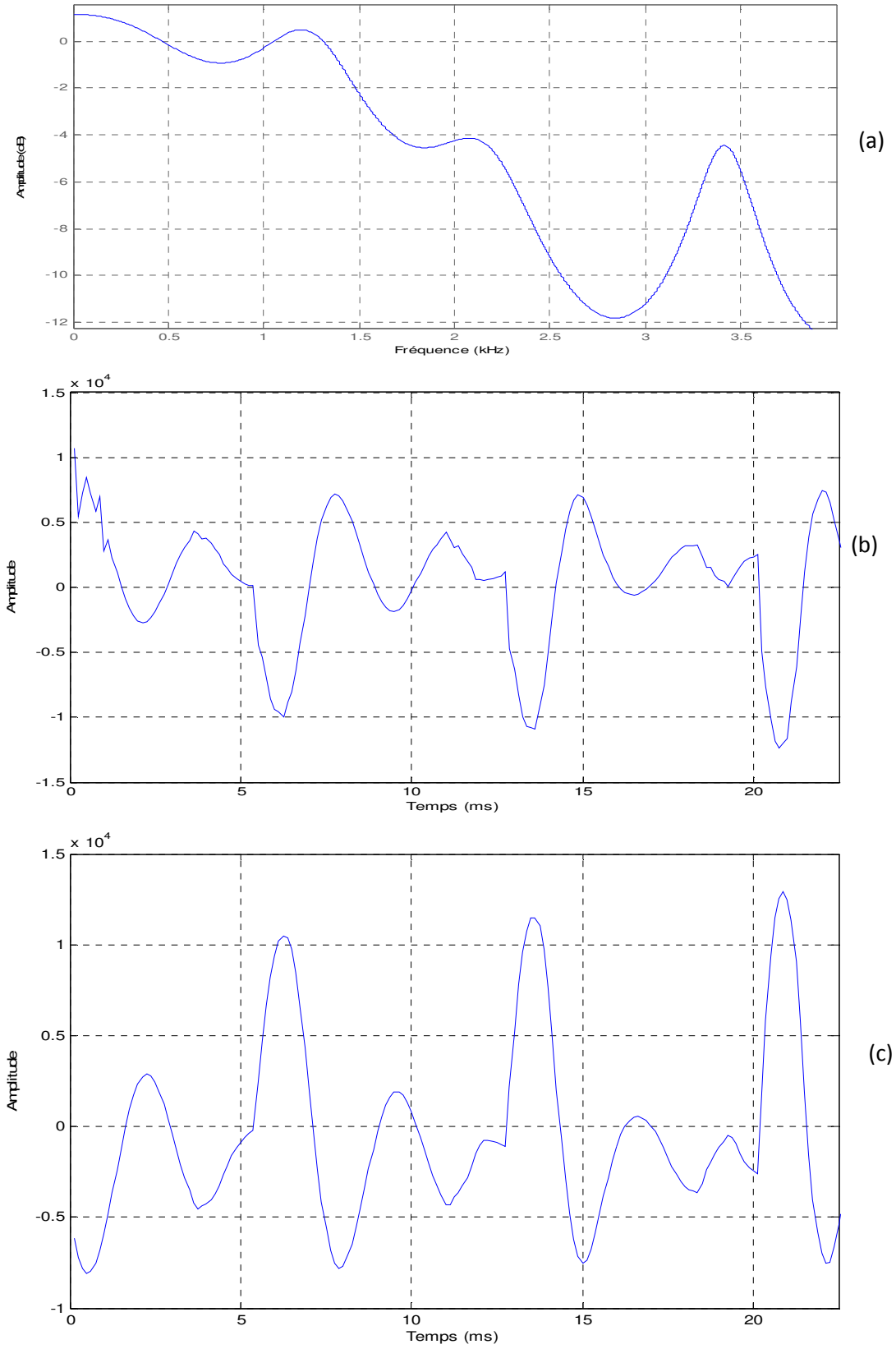


Fig IV.10. Filtre de synthèse pour une trame voisée : a-Réponse fréquentiel
b- signal d'excitation
c- Signal filtré

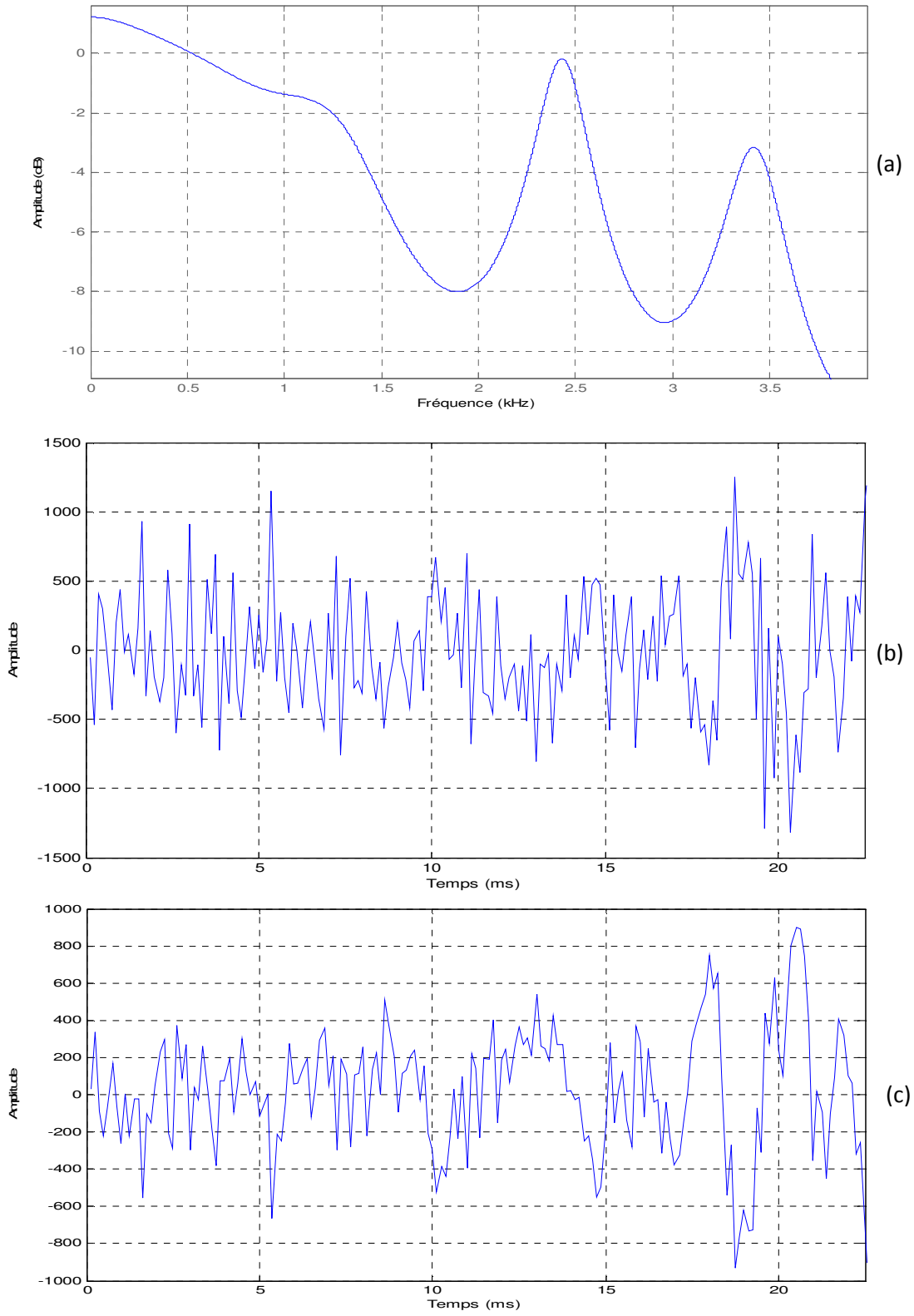


Fig IV.11. Filtre de synthèse pour une trame non-voisée : a-Réponse fréquentiel
b- signal d'excitation
c- Signal filtré

IV.6. Evaluation du signal décodé

La figure IV.12 montre un exemple de résultats où on représente un signal original avec le signal synthétique correspondant. Il s'agit des signaux de la phrase arabe « زحف رمله آذاه » prononcée par un locuteur masculin. La différence dans les allures de ces signaux est prévisible dans le cas du codage paramétrique où on s'intéresse plus à la qualité perceptuelle de la parole qu'à l'allure temporelle des signaux. Ceci justifie le SNR segmental très faible obtenu pour ce signal (-2.75 dB).

La figure IV.13 illustre une zone voisée extraite du signal de la figure précédente (zoom). On remarque des différences dans les allures, mais ces signaux présentent la même périodicité. Ceci confirme la précision du calcul de pitch pour ce signal.

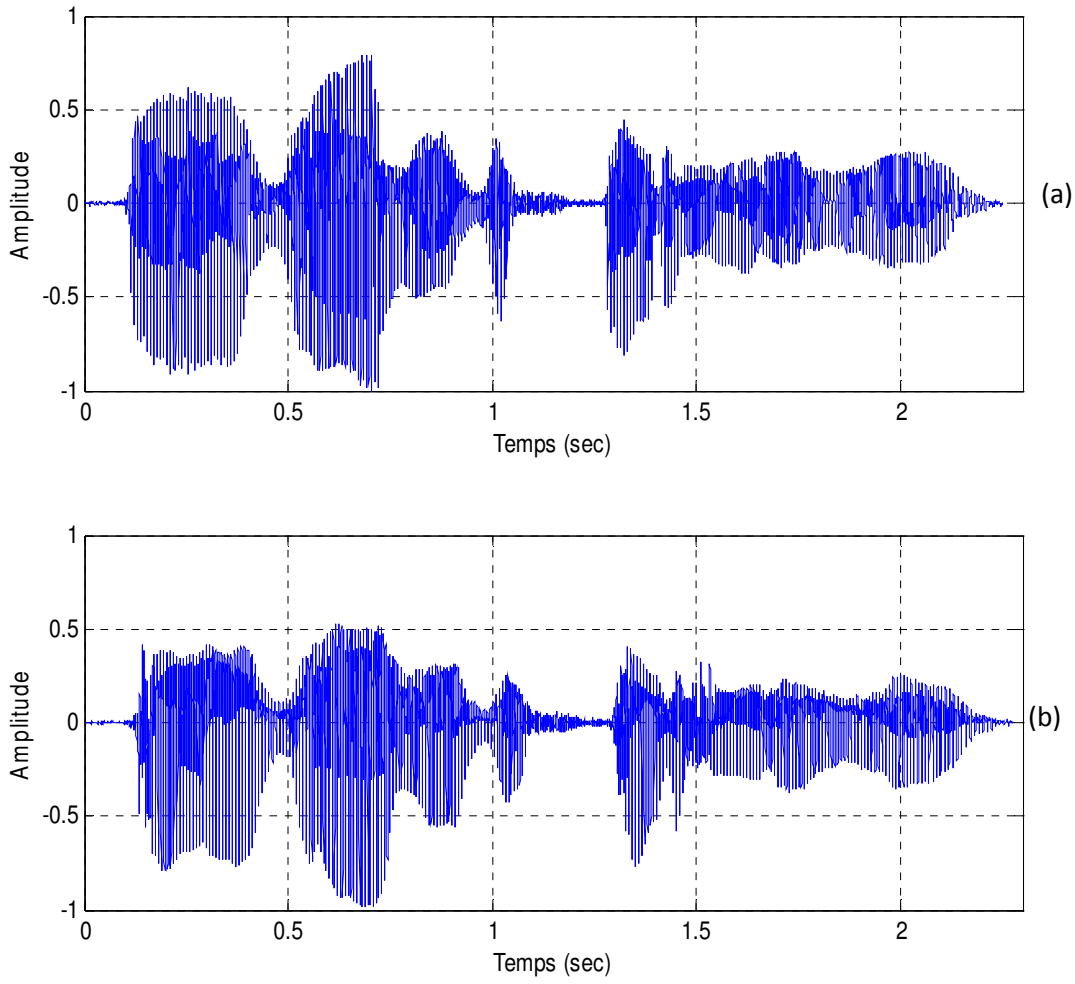


Fig.IV.12. Signaux de la phrase arabe « زحف رمله آذاه » pour un locuteur masculin :
a- Original
b- Synthétique

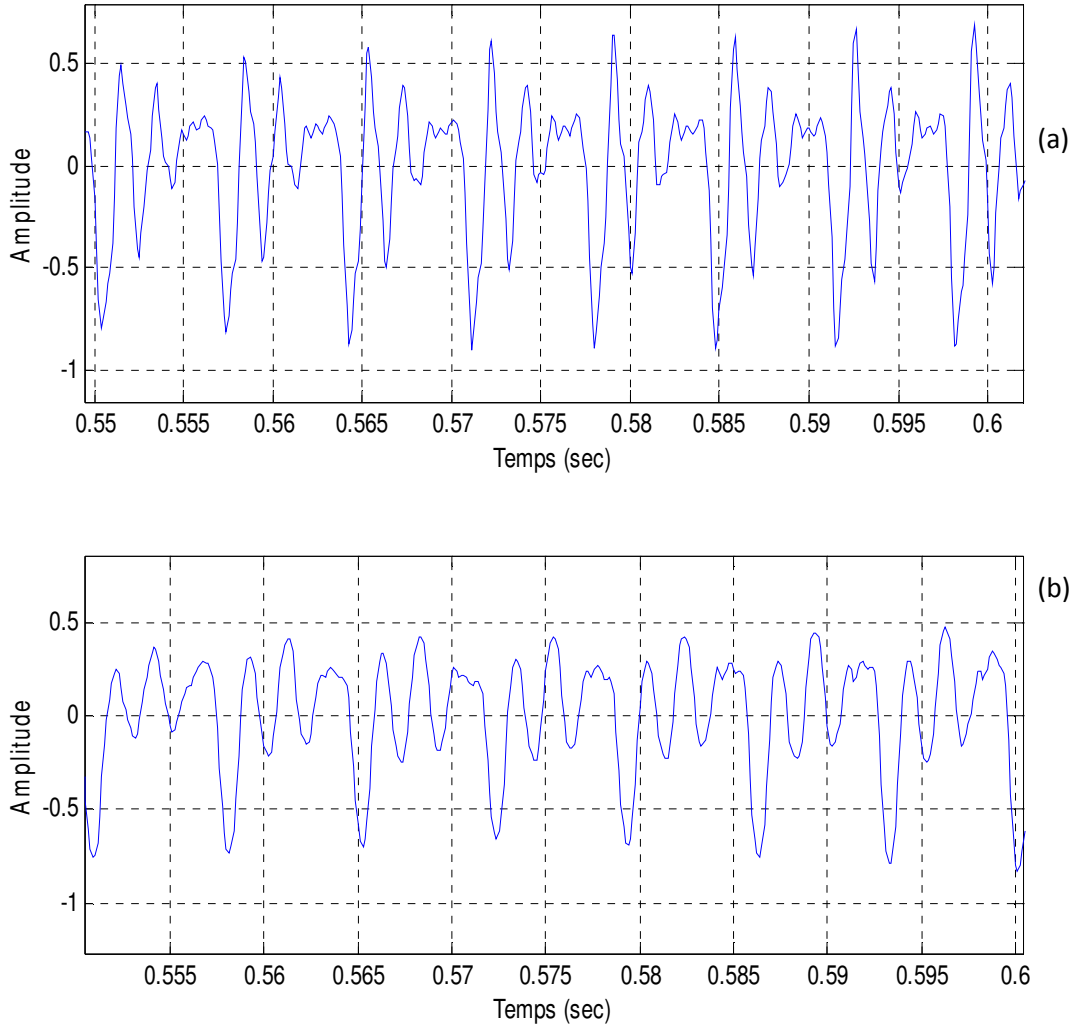


Fig. IV.13. Zone voisée pour un locuteur masculin a- original
b- synthétique

La figure IV.14 montre le signal original et le signal synthétique pour la même phrase arabe mais cette fois-ci prononcée par un locuteur féminin. Une zone voisée de ces deux signaux est donnée par la figure IV.15. Comme précédemment, on remarque une différence dans l'allure temporelle mais la périodicité est conservée pour les deux signaux.

Notons que dans notre cas, ces résultats ne justifient pas une bonne ou une mauvaise qualité perceptuelle. En effet la ressemblance entre les allures confirme certes l'obtention d'une bonne qualité perceptuelle, mais la différence ne peut pas l'infirmer.

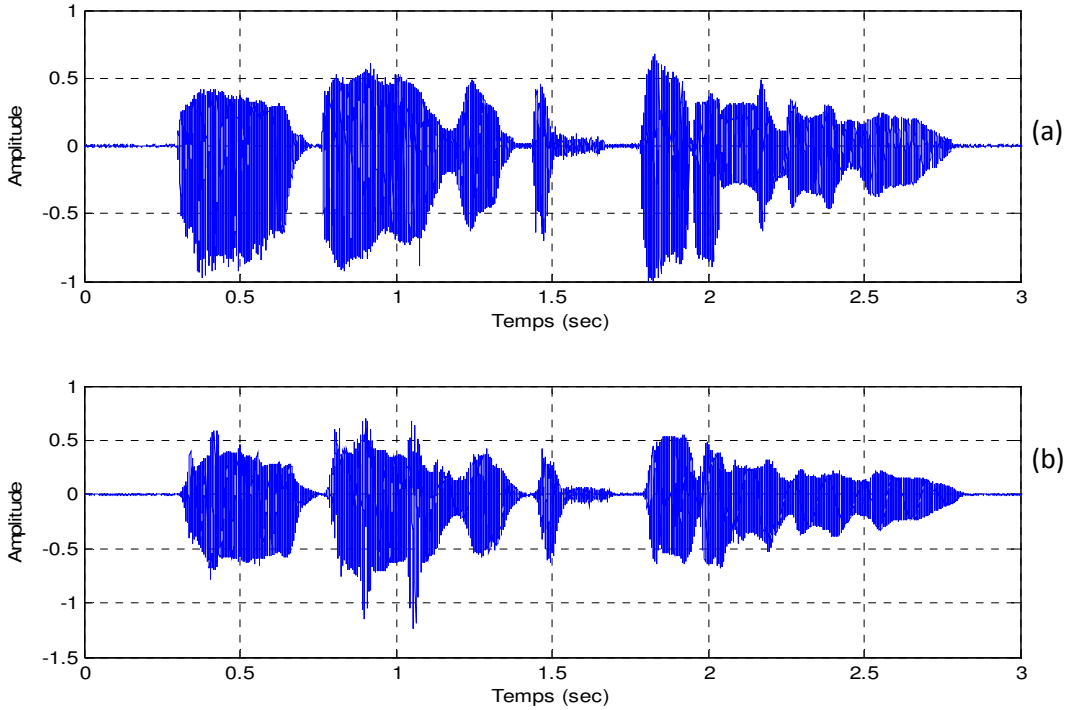


Fig.IV.14. Signaux de la phrase arabe « آذاه زحف رملـه » pour un locuteur féminin
a- Original
b- Synthétique

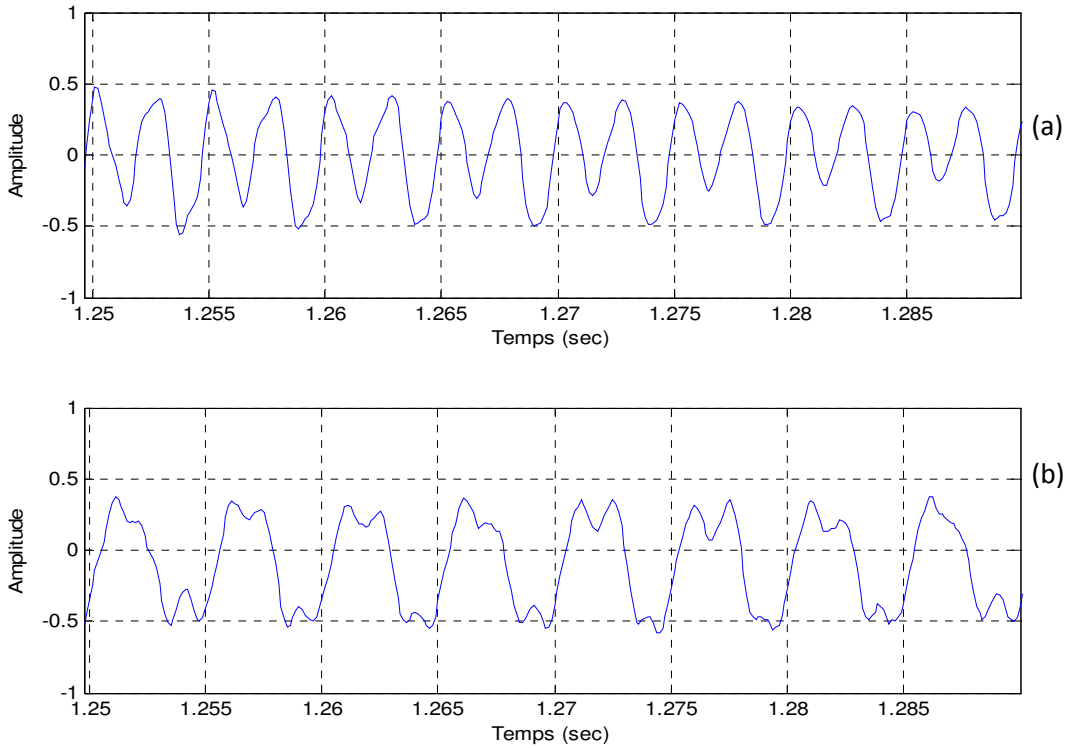


Fig. IV.15. Zone voisée pour un locuteur féminin a- original
b- synthétique

IV.7. Evaluation de la qualité perceptuelle

L'évaluation de la qualité d'un codeur paramétrique fonctionnant à très bas débit comme le MELP par les méthodes objectives telles que le SNR ou le SNR segmental est impossible, à cause de la non ressemblance partielle ou totale entre les allures temporelles des signaux original et synthétique [1]. Pour ces raisons nous avons fait appel à des techniques d'évaluation opérant dans le domaine spectral et qui sont plus adéquates.

Dans ces tests nous avons utilisé le corpus arabe PAPE et des phrases françaises et anglaises. Ces phrases ont été prononcées par des locuteurs masculins et féminins et enregistrées dans les conditions suivantes : milieu calme (quiet room) et milieu bruité. La fréquence d'échantillonnage de ces phrases est de 8 KHz avec une quantification à l'entrée de 16 bits.

Dans le but de chiffrer la qualité perceptuelle de notre codeur, nous avons utilisé les techniques d'évaluation suivantes :

- PESQ
- LLR
- WSS

Nous avons d'abord effectué nos tests sur des phrases acquises en milieu ambiant. Nous avons pris comme exemple les phrases suivantes :

- 3 phrases arabes, extraites d'un large corpus de phrases arabes phonétiquement équilibrées [5] :
  - Ph1 : « صـعد الإمام فـوق المنـــبر »
  - Ph2 : « آذاه زحف رملـه »
  - Ph3 : « نمنــم ماء اليــوم »
- Une phrase extraite du corpus anglais phonétiquement équilibré
  - Ph4 : « than the sun shine up warmly »
- Un enregistrement en langue française
  - Ph5 : « un deux trois »

Le tableau IV.2. montre les résultats obtenus pour ces phrases.

Tableau IV.2. Résultats des tests objectifs

Phrases	**Locuteur**	**PESQ**	**LLR**	**WSS**
Ph1	Masculin	2.54	0.59	42.48
Ph2	Masculin	2.68	0.70	43.54
Ph3	Féminin	2.47	0.50	74.69
Ph4	Masculin	2.38	0.78	52.88
Ph5	Masculin	2.35	0.74	32.28

On constate que le PESQ varie entre 2.68 et 2.35 pour l'ensemble des phrases. Les trois premières phrases font partie de la base de données utilisée dans la mise au point des dictionnaires, ce qui justifie la meilleure qualité obtenue. Les deux autres phrases ne font pas partie de cette base et elles sont prononcées dans une langue différente. Donc la perte de la qualité observée en comparaison avec les trois premières est un résultat prévu.

Le rapport LLR (rapport de vraisemblance logarithmique) mesure la distance entre les enveloppes spectrales des signaux original et synthétique on se basant sur les coefficients LP, donc logiquement une bonne similitude entre les enveloppes spectrales correspond à une faible valeur du LLR. Les résultats que nous avons obtenus avec cette méthode ne sont pas les mêmes que ceux obtenus avec la méthode PESQ. Ils sont même différents des résultats des tests d'écoute. Ce constat peut être justifié par le fait que les coefficients LP sont utilisés pour modéliser la production de la parole et non pas la perception où on utilise d'autres coefficients comme les MFCC (Mel-Frequency Cepstral Coefficients) [56] ou les coefficients PLP (Perceptual Linear Predictive) [57]. Le même constat peut être fait sur les WSS. Donc, même la distance spectrale ne donne pas une idée précise sur la qualité perceptuelle du signal synthétique.

En comparaison avec la qualité perçue lors des tests d'écoute, les résultats obtenus avec la méthode PESQ semblent d'être les plus adéquat.

Pour le corpus arabe PAPE nous avons obtenu un PESQ moyen de 2.5. Ceci classe notre codeur dans la catégorie 'acceptable' selon la norme P862 de l'ITU [29].

IV.8. Test de robustesse

Pour les tests de robustesse, nous avons considéré des phrases extraites du corpus PAPE et nous les avons bruitées avec un bruit blanc additif dont on augmente à chaque fois l'intensité. Nous avons utilisé le SNR segmental pour chiffrer le niveau du bruit ajouté au signal original. Nous avons également utilisé la méthode PESQ pour évaluer la qualité perceptuelle du signal synthétique obtenu à la sortie de notre codeur. Le tableau IV.3 résume les résultats obtenus lors de ces tests.

Le graphe de la figure IV.16 montre l'évolution de la qualité perceptuelle du signal synthétique, représentée par le PESQ, en fonction du niveau du bruit aditif ajouté au signal original. Malgré la diminution observée de la qualité perceptuelle du signal lorsqu'on augmente le niveau du bruit, on constate que le signal synthétique reste intelligible et cela jusqu'à un niveau de bruit de 40 dB (correspondant à un SNR de -12.19). Au delà de 40 dB, le signal original est complètement noyé dans le bruit et le signal synthétique devient non intelligible.

Tableau IV. 3. Tests de robustesse du MELP1.2 mis au point

Niveau de bruit (dB)	**SNR (dB)**	**PESQ**
0	27.8	2.54
5	22.8	2.48
10	17.8	2.42
20	7.8	2.38
40	-12.19	2.29
80	-52.19	1.94
160	-132.19	1.40

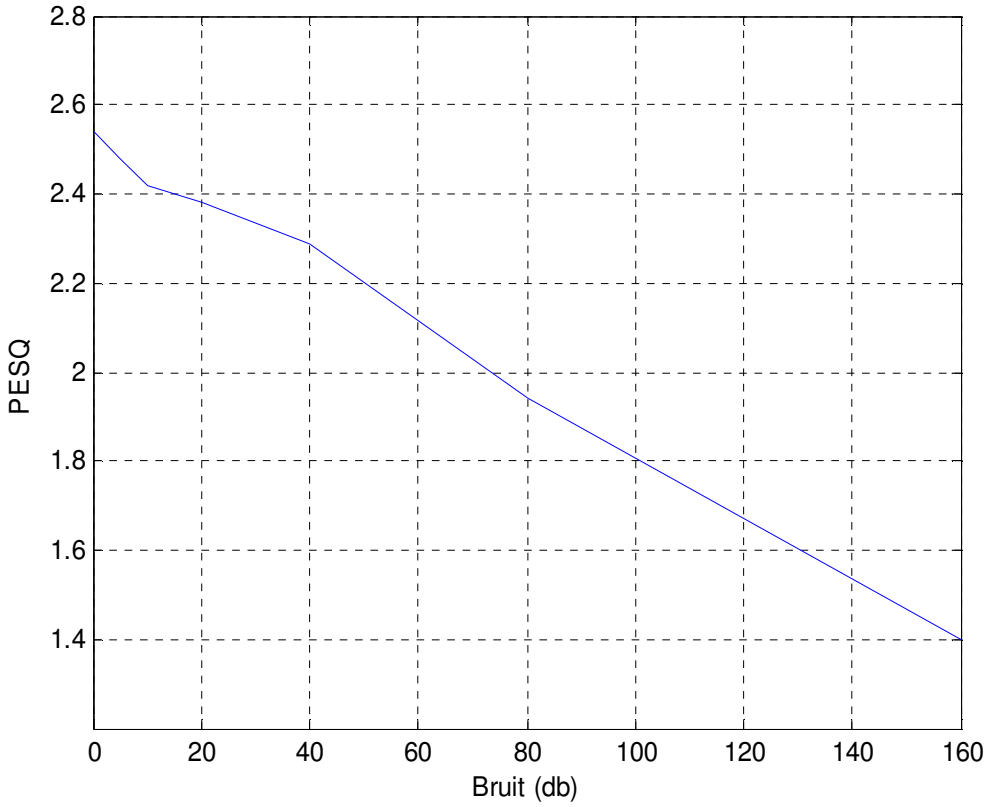


Fig. IV. 16. Robustesse du codeur VLBR mis au point

IV.9. Conclusion

Dans ce chapitre nous avons présenté l'évaluation des résultats obtenus par codeur/décodeur MELP fonctionnant à 1.2 kbps et implémenté sous MATLAB. Le bon fonctionnement de ce codeur prouvé par les tests d'évaluation effectués sur les différents blocs implémentés permet d'affirmer que notre codeur présente une qualité perceptuelle satisfaisante si on prend en considération le gain en débit réalisé. Par rapport au standard MELP 2.4 kbps, une réduction de débit de moitie a été obtenue. Elle est de 75% par rapport au standard FS1016 fonctionnant à 4.8 kbps. De plus, l'algorithme d'analyse utilisé est plus simple. Notre codeur MELP 1.2 kbps est meilleur que le FS1015 en termes de débit (une réduction de 50%) et en termes de qualité.

CONCLUSION GENERALE ET PERSPECTIVES

Le travail que nous avons présenté ici s'inscrit dans les objectifs des travaux de recherche de l'équipe de codage du laboratoire de communication parlée et du traitement du signal (LCPTS).

Dans ce travail nous avons présenté le codage de la parole à bas et très bas débit. Dans ce mémoire nous avons d'abord, présenté l'état de l'art relatif à ce type de codage. Cette étude montre que la mise au point d'un codeur fonctionnant à très bas débit nécessite l'utilisation d'une approche paramétrique. Dans ce cas, on fait souvent recours au codage à base de prédiction linéaire.

Ainsi, nous avons présenté trois algorithmes de codage prédictif utilisés dans trois standard DoD (le FS1015 qui utilise la LPC classique, le FS1016 basé sur l'algorithme CELP et le FS1017 qui utilise le MELP). Notre choix a logiquement été orienté vers les algorithmes MELP qui présentent un rapport débit/qualité intéressant où nous avons fixé comme objectif la mise au point et l'implémentation sous MATLAB d'un codeur MELP fonctionnant à 1.2 kbps.

Pour atteindre notre objectif nous avons utilisé la prédiction linéaire à excitation mixte telle quelle est décrite par le standard FS1017 (MELP fonctionnant à 2.4 kbps). Pour réduire le débit et obtenir un codeur VLBR, nous avons opté pour une approche multi-trames que nous avons appliquées au codeur FS1017. Cette approche a nécessité l'utilisation de plusieurs quantifications scalaires et vectorielles. Ainsi, pour quantifier les coefficients LSF, les amplitudes de Fourier et les gains, nous avons retenu une quantification vectorielle conjointe. Par la suite, une technique d'interpolation a été appliquée aux coefficients LSF. En ce qui concerne le pitch, deux quantifications conjointes ont été employées, l'une scalaire et l'autre vectorielle.

Ainsi, une répartition judicieuse des 81 bits alloués à ce codeur sur une super-trame de 67.5 ms a permis d'obtenir un codeur MELP présentant une qualité PESQ moyenne de 2.5 ms. Ce qui classe notre codeur dans la catégorie « acceptable » selon les recommandations de l'UIT-T.

Les perspectives de ce travail sont nombreuses. Déjà, il est actuellement utilisé pour convoyer les pertes des paquets lors d'une transmission de la voix su IP. En effet, il est exploité dans chaque paquet transmis et permet de recouvrer jusqu'à trois trames perdues.

Par ailleurs, ce codeur va constituer une base pour d'autres mise au point de codeurs VLBR dont les débits seront encore plus bas. Sa structure modulaire et sont langage de programmation MATLAB faciliteront cette mise au point.

REFERENCES BIBLIOGRAPHIQUES

[1] W.C.Chu, "Speech coding algorithms". Ed Wiley, USA, 2003.

[2] M.Jelinek, ''Modélisation spectrale et compression de la parole à bas débit''. Thèse de doctorat, université de Sherbrook, 1998.

[3] J.Holmes, W.Holmes, "Speech synthesis and recognition". Ed Taylor & Francis, USA, 2001.

[4] B.Boudraa, "Analyse et synthèse multi-impulsionnelle". Thèse doctorat d'état, université des sciences et de la technologie Houari Boumediene, 2006

[5] R.Goldberg, L. Riek, "A practical handbook of speech coders". Ed CRC press LCC, USA, 2000.

[6] S.Furui, M.M..Sondhi, "Advences in speech signal processing". Ed Marcel dekker, USA, 1992.

[7] X.Huang, A.Acero, H.W.Hon, "Spoken language processing". Ed Prentice Hall PTR, USA, 2001.

[8] E.Paksoy and all, "An adaptive multi-rate speech coder for digital cellular telephony''. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp.I-193-196, 1999.

[9] L.R.Rabiner, R.W.Schafer, "Digital processing of speech signal". Prentice Hall, USA, 1978

[10] A.Spanias, T.Painter, V.Atti, "Audio signal processing and coding". Ed Wiley, USA, 2007.

[11] A.S. Spanias, "Speech Coding : A Tutorial Review", Proc.IEEE, 82 (1994), pp. 1541–1582.

[12] I. Mcloughlin, "Applied speech and audio processing with MATLAB examples". Ed Cambridge, Singapore, 2009.

[13] J.D.Gibson, "The communication handbook". Ed CRC press LCC, USA, 2002.

[14]A.M.Kondoz, "Digital speech for low bit rate communication systems". Ed Wiley, England, 2004.

[15] L.Hanzo, F.C.Somerville, J.Woodard, "Voice and audio compression for wireless communication" ". Ed Wiley, England, 2007.

[16] J.Makhoul, R.Viswanathan, R.Schwartz, A.W.F.Huggins, "A mixed-source model for speech compression and synthesis", IEEE International Conference on Acoustics, Speech, and Signal Processing, pp.III-163-166, 1978.

[17] A.V.Mcree, T.P.Barnwell, "A Mixed excitation LPC vocoder model for low bit rate speech coding". IEEE Transaction on Speech and Audio Processing, pp. III-242-250, 1995.

[18] D.G.Row, "Techniques for harmonique sinusoidal coding'', Thèse de doctorat, School of physics and electronic systems engineering, 1997.

[19] R.J.Mcaulay, T.F.Quatieri, "The sinusoidal transforme coder at 2.4 kbps'', Military Communications Conference, 1992. MILCOM '92, Conference Record. 'Communications - Fusing Command, Control and Intelligence', IEEE, pp. I-378-380

[20] D.W.Griffin, J.S.Lim, "Multiband excitation vocoder''. IEEE Transaction on Speech and Audio Processing, pp.XXXVI-1223-1235, 1995.

[21] N.Moreau, "techniques de compression des signaux''. Masson, France, 1995.

[22] G.Madre, "Application de la transformée en nombres entiers à l'étude et au développement d'un codeur de parole pour transmission sur réseaux IP". Thèse de doctorat, université de Bretagne occidentale, 2000.

[23] J.Makhoul, "linear prediction : A tutorial review ". Proceedings of the IEEE, pp. LXIII-561-580, 1975.

[24] M.Djamah, M.Boudraa, B.Boudraa, M.Bouzid, "Quantification adaptative des coefficients LSF pour le codage de la parole a bas débit". 25é édition des journées d'étude sur la parole, palais des congrès de Fès, Maroc, 19-22 avril 2004.

[25] P.Vary, R.Martin, "Digital speech transmission enhancement, coding and error concealement". Ed Wiley, England, 2006.

[26] J.Ma, Y.Hu, P.C.Loizou, 'Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions'. Journal of Acoustical Society of America, 2009.

[27] ITU-T, "Mesure objective de la qualité des codecs vocaux fonctionnant en bande téléphonique (300 - 3400 Hz) ". ITU-T recommendation P.861, 1996.

[28] ITU-T, "Mesure objective de la qualité des codecs vocaux fonctionnant en bande téléphonique (300 - 3400 Hz) ". ITU-T recommendation P.861, 1998.

[29] ITU-T, 'Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs'. ITU-T recommendation P.862, 2001

[30] B.S.ATAL, S.L.HANAUER, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave". Journal of the Accoustic Society of America, 1972.

[31] M.R.Schroeder and B.S.Atal, "Code-excited linear prediction (CELP): high-quality speech at very low bit rates". IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. X.937–940, 1985.

[32] Fcderal Standard 1015, "Telecommunications: Analog to Digital Conversion of Radio Voice by 2400 bit/stcond Linear Rcdictive Coding". National Communication System Office of Ttchnology and Standards, 1984

[33] J. P. CampbeI1, V. C. Welch et T. E. Tremain, 'The New 4800 bps Voice Coding Standard," Roc. Military Speech Tech., p. 64-70, 1989.

[34] L.M.Supplee, R.P.Cohn, J.S.Collura, A.V. McCree, 'MELP: The new federal standard at 2400 BPS". IEEE International Conference on Acoustics Speech and Signal Processing, pp.II-1591-1594, 1997.

[35] B. **S.** Atal, M. R. Schroeder, "Predictive coding of speech signals," in Proc. Conf. Commun., Processing, Nov. 1967, pp.360-361.

[36] B.S.Atal, "predictive coding of speech at low bit rates". IEEE transactions on communications, 1982

[37] J. P. CAMPBELL, T. E.TREMAIN, "Voiced/Unvoiced Classication ofSpeech with application to the U.S. Government LPC10e Algorithm." IEEEICASSP, 1986.

[38] J.D.Gibson, "Speech coding methods, standards and applications". Circuits and systems magazine, IEEE, pp. V-30-49, 2005

[39] A.V. Mcree, T.P.Barnwell, 'Implementation and evaluation of a 2400 BPS mixed excitation LPC vocoder'. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. II-159-162, 1993.

[40] T.T.Teo, E.C.Tan, 'Implementation of 2400 bps melp vocoder on tms320c44'. Proceedings of the IEEE International Conference on Signal Processing, pp. I-576-579, 1998.

[41] A.V. Mcree, J.C.de Martin '1.7 KBPS MELP vocoder with improvement analysis and quatization'. Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing, pp. II-593-596, 1998.

[42] A.V. Mcree, J.C.de Martin, '1.6 KBPS MELP vocoder for wireless communication'. Proceedings of the IEEE Workshop on Speech Coding for Telecommunications, pp. 23-24, 1997.

[43] T.Wang, K.Koishida, V.Cuperman, A.Gersho, J.C.Collura, 'A 1200 BPS speech coder based on MELP'. Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing, pp. III-1375-1378, 2000.

[44] M.Chamberlain, 'A 600 bps MELP vocoder for use on HF channels'. IEEE Military Communications Conference, pp. I-447- 453, 2001.

[45] F.Khlili, M.Ardeliripour, H.Sqmeti, "design and implementation of vector quantizer for 600 bps vocoder based on MELP", IEEE 2009.

[46] Y.Medan, E.Yair, D.Chazan, 'Super resolution of pitch determination for speech signal'. IEEE Transactions on Signal Processing, pp. XXXIX-40-48, 1991

[47] T.T.Teo, E.C.Tan, 'real time implementation of MELP vocoder'. Journal of The Institution of Engineers, Singapore Vol. 44 Issue 3 2004

[48] M. Boudraa, B. Boudraa, B. Guerin, 'Mise en place de phrases arabes phonétiquement équilibrées'. *XIX ème JEP*, Bruxelles, 1992.

[49] Mathworks, 'Vector quantization disign', site internet: mathworks.com

[50] B.Boudraa, 'Analyse et synthèse multi-impulsionnelle'. Thèse doctorat d'état, université des sciences et de la technologie Houari Boumediene, 2006.

[51] S.Grassi, 'Optimized implementation of speech processing algorithm'. Thèse de doctorat, université de Neuchatel, 1998.

[52] K.Paliwal, B.S.Atal, 'Efficient vector quantization of LPC parameters at 24bit/frame' IEEE Transaction on Speech and Audio Processing, pp. I-3-14, 1993.

[53] J.Hansen, B.Pellom, 'An effective quality evaluation protocol for speech enhancement algorithms'. Proceedings of the International Conference on Spoken Language Processing, pp.VII-2819–2822, 1998.

[54] S.R.Quackenbush, T.P.Barnwell, M.A.Clements, 'Objective Measures of Speech Quality', Prentice-Hall, Englewood Cliffs, 1988.

[55] D.H.Klatt, 'Prediction of perceived phonetic distance from critical band spectra: A first step'. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. II-1278–1281, 1982.

[56] L.R.Rabiner, R.W.Schafer, 'Digital processing of speech signal'. NOW, USA, 2007

[57] H. Hermansky. 'Perceptual linear predictive (PLP) analysis of speech'. Journal of Acoustical Society of America, pp.IV-1738–1752, 1990.