REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE DE BLIDA
INSITITUT D'ELECTRONIQUE

MEMOIRE

Présenté par : Mr AIT SAADI HOCINE

EN VUE DE L'OBTENTION DU DIPLOME
DE MAGISTER EN ELECTRONIQUE

OPTION : COMMUNICATION

THEME

Amélioration des performances des codeurs CELP par l'utilisation d'un filtre prédicteur long-terme à haute résolution

Devant le jury :

Mr M.BENSEBTI	Maître de conférence (USTB)	Président
Mme M.GUERTI	Maître de conférence (ENP d'Alger)	Examinateur
Mr M.AREZKI	Chargé de cours (USTB)	Examinateur
Mr M.HALIMI	Chargé de recherche (CDTA)	Rapporteur
Mr A.GUESSOUM	Professeur (USTB)	Co-Rapporteur

Introduction Générale

Le monde a connu dernièrement des progrès et des activités de recherches énormes dans le domaine de la compression de la parole. Les applications liées à ce domaine telles que la communication avec les mobiles ne cessent de croître car elles peuvent être réalisées maintenant à coût réduit.
Les systèmes de compression de la parole sont caractérisés par quatre éléments essentiels :

- Le débit de transmission qu'il requiert entre le codeur (émetteur) et le décodeur (récepteur).
- La qualité d'écoute du signal vocal reconstruit.
- La complexité algorithmique nécessaire à sa mise en œuvre.
- L'efficacité spectrale du codeur.

Le but des recherches est alors de trouver le meilleur compromis *qualité /débit / complexité* . On parle dans ce cas d'un codage efficace.
Le codeur prédictif linéaire et excité par codes CELP (Code-Excited Linear Prediction) depuis son introduction en 1985 par M.Schroeder et B.Atal est le codeur le plus largement utilisé pour compresser les signaux de parole en bande téléphonique (200 Hz – 3.4 kHz) et à des débits inférieurs ou égaux à 16 kbps. Ce codeur permet de réduire le débit de transmission tout en conservant une bonne qualité auditive du signal reconstruit. L'amélioration des performances de ce type de codeur est cependant acquise au prix d'une importante complexité algorithmique. De nombreux travaux ont été entrepris autour du codage de type CELP avec le même objectif maintenir ou améliorer si possible la qualité tout en réduisant la complexité. La plupart des méthodes de compression cherchent à exploiter la redondance propre du signal. Il suffit de transmettre l'information non prédictible. Les codeurs de parole normalisés ou en cours de développement sont généralement basés sur une quantification vectorielle du signal résiduel de parole.
En 1991 est apparu le standard FS 1016 à 4.8 kbps (proposé par le département de la défense américaine (US DoD)). Ce dernier présente un signal de parole reconstruit de bonne qualité à un

faible débit de transmission. Les progrès réalisés dans les algorithmes de codage et l'augmentation des capacités de mémorisation et de calcul des microprocesseurs de traitement du signal même si elles restent inférieures aux performances réclamées, permettent d'envisager le traitement en temps réel.

Les codeurs CELP se basent principalement sur la prédiction linéaire et les techniques d'analyse par synthèse (LPAS). On essaie donc par ces codeurs d'imiter le processus de reproduction du son humain où la source et le conduit vocal sont remplacés respectivement par une séquence d'excitation et un filtre de synthèse. Le signal de parole est découpé en trames ou blocs d'échantillons qui sont codés séparément. Les séquences d'excitation sont représentées par des vecteurs stockés dans un dictionnaire connu de l'émetteur et du récepteur. Le critère de choix est le critère des moindres carrés. Le signal d'excitation qui minimise l'énergie de l'erreur perceptuelle est sélectionné. Le dictionnaire étant connu de l'émetteur et du récepteur, la séquence d'excitation choisie est codée par son numéro ou index. Dans le cas d'un son voisé, la séquence d'excitation réelle est périodique et cette période correspond à la fréquence du fondamental (fréquence de vibrations des cordes vocales). La reproduction de la périodicité du signal est réalisée par un filtre de prédiction long-terme ou par un dictionnaire adaptatif. Pour cela, on reproduit la périodicité en allant chercher dans la mémoire des excitations des sous-trames précédentes, celle qui a un retard égal à la période. Comme l'oreille humaine est très sensible aux variations du fondamental, une mauvaise reproduction de la périodicité causera une dégradation de la qualité. Une estimation plus précise de la période s'avère donc nécessaire surtout pour les locuteurs féminins (fréquences hautes). Généralement le filtre prédicteur long-terme utilisé est d'ordre 1 mais on peut envisager d'augmenter l'ordre du filtre comme dans [1],[2]. Une autre méthode consiste à augmenter la résolution temporelle par une interpolation des échantillons composants le dictionnaire, c'est la méthode du pitch fractionnaire.

Le but de ce mémoire est de montrer l'amélioration apportée aux codeurs de type CELP, par une meilleure reproduction de la périodicité en utilisant la méthode du pitch fractionnaire. Cette méthode permet d'avoir une meilleure qualité de parole synthétique.

Ce mémoire est composé de cinq chapitres

- Le premier chapitre donne quelques généralités sur le modèle de production de la parole humaine et sur le codage.
- Le deuxième chapitre fournit les principes de la prédiction linéaire, les techniques d'analyse par synthèse utilisées par une large classe de codeurs en particulier les codeurs de type CELP, et une description de l'algorithme du codage CELP.
- Dans le troisième chapitre, on trouve des définitions sur la quantification scalaire et vectorielle. On verra que cette dernière n'est pas uniquement une généralisation de la quantification scalaire mais elle permet de prendre en compte la corrélation du signal.
- Le quatrième chapitre est consacré à la description des méthodes de reproduction de la périodicité du signal dans les codeurs CELP.
- Le cinquième et dernier chapitre décrit la simulation et l'interprétation des résultats trouvés qui montrent clairement l'amélioration de la reproduction de la périodicité grâce à l'utilisation du pitch fractionnaire.

Chapitre 1

Généralités

1.1 Introduction

La parole est le moyen de communication privilégié entre les humains. Pour réaliser une représentation paramétrique du signal vocal, il faut exploiter une connaissance a priori du mécanisme qui produit ce signal. Ce chapitre expose quelques généralités relative au signal vocal et à ses caractéristiques et aussi sur le codage et la classification des codeurs.

1.2 Aspects physiologiques de la phonation

Les principaux organes composants l'appareil phonatoire sont les poumons, la trachée artère le larynx, le pharynx et les cavités buccales et nasales.

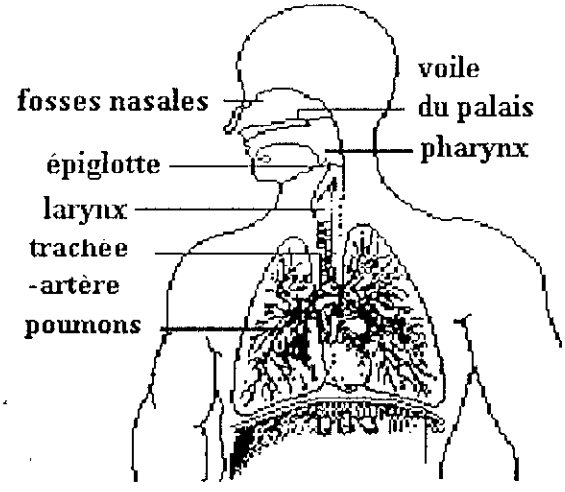


Figure 1.1 Appareil phonatoire humain

La parole est une onde de pression acoustique qui est due à des mouvements physiologiques volontaires de l'ensemble du mécanisme de production.

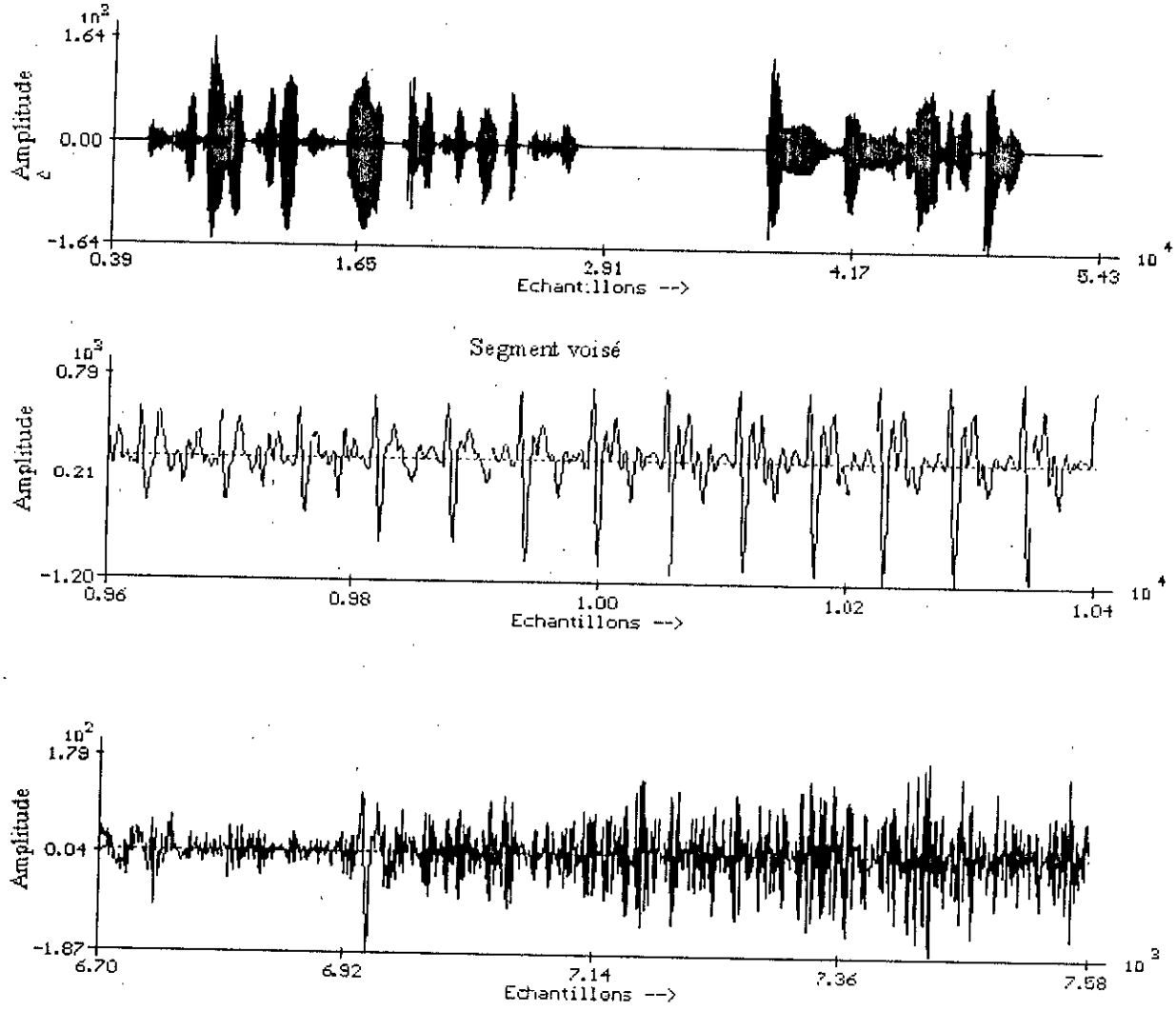


Figure 1.2 Exemple de deux segments de signaux de parole (voisé et non voisé)

L'exemple de la figure 1.2 comporte deux segments tirés des deux phrases suivantes :
« Il se garantira du froid avec un bon capuchon . Annie s'ennuie loin de mes parents ».

Les poumons débitent un courant d'air dans la trachée. Ce courant d'air est forcé à travers les cordes vocales. La hauteur de la voix, au cours d'une conversation, varie selon les personnes.

Essentiellement déterminée par la longueur, la forme et la position des cordes vocales, elle peut être volontairement modifiée, dans certaines limites, par l'intermédiaire des muscles respiratoires et de ceux du larynx, lesquels font varier la pression d'air. L'association de ces éléments détermine la fréquence de la vibration des cordes vocales : plus celle-ci est élevée, plus la voix est aiguë. L'articulation est la prononciation des sons en les différenciant pour former les mots. La parole est articulée en interrompant et en modulant le flux d'air à l'aide des lèvres, de la langue, des dents, de la mâchoire et du palais.

Un son est dit voisé lorsque les cordes vocales sont excitées. Lorsqu'un son voisé est prononcé, les cordes vocales s'ouvrent progressivement, sous l'action de la pression de l'air. Elles laissent apparaître une ouverture de forme triangulaire et d'étendue variable : la glotte. Elles se mettent alors à vibrer donnant ainsi naissance à une onde glottique caractérisée par des variations impulsionnelles périodiques de la pression et du débit de l'air.

Les sons voisés sont générés par une onde glottique (pharynx, cavités buccales et nasales). L'ouverture de la glotte changera en fonction des longueurs, épaisseur et tension des cordes vocales. Plus la tension des cordes vocales est élevée, plus la fréquence fondamentale ou fréquence de mélodie (pitch en anglais) sera élevée. Cette fréquence varie en moyenne de 70Hz à 150Hz pour les hommes et de 150 à 400Hz pour les femmes et les enfants.

Les sons fricatifs résultent de l'écoulement de l'air dans une constriction étroite située en un point du conduit vocal, en particulier au niveaux des lèvres et des dents. Les sons fricatifs sont non voisés (f, s,...) ou voisés (v, ...). Un son plosif (ou occlusif) est produit par une ouverture brusque ; il peut être voisé (b, d, ...) ou non voisé (p, t,...) [3], [4].

1.3 Spectre d'un signal de parole

1.3.1 La source

Le signal de source est l'onde que l'on obtient à la sortie de la glotte, il peut être périodique ou non. Lorsque l'onde glottique est périodique, elle peut être développée en série de Fourier. La première raie sera la fréquence fondamentale F_0.

Les autres raies espacées de F_0, sont les harmoniques. Ce train d'impulsions est très étendu en fréquence.

1.3.2 Cavités supra glottiques

L'ensemble des cavités supra-glottiques constitue un filtre assimilable à un ensemble de résonateurs qui renforcent ou atténuent sélectivement les composants du spectre du signal de source, c-à-d l'onde glottique. Les résonances sont appelées formants. Ces formants sont modifiés en fonction de la position des articulateurs (langue, voile du palais et lèvres). Un son est donc le résultat du filtrage de l'onde glottique par les cavités supra-glottiques. On observe généralement dans le spectre du son entre 50 Hz et 4 kHz quatre formants principaux notés F1, F2, F3, F4.

1.4 Propriétés statistiques d'un signal de parole

Le signal de parole peut être considéré comme un processus aléatoire non-stationnaire, c'est à dire que ses propriétés statistiques changent au cours du temps. La non-stationnarité de la parole résulte des changements au cours du temps aussi bien de la source que de la forme et des dimensions du conduit vocal. D'une manière générale, le signal de sortie d'un système linéaire est stationnaire si les caractéristiques du système sont invariables en fonction du temps et si le signal d'excitation est stationnaire.

Dans la pratique, les caractéristiques du conduit vocal et de la source évoluent lentement (sauf pour certains sons comme les plosives) et on fera l'hypothèse de quasi-stationnarité sur des périodes allant de 10 à 30 ms. On pourra donc appliquer pour le traitement des signaux de parole les méthodes classiques du traitement du signal en prenant certaines précautions et toujours au prix de certains compromis [3], [4].

On verra plus loin la méthode d'analyse par synthèse qui consiste à représenter le signal vocal original sous forme d'un ensemble de paramètres définissant un modèle. Les paramètres sont déterminés afin que le modèle fournisse un signal de parole synthétique le plus proche possible au sens d'un critère prédéfini, du signal vocal original.

1.5 Le codage

Un signal de parole numérique offre de nombreux avantages tels que l'immunité au bruit, la facilite de stockage, la commodité d'emploi que ce soit pour le multiplexage, le cryptage ou la synthèse.

1.5.1 Système de transmission élémentaire

Nous allons définir les différents éléments constituant un système de compression élémentaire de la parole .

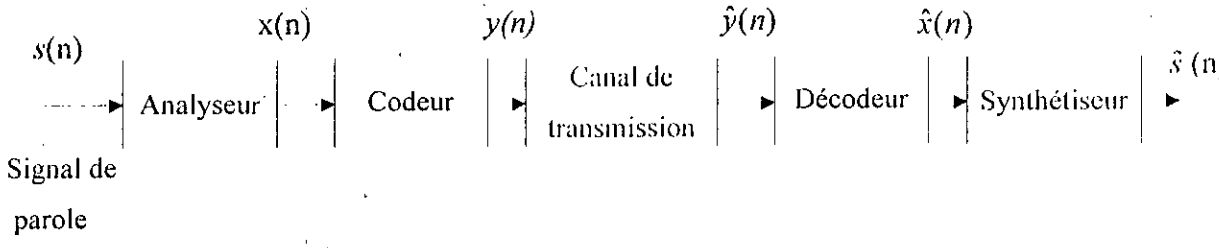


Figure 1.3 Système de transmission de parole

- Le premier élément dans la figure 1.3, analyse le signal de parole qui aura été filtré et échantillonné au préalable, la sortie de l'analyseur est un vecteur x d'éléments non quantifiés.
- Le codeur quantifie puis code le vecteur x pour la transmission.
- Le canal de transmission transmet le vecteur y des éléments codés
- Le décodeur décode le vecteur reçu $\hat{y}$ et on extrait un ensemble de paramètres.
- Le synthétiseur utilise ce vecteur de paramètres $\hat{x}$ (n) pour reconstruire le signal de parole. L'objectif d'un système de compression est de réduire le débit binaire exprimé en bits par secondes (bps) de l'information y(n) à transmettre tout en gardant une qualité satisfaisante de la parole synthétique.

L'analyse du signal détermine l'efficacité du système de compression de parole. Pour les codeurs les plus élémentaires, elle est inexistante. Le type d'analyse à réaliser sur un signal est fixé par le type de synthèse. Le bloc "synthèse" est généralement constitué d'une fonction d'excitation et d'une fonction de transfert. Le synthétiseur détermine le nombre de paramètres nécessaires à la synthèse. Une réduction supplémentaires de débit pourra se faire par un meilleur codage des paramètres.

Les deux fonctions principales du codeur sont la quantification et le codage.
On distingue la quantification scalaire qui attribue à une valeur de paramètre un nombre fini choisi dans un ensemble fini et connu de nombre, de la quantification vectorielle qui attribue à un groupe de valeurs un vecteur choisi dans un ensemble fixé de vecteurs appelé dictionnaire.

La quantification vectorielle est présentée comme une méthode de suppression de redondances c'est à dire des liens qui existent entre les différents paramètres du vecteur [5]. Elle utilise quatre principales propriétés :

- Dépendance linéaire (corrélation)
- Dépendance non linéaire.
- Forme de la fonction de densité de probabilité.
- Dimension des vecteurs .

Le codage traduit ces nombres choisis en séquences de nombres binaires qui seront transmises au décodeur. Selon les besoins, chacune de ces opérations peut être améliorée.
La réalisation d'un système de codage efficace dépendra de quatre paramètres :

- Les caractéristiques du signal de parole.
- Le débit de transmission.
- La qualité de parole synthétique.
- Le coût du système.

1.5.2 Suppression de la redondance

La suppression partielle des redondances permet une représentation plus efficace des données. La compression des données peut se faire sans perte d'information (exemple code de Huffman) ou avec perte d'information en exploitant dans ce cas la tolérance de l'organe récepteur (exemple l'oreille). Le signal de parole a des caractéristiques très particulières ; la compression du signal consistera à réduire les redondances. Ces redondances sont essentiellement : le manque de platitude du spectre court-terme, la quasi-périodicité des signaux voisés, la limitation des formes et des vitesses de mouvement possibles du conduit vocal, les distributions de probabilités non uniformes

des valeurs de paramètres de transmission. Les trois premières sont dues à des propriétés physiques du mécanisme de production de la parole. La dernière est fonction du codage utilisé.

Le manque de platitude du spectre court-terme est lié au fait que les échantillons de parole adjacents sont corrélés entre eux. On peut décorréler ces échantillons de parole par un filtrage spectral adapté. La quasi-périodicité des signaux de parole voisé peut être supprimée en utilisant un prédicteur long-terme. La lenteur du conduit vocal permet d'envoyer les paramètres des filtres toutes les 10-30 ms. La dernière des redondances citées peut être exploitée par un codage approprié.

1.5.3 Classification des codeurs

Le classement des codeurs de parole peut se faire selon différentes approches : débit obtenu, type de codage. Nous distinguerons les codeurs par formes d'ondes des codeurs de source ou vocodeurs et les techniques hybrides [4].

1.5.3.1 Les codeurs par formes d'ondes : Dans cette catégorie, on distingue les codeurs temporels et les codeurs fréquentiels. Ces codeurs n'utilisent aucune connaissance a priori sur la façon dont le signal est généré. Le codeur temporel fera correspondre à l'amplitude du signal analogique une suite d'éléments discrets. Le signal reconstruit est sans doute le plus proche du signal original. Ces codeurs sont conçus pour être indépendants du signal codé et peuvent coder n'importe quel son. Le débit de codage est généralement élevé. En utilisant les propriétés de corrélation du signal, il est possible de diminuer ce débit jusqu'à une certaine limite.

1.5.3.2 Les vocodeurs : Ces codeurs utilisent une méthode dite par analyse et synthèse, où l'on essaie d'extraire du signal de parole un ensemble de paramètres liés à un modèle simplifié. Ces paramètres sont l'enveloppe spectrale du spectre court-terme et des informations sur le signal d'excitation (pitch, amplitude, voisement). On suppose donc qu'on a des connaissances a priori sur le signal de parole. Ces codeurs sont sensibles aux bruits de transmission et la qualité de la parole est limitée. Le débit de transmission est généralement faible.

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE DE BLIDA
INSITITUT D'ELECTRONIQUE

MEMOIRE

Présenté par : Mr AIT SAADI HOCINE

EN VUE DE L'OBTENTION DU DIPLOME
DE MAGISTER EN ELECTRONIQUE

OPTION : COMMUNICATION

THEME

Amélioration des performances des codeurs CELP par l'utilisation d'un filtre prédicteur long-terme à haute résolution

Devant le jury :

Mr M.BENSEBTI	Maître de conférence (USTB)	Président
Mme M.GUERTI	Maître de conférence (ENP d'Alger)	Examinateur
Mr M.AREZKI	Chargé de cours (USTB)	Examinateur
Mr M.HALIMI	Chargé de recherche (CDTA)	Rapporteur
Mr A.GUESSOUM	Professeur (USTB)	Co-Rapporteur

1.5.3.3 Les codeurs hybrides : Ces codeurs font intervenir les techniques d'analyse par synthèse et les techniques de codages par formes d'ondes. Au prix d'une complexité parfois élevée, ils permettent d'obtenir une bonne qualité de signal à des débits intermédiaires (codeurs CELP).

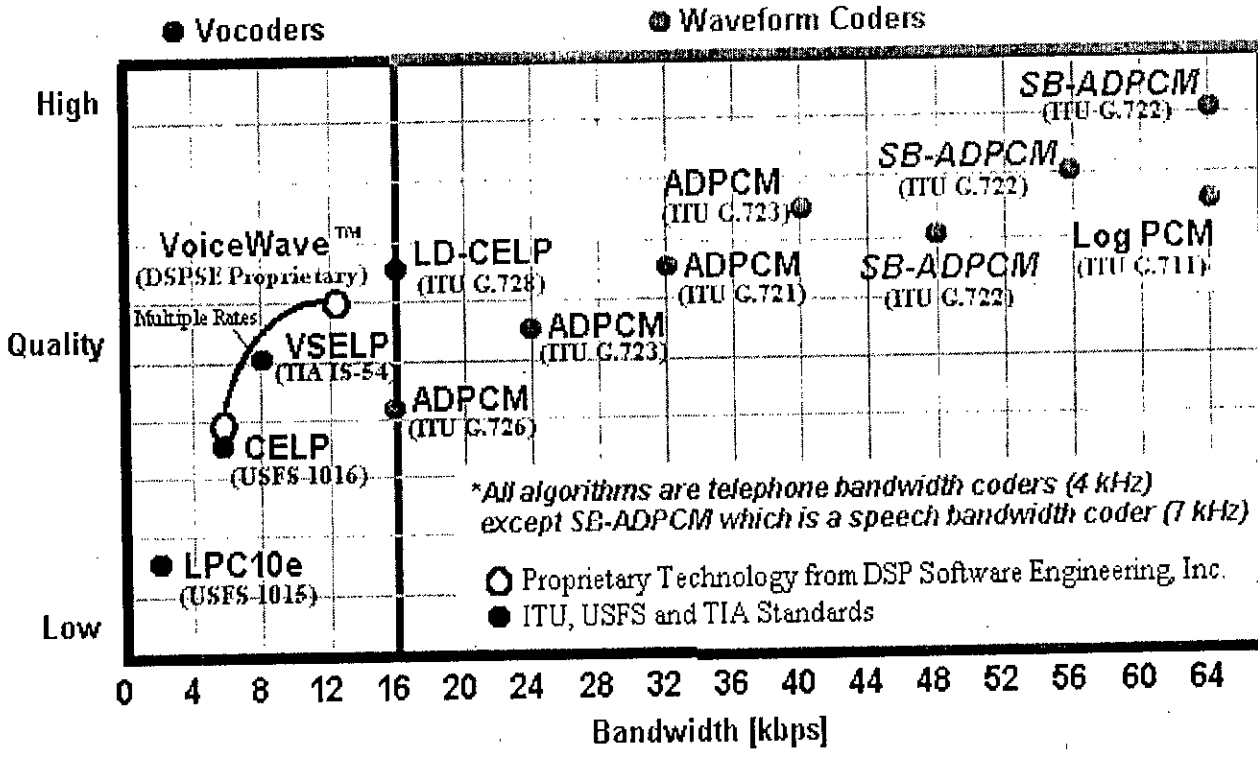


Figure 1.4 Relation entre débit de codage et la qualité de parole obtenue pour différents codeurs [24].

1.6 Conclusion

L'étude du mécanisme de la phonation permet de réaliser une modélisation de l'appareil phonatoire humain en tenant compte des caractéristiques du signal vocal. Le codeur hybride et tout particulièrement le codeur CELP permet de restituer un signal de parole de bonne qualité.

Chapitre 2

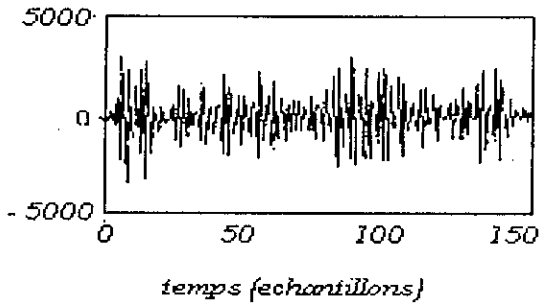
Technique d'Analyse par Synthèse et la Prédiction Linéaire

2.1 Introduction

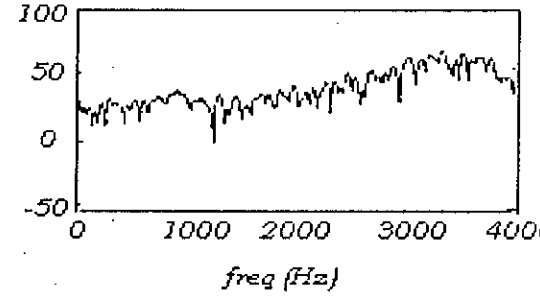
La production de la parole peut être vue comme une opération de filtrage dans laquelle une source sonore excite un filtre (conduit vocal). La source de son représente un bruit généré par une constriction du conduit vocal durant les sons non voisés ou des impulsions glottales durant les sons voisés ou une combinaison des deux. Le spectre de la source sonore pour les sons voisés contient des harmoniques espacées de F_0 avec une énergie plus concentrée aux fréquences basses alors que pour les sons non voisés le spectre est approximativement plat et sans structure harmonique (voir figure 2.1).

Le conduit vocal modifie la distribution d'énergie du spectre de la source sonore et introduit des résonances (formants) et des anti-résonances. Représentant le conduit vocal comme un filtre variant dans le temps, les résonances et les anti-résonances sont dues respectivement aux pôles et aux zéros de la réponse fréquentielle du conduit vocal. Les codeurs à faible débit tentent de réduire le débit, tout en préservant la qualité du son.

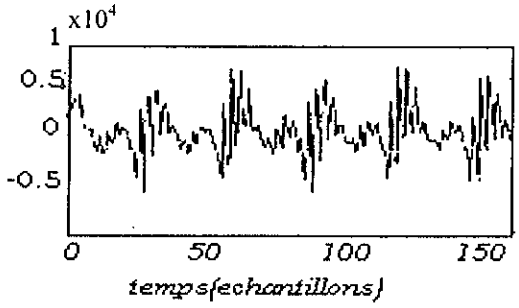
La redondance dans le signal de parole conduit à la conclusion que les échantillons de la parole sont corrélés. L'enveloppe spectrale correspond aux corrélations à court-terme et la structure harmonique correspond aux corrélations à long-terme.



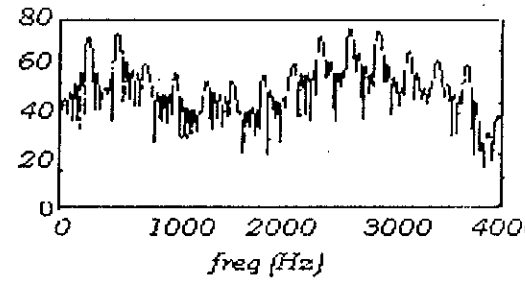
(a) Section d'un signal parole non voisé



(b) Spectre d'un segment non voisé



(c) Section d'un signal parole voisé



(d) Spectre d'un segment voisé

Figure 2.1 Exemples de segments de parole [7].

2.2 La prédiction linéaire

La prédiction linéaire est très largement utilisé dans les systèmes de codage de la parole. C'est une technique qui exploite les formes de redondances présentes dans le signal vocal, pour différentier la partie innovation de la partie prédictible. Elle est utilisée aussi bien en analyse qu'en synthèse de la parole.

Avec l'augmentation croissante de la puissance de calcul des microprocesseurs, la prédiction linéaire est employée dans de nombreux systèmes de production ou de codage du signal vocal.

L'idée de base de la prédiction linéaire est qu'un échantillon S_n peut être approximativement prédit par une combinaison linéaire des échantillons passés.

La prédiction linéaire est ainsi utilisée pour enlever les redondances du signal de parole. La suppression des redondances est réalisée avec un filtre prédicteur linéaire (LP) ou filtre d'analyse.

Le filtre d'analyse (LP) enlève la structure formantique du signal de parole. Le filtre d'analyse inverse (filtre de synthèse) modélise le conduit vocal et sa fonction de transfert décrit l'enveloppe spectrale du signal de parole. Aussi on peut utiliser un autre filtre qui exploite la périodicité du signal. L'inverse de ce filtre s'appelle prédicteur long-terme. Ce dernier modélise l'effet de la glotte et sa fonction de transfert décrit la structure harmonique du signal de parole.

2.2.1 Prédiction Court-terme

Le signal de parole peut être considéré comme la sortie d'un certain système avec une entrée inconnue d'excitation *u(n)*.

$$S(n) = \sum_{k=1}^{p} a_k \, S(n-k) + G \sum_{l=0}^{q} b_l \, U(n-l) \qquad \text{avec } b_0 = 1, \qquad (2.1)$$

où $\{a_k\}, \{b_l\}$ et le gain G sont les paramètres du système.

Comme le montre l'équation (2.1), le signal de parole est prédit comme une combinaison linéaire des sorties passées et des excitations courantes et passées.

La transformée en z du système, est ainsi donnée par :

$$H(z) = \frac{S(z)}{U(z)} = G \frac{1 + \sum_{l=1}^{q} b_l Z^{-l}}{1 - \sum_{k=1}^{p} a_k Z^{-k}}, \qquad (2.2)$$

où $S(z)$et $U(z)$ sont respectivement les transformées en Z de $S(n)$et$U(n)$.

$H(z)$ dans l'équation (2.2) correspond à un modèle Auto-Regressif à Moyenne Ajustée (ARMA). Les racines du numérateur et du dénominateur correspondent respectivement, aux zéros et aux pôles du système.

Ce modèle peut prendre deux cas particulier :

(1) Quand $a_k = 0$ pour $k = 1, \ldots, p$, $H(z)$ devient un modèle tous zéros ou bien un modèle à moyenne ajustée (MA).

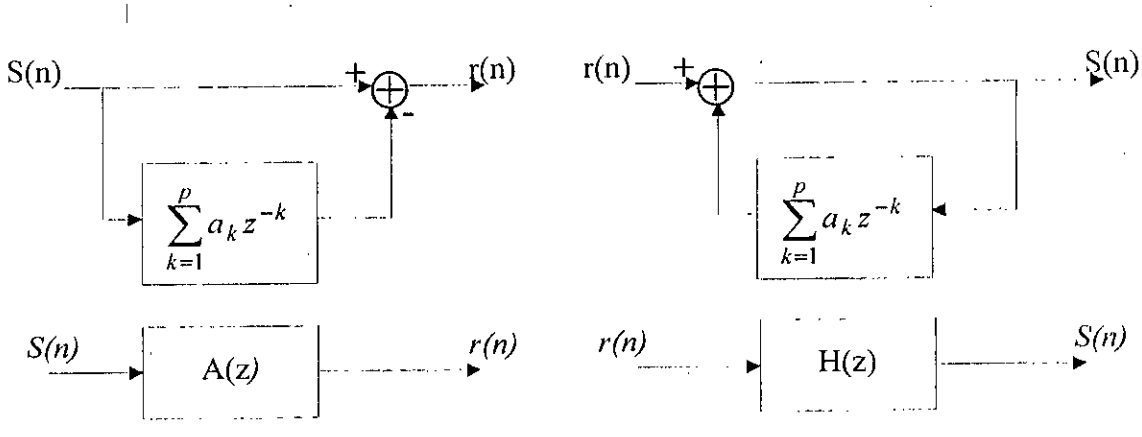
(2) Quand $b_l = 0$ pour $l = 1, \ldots, q$, $H(z)$ devient un modèle tous pôles ou bien un modèle auto-regréssif (AR).

Le modèle (AR) est utilisé pour la synthèse de la parole puisqu'il fournit une bonne représentation des effets du conduit vocal pour les voyelles, bien que ce modèle ne soit pas valable pour les sons nasalisés où la connexion du conduit nasal en parallèle avec le conduit vocal introduit des zéros dans la fonction de transfert. Il ne permet pas aussi de représenter le signal sur des zones fortement non stationnaires . Néanmoins, l'oreille humaine est plus sensible aux formants qu'aux vallées ce qui rend la simplification acceptable. De plus ; il a été montré que l'effet d'un zéro dans la fonction de transfert peut être obtenu en incluant plus de pôles [4], [6], [7].

En se basant sur le modèle AR (modèle tous pôles), l'échantillon parole courant est prédit par une combinaison linéaire de p échantillons passés. Ainsi on peut écrire l'équation suivante :

$$\hat{S}(n) = \sum_{k=1}^{p} a_k S(n-k) \tag{2.3}$$

où P est l'ordre du filtre AR et $\hat{S}(n)$ est le signal prédit



(a) Prédiction de la structure formantique (b) Synthèse de la structure formantique

Figure 2.2 Les étapes d'Analyse et de Synthèse Formant (a) et (b)

La figure 2.2 montre les deux étapes d'analyse et de synthèse lors de la prédiction linéaire court-terme (prédiction formant).

La sortie r (n) appelée erreur de prédiction ou signal résiduel est donnée par :

$$r(n) = S(n) - \sum_{k=1}^{p} a_k S(n-k). \tag{2.4}$$

En prenant la transformée en Z de l'équation (2.4) on aura :

$$R(z) = A(z)S(z) \tag{2.5}$$

où R(z) est la transformée en Z du signal résiduel r(n) et

$$A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k}. \tag{2.6}$$

Le filtre A(z) s'appelle filtre d'analyse. Le filtre tous pôles H(z) est le filtre de synthèse donné par :

$$H(z) = \frac{1}{A(z)} \tag{2.7}$$

Ce filtre modélise l'enveloppe spectrale de puissance court-terme du signal de parole.

Le choix de l'ordre de prédiction p résulte d'un compromis. Il doit être suffisamment grand pour reproduire correctement la structure formantique du signal de parole. En général une paire de pôles est allouée pour chaque formant, un ordre 8 est nécessaire pour créer quatre pics dans le spectre puisque le signal de parole comporte généralement quatre formants. Inversement, l'ordre doit être le plus faible possible pour économiser le débit. Pour une fréquence d'échantillonnage de 8 kHz, l'ordre p est généralement pris égal à 10.

Les coefficients $\{a_k\}$ connus sous le nom de LPC (linear prediction coefficients) sont estimés sur des intervalles de temps de courte durée, dans lesquelles on considère le signal de parole comme quasi-stationnaire. La détermination de ces coefficients est réalisée par minimisation au sens du critère des moindres carrés.

Dans la méthode des moindres carrés, le signal parole ou le signal erreur est multiplié par une fenêtre de pondération et l'ensemble des coefficients $\{a_k\}$ est choisi de façon à minimiser l'énergie du signal erreur.

Selon la pondération on aboutit à la méthode de covariance ou à la méthode d'autocorrelation. Dans la méthode d'autocorrelation le signal de parole est pondéré tandis que dans la méthode de

covariance c'est le signal erreur qui est pondéré. La méthode d'autocorrelation garantit la stabilité du filtre de synthèse H(z). Cette propriété fait que cette méthode d'autocorrelation est la plus utilisée pour l'estimation des coefficients du filtre.

Les coefficients du filtre doivent être codés. Dans la pratique, on ne quantifie pas directement les coefficients $a_1,\ldots,a_p$ du filtre A(z) car ils ont de mauvaises propriétés de codage. Il existe une autre représentation de ces coefficients qui offre de meilleures propriétés de codage (moins sensible au problème de la quantification). Il s'agit des paires de raies spectrales ou Line Spectrum Pairs (LSP). On en donne une représentation très succincte. A partir de l'équation (2.6) on construit ces deux polynômes d'ordre p+1.

$$B_1(z) = A(z) + z^{-p-1}A\left(z^{-1}\right)$$
$$B_2(z) = A(z) - z^{-p-1}A\left(z^{-1}\right) \tag{2.8}$$

Ces polynômes possèdent les propriétés suivantes :

- Le polynôme $B_1(z)$ est un polynôme symétrique. Le polynôme $B_2(z)$ est un polynôme antisymétrique.
- Si toutes les racines de A(z) sont à l'intérieur du cercle unité, toutes les racines de $B_1(z)$ et $B_2(z)$ sont sur le cercle unité.
- Les racines de $B_1(z)$ et $B_2(z)$ apparaissent de façon alternée sur le cercle unité.
- Si p est pair, on peut écrire $B_1(z)$ et $B_2(z)$ sous la forme :

$$B_1(z) = \left(1 + z^{-1}\right)\prod_{i=1}^{p/2}(1 - 2\cos\phi_{2i-1}\,z^{-1} + z^{-2}) \tag{2.9}$$
$$B_2(z) = \left(1 - z^{-1}\right)\prod_{i=1}^{p/2}\left(1 - 2\cos\phi_{2i}\,z^{-1} + z^{-2}\right)$$

- Si p est impair, on obtient :

$$B_1(z) = \prod_{i=1}^{(p+1)/2}(1 - 2\cos\phi_{2i-1}\,z^{-1} + z^{-2}) \tag{3.10}$$
$$B_2(z) = \left(1 - z^{-2}\right)\prod_{i=1}^{(p-1)/2}(1 - 2\cos\phi_{2i}\,z^{-1} + z^{-2}).$$

Résumé

La technique de la prédiction long-terme consiste à rechercher dans le passé du signal, une séquence qui, à un facteur prés, est la plus proche de la séquence traitée, au sens d'un critère choisi (ici de type quadratique).

La valeur trouvée M est la valeur entière la plus proche (comprise entre 20 et 147) de la valeur réelle de la période du fondamental. La numérisation du signal est la cause de cette approximation. Cette numérisation peut également engendrer une détermination d'un multiple de la période.

L'oreille humaine étant très sensible aux variations de la période du fondamental, une estimation précise de cette période s'avère nécessaire. Il s'offre alors à nous deux possibilités ; ou bien on augmente l'ordre du prédicteur (2 ou 3), ou bien on interpole les mots de code candidats dont l'index est proche de la valeur de la période.

La première option est généralement peu utilisée pour deux raisons ; premièrement parce que l'augmentation de l'ordre du prédicteur accroît le débit rendant le codage moins efficace, et deuxièmement parce que ces prédicteurs posent des problèmes de stabilité.

La deuxième option consiste à augmenter la résolution temporelle par l'utilisation de filtres interpolateurs. Le coût de codage est plus faible puisqu'il vaut un bit par sous-trame. On parle ainsi de pitch fractionnaire car le retard n'est plus nécessairement entier mais possède une fraction de la période d'échantillonnage.

Nous vous proposons dans ce mémoire d'étudier les performances atteintes par l'utilisation d'un filtre prédicteur long-terme (fractionnaire) dans des codeurs/décodeurs de parole de type CELP.

Abstract

The technique of the long-term prediction consists in searching in the past of the signal, a sequence that is nearest of the sequence treated in the sense of a chosen criterion (here of quadratic form).
The value found M is the nearest integer (between 20 and 147) of the real value of the period of the pitch. The digitalization of the signal is the reason of this approximation. This digitalization can also generates a determination of a multiple of the period.
The human ear being very sensitive to variations of the pitch, a precise evaluation of this period turns out to be necessary. It offers to us two possibilities then; the first one consists of increasing the order of the predictor (2 or 3), the second of interpolating words of code candidates whose index are near of the value of the period.
The first option is not generally used a lot for two reasons; first because the increase of the order of the predictor increases the bit-rate transforming the coding less efficient, and secondly because these predictors create problems of stability.
The second option consists in increasing the temporal resolution by the utilization of interpolators filters. The cost of coding is weaker since it is requires only one additional bit by sub-frame. This is known as fractional pitch because the delay is not anymore necessarily integer but possesses a fraction of the period of sampling.

This memory deals with studying the performances reached by the utilization of a fractional filter long-term predictor in encoders/decoders of speech of CELP type.

Table des Matières

Liste des Figures

Liste des Tableaux

- Les line Spectrum pairs ϕ_1..... ϕ_p vérifient la relation :

$$0 < \phi_1 < .....<\phi_p < \pi. \tag{2.11}$$

Connaissant les a_k, on en déduit les coefficients ϕ_i. Réciproquement, connaissant les coefficients ϕ_i on peut déduire les coefficients a_k puisque :

$$A(z) = \frac{B_1(z) + B_2(z)}{2}. \tag{2.12}$$

La quantification des coefficients ϕ_i doit conserver la relation (2.11) pour maintenir la stabilité du filtre de synthèse. On peut aussi utiliser les coefficients de corrélation partielle $k_1,\ldots,k_p$ (PARCOR) calculés en utilisant l'algorithme de Levinson ou de Schur. Ces coefficients possèdent la propriété d'être toujours compris entre –1 et +1. Les coefficients k_i subissent au préalable une transformation non linéaire avant d'être quantifiés.

$$K_i = \log \frac{1 + k_i}{1 - k_i}. \tag{2.13}$$

Ces nouveaux coefficients sont appelés les Log Area Ratio.

Mais la représentation LSP reste la plus utilisée et la meilleure pour le codage des coefficients a_k. [1] [8].

2.2.2 Prédiction Long-terme

Lorsque le signal de parole est voisé, il existe une forte corrélation qui est due à la périodicité du signal. Cette redondance est exploitée par l'utilisation d'un prédicteur pitch ou prédicteur long-terme. Dans ce contexte un filtre long-terme d'ordre 1 peut être employé :

$$P(z) = \beta z^{-D}. \tag{2.14}$$

β et D sont respectivement, le coefficient prédicteur (gain) et la période du pitch.

Le signal erreur est exprimé par :

$$e(n) = r(n) - \beta\, r(n - D), \tag{2.15}$$

[1] Le problème de la quantification est traité dans le prochain chapitre.

2.4.1 Sélection des paramètres de synthèse

Les paramètres du filtre de synthèse formant sont les premiers à être déterminés par une analyse du signal de parole original. Ce filtre est spécifié par ces p coefficients qui sont réactualisés chaque trame.

Les paramètres restants sont l'index i du vecteur optimal et son gain G, la période pitch D et son coefficient β. La procédure d'optimisation est basée sur le critère des moindre carrés pondérés comme le montre la figure 2.7.

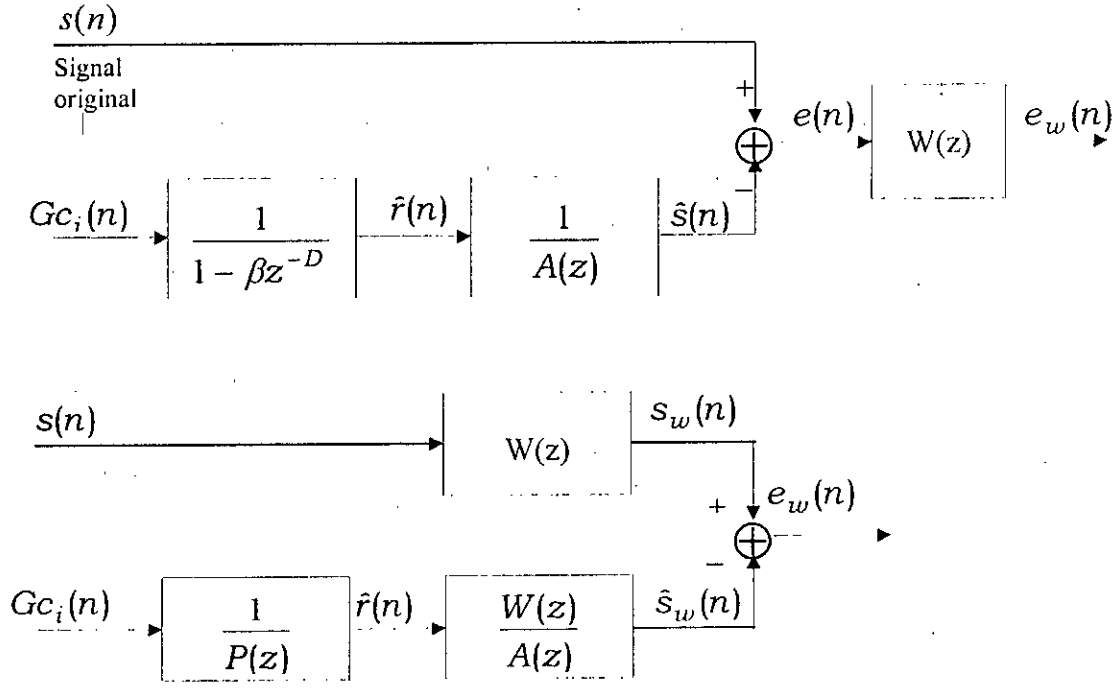


Figure 2.7 Modification de la structure du codeur CELP

La figure 2.7 illustre une autre structure CELP dont la modification est la distribution du filtre de pondération W(z) dans chacune des deux branches du codeur. Ainsi dans la première branche de la figure 2.7, on effectue d'abord une convolution du signal de parole original avec la réponse impulsionnelle du filtre A(z). Le signal résiduel engendré est ensuite filtré par le filtre de fonction de transfert 1/A(z/γ).

Dans la seconde branche, la mise en cascade permet le remplacement des deux filtres par un seul filtre de fonction de transfert 1/A(z/γ). Cette modification permet de réduire les calculs.

L'erreur calculée au sens des moindres carrés est comme suit :

$$\varepsilon = \sum_{n=0}^{N-1} (s_w(n) - \hat{s}_w(n))^2 \quad . \tag{2.34}$$

où N est la taille de la sous-trame.

Il existe différentes méthodes pour la détermination des paramètres selon la complexité et les performances. On peut citer quelques-unes :

- On cherche les valeurs optimales *i*, *G*, β et *D*. dans cette méthode *G* et β sont déterminés pour toutes les combinaisons de paires (*i* , *D*), les valeurs choisies sont celles qui minimisent l'erreur quadratique moyenne pondérée.
- Une autre méthode consiste à utiliser une approche séquentielle qui suggère que la période pitch soit optimisée pour un signal d'excitation nul, c'est à dire $G = 0$. Le retard *D* est déterminé et est gardé constant. Les paramètres restants sont déterminés de deux manières :

1. Le gain optimum et le coefficient pitch sont déterminés pour chaque valeur index *i*.
2. Le coefficient pitch est aussi déterminé pour une excitation nulle. Gardant les paramètres du filtre fixés, le gain optimum est déterminé pour chaque index de l'excitation.

En conclusion, le choix de l'approche séquentielle pour la détermination des paramètres du filtre pitch est très attractif (moins de complexité). Avec cette approche le signal d'excitation utilisé pour exciter le filtre de synthèse formant, consiste en deux composantes dont la première est une version retardée et pondérée du signal d'excitation passée, la seconde celle donnée par le dictionnaire d'excitation.

2.4.2 Dictionnaire d'excitation

Outre le dictionnaire aléatoire gaussien (dictionnaire stochastique) proposé initialement pour le codeur CELP, le contenu du dictionnaire peut être adapté à un algorithme particulier pour arriver à une méthode de recherche efficace. Dans la tâche de réduire la charge excessive des calculs lors de la recherche du vecteur d'excitation optimum, plusieurs structures du dictionnaire d'excitation ont été proposées. Le but est d'atteindre l'un des objectifs suivants :

- Procédure de recherche rapide
- Réduire l'espace de stockage
- Augmenter la qualité du signal de parole reconstruit.

Divers types de dictionnaire ont été définis :

- Des dictionnaires lacunaires.
- Des dictionnaires d'impulsions régulièrement espacées.
- Des dictionnaires construits avec des séquences mutipulsionnelles.
- Des dictionnaires binaires où les échantillons prennent des valeurs –1 et +1.
- Des dictionnaires ternaires où les échantillons prennent des valeurs –1, 0 et +1.
- etc…

2.5 Conclusion

Le but de codage est de réduire le nombre d'informations à envoyer chaque seconde tout en gardant une qualité adéquate aux besoins. Dans le codage CELP, le couple "source - conduit vocal" se modélise en codage par le couple "signal d'excitation - filtre LPC" ; le principe de masquage auditif a permis de définir un filtre perceptuel de mise en forme de bruit.

La majorité des codeurs de parole récents sont de type CELP. Ce dernier a beaucoup évolué depuis le premier modèle de Schroeder et d'Atal entraînant des modifications de structures et de dictionnaires afin de réduire la complexité et le débit.

Chapitre 3

La Quantification

3.1 Introduction

L'objet de ce chapitre est de rappeler les résultats fondamentaux de la quantification scalaire et vectorielle, nous verrons ainsi la supériorité théorique du cas vectoriel par rapport au cas scalaire. Lorsque le débit devient faible, inférieur ou égal à 2 bits par échantillon, il est nécessaire de regrouper des échantillons avant l'opération du codage pour former un vecteur. Le quantificateur scalaire traite chaque échantillon indépendamment des précédents. Le quantificateur vectoriel prend en compte directement la corrélation qui peut exister entre les échantillons successifs du signal. C'est la propriété fondamentale de ce quantificateur.

3.2 Quantification scalaire

3.2.1 Définitions et principes

La quantification consiste en l'approximation d'un signal d'amplitude continue par un signal d'amplitude discrète.

Considérons un signal à temps continu s(t) et à bande limitée [-B, +B]. La fréquence de Nyquist f_c = 2B définit la fréquence d'échantillonnage la plus basse n'entraînant aucune perte d'information. Cette procédure d'échantillonnage fournit la suite des échantillons s(n).

Il faut maintenant réaliser une discrétisation des amplitudes pour obtenir une représentation numérique du signal. On parle alors de quantification.

La quantification scalaire (QS) assigne à une valeur d'entrée x sa valeur approximée d'un ensemble fini prédéterminé ou dictionnaire de L valeurs de sorties acceptables $C=\{ y_k / k = 1, \cdots L \}$.

On définit L points scalaires $Q : R \rightarrow C$, $C = \{ y_1, y_2, \ldots y_L \} \subset R$, où L est la taille du dictionnaire ou codebook. Les valeurs de sorties y_i sont appelées niveaux de sorties, valeurs de sorties ou encore valeurs de reproduction. Dans tout les cas L est fini et on admet que :

$$y_1 < y_2 < \cdots < y_L .$$

Le quantificateur partitionne la sortie en L intervalles [1] ou classes R_i, la ième classe est donnée par :

$$R_i = \langle x \in R : Q(x) = y_i \rangle \equiv Q^{-1}(y_i) \qquad (3.1)$$

Suivant cette définition $\bigcup_i R_i = C$ et $R_i \cap R_j = \phi$ pour i≠j. les classes ou les cellules limites sont appelées cellules granulaires, l'ensemble de ces cellules est appelé région granulaire.

Un quantificateur régulier a la caractéristique que si deux valeurs entrées, a et b avec a < b, sont quantifiées avec la même valeur sortie y, alors n'importe quelle autre entrée comprise entre a et b sera quantifiée elle aussi avec la même sortie y.

Lorsque le nombre de valeurs quantifiées est une puissance de 2 soit $L = 2^b$, chaque valeur quantifiée peut être représenter par un mot de b bits.

La loi de quantification Q(x) peut prendre deux formes :

- L pair → dans ce cas $x = 0$ est un niveau de décision (midrise)
- L impair → y = 0 est une valeur de sortie (midtread)

Chaque quantificateur peut être considéré comme l'effet de deux opérations successives : un encodeur ξ et un décodeur D.

[1] Les intervalles de la partition sont aussi appelés, les régions de voronoï

L'encodeur est donné par ξ :R $\rightarrow$ I, ou I={ 1,2,...L } et le décodeur est donné par D : I $\rightarrow$ C. Ainsi si Q(x) = y_i, alors ξ(x) = i et D(i) = y_i. Avec ces conditions on a Q(x) = D(ξ(x)).

Dans les systèmes de communication comme dans le codage de parole, le codeur transmet l'index i du niveau sélectionné y_i et non pas la valeur y_i. Au niveau du décodeur qui possède un dictionnaire (codebook) des valeurs quantifiées ,on récupère la valeur y_i qui correspond à l'index reçu [10].

3.2.2 Mesure de performance d'un quantificateur

On définit une mesure de distorsion entre deux nombres s et $\hat{s}$ que l'on note $d(s,\hat{s})$ avec :

$$d(s,\hat{s}) = |s-\hat{s}|^2 \tag{3.2}$$

puis la distorsion moyenne

$$D = E\{d(s,\hat{s})\} = \int_{-\infty}^{+\infty} d(x,\hat{s}) p_s(x) dx \tag{3.3}$$

(où E est l'espérance mathématique).

Pour un quantificateur Q ={ y_i, R_i, i=1,2... L} et une variable aléatoire d'entrée s, la distorsion moyenne a pour expression :

$$D = \sum_{i=1}^{L} \int_{x \in R_i} d(x,\hat{s}_i) p_s(x) dx \tag{3.4}$$

où R_i spécifie partition associé au numéro i et p_s(x) est la densité de probabilité du processus s.

3.2.3 Quantificateur scalaire uniforme

Considérons un signal à temps discret s(n) prenant des valeurs uniformément distribuées dans l'intervalle [-A,+A].

La démarche la plus naturelle pour définir un quantificateur consiste à :

1. Partitionner l'intervalle [-A,+A] en L =2^b intervalles { $R_1 \dots R_L$ } (intervalles de décision) distincts de même longueur Δ =2A/2^b.
2. Numéroter chaque intervalle.
3. Définir un représentant par intervalle, on aura donc L valeurs de sorties $y_1, y_2, \dots, y_L$, qui sont les centres de masse de chacun des intervalles.

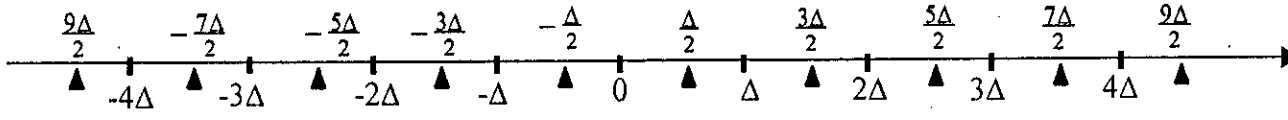


Figure 3.1 Exemple d'un quantificateur scalaire uniforme.

Le débit de ce quantificateur est donc donné par l'expression suivante :

$$R = \log_2 L \tag{3.5}$$

Il apparaît deux sortes d'erreurs ou bruits de quantification :

- Le bruit granulaire qui se produit lorsque la valeur d'entrée s se situe dans l'une des cellules $[s_i, s_{i+1}]$, l'erreur résultante qui est la différence entre s et $Q(s)$ peut être majorée par un demi-pas de quantification.
- Le bruit de surcharge ou de dépassement qui se produit lorsque la valeur d'entrée se situe hors de l'intervalle $[s_0, s_L]$. La valeur de reproduction est alors soit y_1 ou y_L, et l'erreur résultante est supérieure à un demi-pas de quantification.

La procédure d'encodage consiste à décider à quel intervalle appartient s(n) puis à lui associé le numéro *i(n)* ∈ [1,...,L=2^b] correspondant qui sera transmis.

La procédure de décodage consiste à associer au numéro i(n) le représentant correspondant choisi parmi l'ensemble des représentants $[y_1 \dots y_L]$.

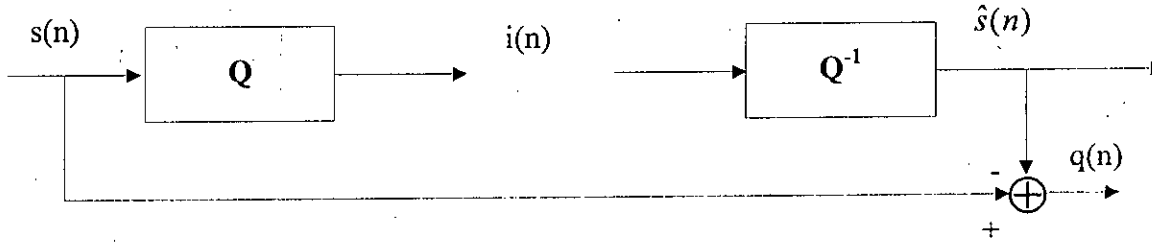


Figure 3.2 Schéma simple illustrant l'encodage et le décodage

Les procédures d'encodage et de décodage de ce quantificateur sont schématisées par la figure 3.2. L'erreur de quantification a pour expression :

$$q(n) = s(n) - \hat{s}(n) \tag{3.6}$$

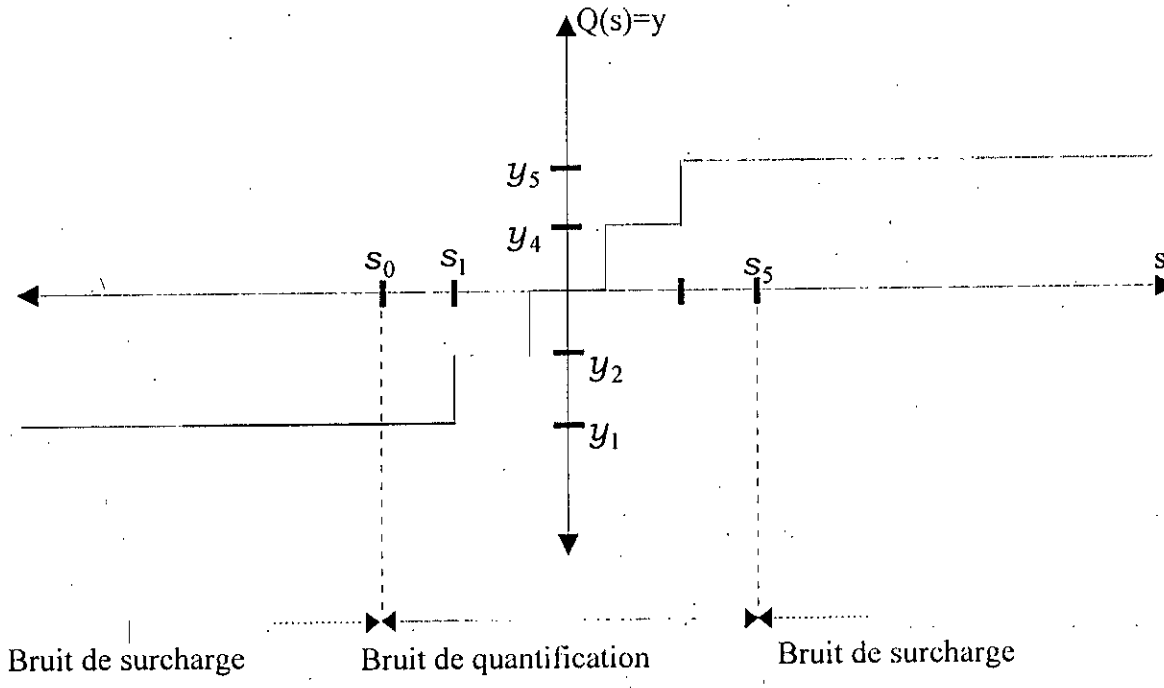


Figure 3.3 Exemple d'un QS uniforme L=5

Pour caractériser la dégradation apportée par l'opération de quantification, il faut définir un critère et proposer un modèle simple pour les signaux intervenant dans ce critère.

Supposons que s(n) soit la réalisation d'un processus aléatoire S(n). On choisit habituellement de minimiser l'erreur quadratique moyenne donnée par l'équation (3.3) .

On peut également chercher à maximiser le rapport signal sur bruit (*RSB*) :

$$RSB = \frac{E\{s^2(n)\}}{E\{(s(n) - \hat{s}(n))^2\}} \tag{3.7}$$

L'erreur de quantification q(n) peut être considéré comme un processus aléatoire uniformément distribuée dans l'intervalle $[-\Delta/2, +\Delta/2]$ [2].

La densité de probabilité de l'erreur est donnée par :

$$P_Q(x) = \begin{cases} 1/\Delta & |x| \le \Delta/2 \\ 0 & |x| > \Delta/2 \end{cases} \tag{3.8}$$

sa variance est donnée par $\sigma_Q^2 = E\{Q(n)^2\} = \int_{-\Delta/2}^{+\Delta/2} x^2 \frac{1}{\Delta} dx = \frac{\Delta^2}{12} = \frac{1}{12}\left(\frac{2A}{2^b}\right)^2$

$$\sigma_Q^2 = \frac{A^2}{3} 2^{-2b}. \tag{3.9}$$

La moyenne de l'erreur est nulle. Pour le signal S(n) uniformément distribué dans l'intervalle [-A,+A], sa moyenne est nulle et sa variance a pour expression

$$\sigma_s^2 = E\{S^2(n)\} = \int_{-A}^{+A} x^2 \frac{1}{2A} dx = \frac{A^2}{3} \tag{3.10}$$

Les processus aléatoires ici sont stationnaires (au deuxième ordre), ergodiques, centrés et la variance de la variable aléatoire S(n) est égale à la puissance du signal.

On obtient la relation fondamentale qui donne la puissance de l'erreur de quantification en fonction de la puissance du signal et de la résolution b

$$\sigma_Q^2 = \sigma_s^2 2^{-2b} \tag{3.11}$$

Le rapport signal sur bruit a pour expression :

$$10 \log_{10}\left(\frac{\sigma_s^2}{\sigma_Q^2}\right) = 10 \log_{10}(2^{2b}) = 6.02b \tag{3.12}$$

[2] Relation valable pour les débits élevés

Le fait de rajouter un bit revient donc à augmenter le rapport signal sur bruit de 6dB. Pour les signaux de parole à bande étroite (bande téléphonique), la quantification est faite sur 8bits (avec la loi A ou μ) ce qui est équivalent à une quantification linéaire sur 12 bits [8].

3.2.4 Quantificateur scalaire non-uniforme

Le quantificateur scalaire uniforme n'est pas le quantificateur optimal pour une source non uniforme. Le quantificateur optimal sera déterminé en fonction de la densité de probabilité de cette source.

3.2.4.1 Conditions nécessaires d'optimalité

Pour définir un quantificateur il s'agit de trouver la partition $\{R_1 \dots R_L\}$ et les représentants $\{\hat{s}_1 \dots \hat{s}_L\}$qui minimisent la distorsion D. Cette minimisation conjointe n'admet pas de solution simple. Il n'existe que deux conditions nécessaires d'optimalité. Si l'on connaît les représentants $\{\hat{s}_1 \dots \hat{s}_L\}$, on peut calculer la meilleure partition $\{R_1 \dots R_L\}$. Si l'on se donne la partition, on peut en déduire les meilleurs représentants. La partie encodage du quantificateur doit être optimale étant donné la partie décodage et réciproquement.

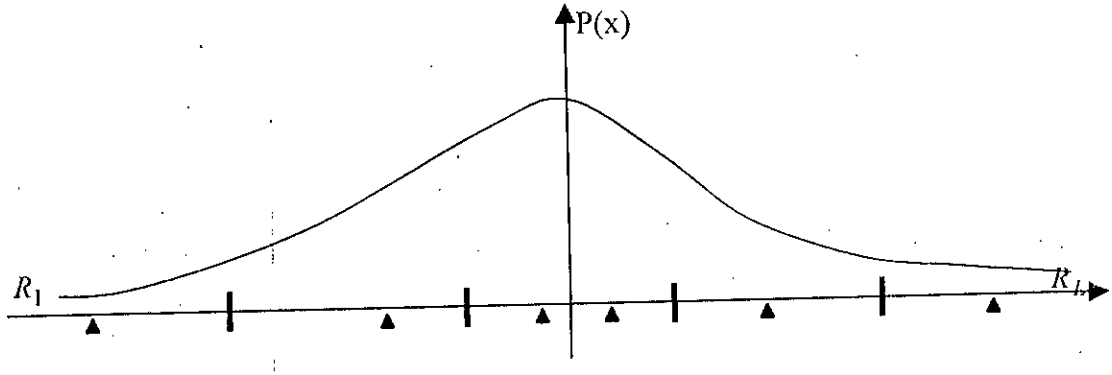


Figure 3.4 Quantificateur scalaire non uniforme

Une démonstration sur les deux conditions nécessaires d'optimalité, est donnée par [8].

1. Etant donné un dictionnaire $\{\hat{s}_1 \dots \hat{s}_L\}$, la meilleure partition est celle qui vérifie

et il est appelé signal résiduel pitch. La transformée en Z des deux cotés de l'équation (2.15) permet d'obtenir l'expression du filtre d'analyse pitch :

$$P_a(z) = 1 - \beta z^{-D}, \tag{2.16}$$

Dans le domaine temporel, le filtre d'analyse $P_a(z)$ soustrait de l'échantillon courant de parole l'échantillon distant d'un retard égal à la période estimée (pondérée par β). Dans le domaine fréquentiel, ce filtre enlève la structure harmonique du signal d'entrée (dans notre cas le résiduel). L'analyse pitch n'aura pas un effet utile au niveau du signal non voisé puisque son excitation est aléatoire (pas de structure harmonique). Le coefficient predicteur β correspond au degré de périodicité de la forme d'onde et prend les valeurs $0 \leq \beta \leq 1$.

Ainsi β est proche de 0 pour une structure non périodique (dans ce cas la valeur de D est sans signification) et il est pratiquement égal à l'unité pour l'état stable de la parole voisée.

Au décodeur le filtre de synthèse du pitch $P_s(z) = 1 / P_a(z)$ est utilisé pour introduire une structure harmonique du signal de parole synthétisé.

Il existe plusieurs méthodes pour la détection du pitch et la localisation des zones voisées, on peut en citer quelques-unes :

- Méthode d'autocorrélation.
- Méthode basée sur une forme simplifiée de l'autocorrection AMDF (Average Magnitude Difference Function).
- Méthode de Gold Rabiner (Gold Rabiner Pitch Tracker).
- Méthode du cepstre ou méthode homomorphique
- Méthode mettant en œuvre le filtre inverse SIFT (Simplified Inverse Filter-Tracking algorithm).

Ces méthodes peuvent être consultées dans [9].

2.2.3 Estimation des paramètres des prédicteurs

On peut faire une formulation générale pour la détermination des coefficients (paramètres) des prédicteurs court-terme et long-terme [1], [7].

$$R_i = \left\{ s : (s - \hat{s}_i)^2 \leq (s - \hat{s}_j)^2 \quad \forall j \in \{1 \ldots L\} \right\} \tag{3.13}$$

C'est la règle dite du plus proche voisin.

2. Etant donné une partition $\{R_1 \ldots R_L\}$, les meilleurs représentants sont obtenus par la condition dite du centroïde (ou centre de gravité) de la partie de la densité de probabilité placée dans la région R_i

$$\hat{s}_i = \frac{\int_{x \in R_i} x P_s(x) dx}{\int_{x \in R_i} P_s(x) dx} = E\{S / S \in R_i\}. \tag{3.14}$$

La valeur que l'on doit choisir est la valeur moyenne de S dans l'intervalle considéré. Ce résultat a une interprétation en mécanique : le moment d'inertie d'un objet par rapport à un point est minimum lorsque ce point est le centre de gravité.

On remarque que la connaissance explicite de la partition n'est pas nécessaire. Cette partition est entièrement déterminée par la connaissance de la mesure de distorsion, par l'application de la règle du plus proche voisin et par l'ensemble des représentants. Le schéma de l'encodeur et du décodeur est donné par la figure 3.5

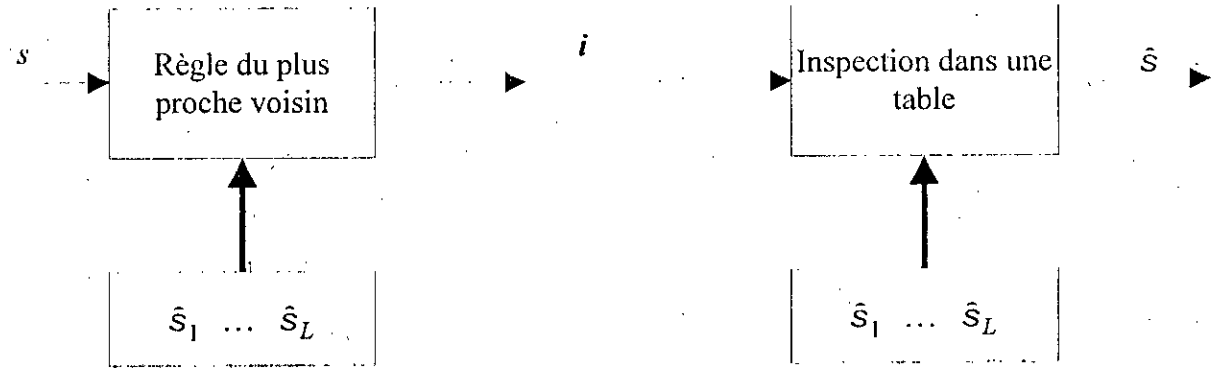


Figure 3.5 Encodeur et décodeur

3.2.4.2 Algorithme de Lloyd-Max

Dans la pratique, on ne connaît pas la densité de probabilité $P_s(x)$. Pour construire un quantificateur, on utilise des données empiriques (une base d'apprentissage), en associant à chaque valeur le même

poids. La base d'apprentissage doit être composée d'un grand nombre d'échantillons représentatifs de la source.

On donne ici une description sommaire de l'algorithme de Lloyd-Max permettant de construire un quantificateur optimal. C'est un algorithme itératif vérifiant successivement les deux conditions d'optimalité

1. On initialise le dictionnaire $\{\hat{s}_1 \dots \hat{s}_L\}$, par exemple par tirage aléatoire.
2. Connaissant ce dictionnaire, on étiquette chaque échantillon de la base d'apprentissage, par le numéro de son plus proche voisin. On détermine la partition optimale $\{R_1 \dots R_L\}$.
3. A partir de tous les échantillons étiquetés par le même numéro, on en déduit un nouveau représentant par un calcul de moyenne.
4. On calcul la distorsion moyenne associée à cette base d'apprentissage et on arrête cet algorithme si là distorsion ne décroît presque plus, c'est-à-dire si la décroissance devient inférieure à un seuil sinon on reprend les deux étapes précédentes.

3.2.4.3 Hypothèse dite haute résolution

Cette méthode consiste à admettre que le nombre de niveaux de quantification est très élevé (i.e. la partie de la densité de probabilité du processus contenue dans chaque élément de la partition de voronoï est approximativement constante), il est possible d'obtenir explicitement l'expression de la puissance de l'erreur de quantification uniquement en fonction de $p_S(x)$ avec :

$$\sigma_Q^2 = \frac{1}{12}\left(\int_{-\infty}^{+\infty}(p_S(x))^{1/3}dx\right)^3 2^{-2b} \tag{3.15}$$

Pour une source stationnaire gaussienne centrée de variance σ_S^2 nous obtenons :

$$\sigma_Q^2 = c_1\sigma_S^2 2^{-2b} \qquad \text{avec} \qquad c_1 = \frac{\sqrt{3}}{2}\pi \tag{3.16}$$

Ces équations vont nous servir de références pour les comparaisons entre quantificateurs. Ajoutons aussi que des études [10] ont montré qu'un QS non-uniforme est équivalent à un QS uniforme précédé d'une transformation non linéaire.

3.2.5 Quantification scalaire prédictive

Le quantificateur scalaire prédictif cherche à décorréler le signal avant de le quantifier. Considérons la procédure suivante qui consiste à :

- Retrancher à s(n) une valeur quelconque v(n)
- Réaliser l'encodage et le décodage
- Ajouter v(n) à la valeur décodée

$$s(n) - \hat{s}(n) = [s(n) - v(n)] - [\hat{s}(n) - v(n)] = r(n) - \hat{r}(n). \tag{3.17}$$

On remarque que les deux distorsions $s(n) - \hat{s}(n)$ et $r(n) - \hat{r}(n)$ sont identiques.

En principe il faut distinguer l'erreur de quantification $q(n) = r(n) - \hat{r}(n)$ de l'erreur de reconstruction $\bar{q}(n) = s(n) - \hat{s}(n)$. Ici, les deux erreurs sont égales à chaque instant n ($q(n) = \bar{q}(n)$). On peut prendre v(n) comme une combinaison linéaire des échantillons passées.

$$v(n) = -\sum_{i=1}^{p} a_i s(n-i) \tag{3.18}$$

$$r(n) = s(n) - v(n) = s(n) + \sum_{i=1}^{p} a_i s(n-i). \tag{3.19}$$

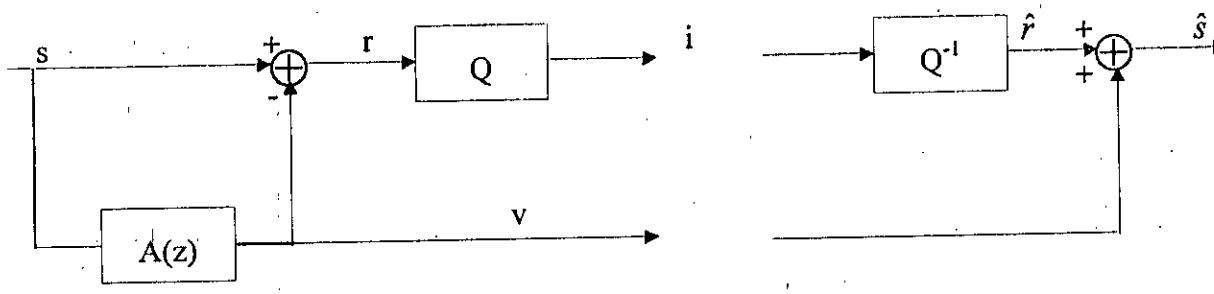


Figure 3.6 Quantificateur prédictif

Le filtre qui permet d'obtenir r(n) à partir de s(n) s'appelle filtre d'analyse A(z). La minimisation de la puissance de l'erreur de prédiction nous permet de déterminer les coefficients a_i.

$$\sigma_R^2 = E\left\{\left[s(n) + \sum_{i=1}^{p} a_i s(n-i)\right]^2\right\} \tag{3.20}$$

Le développement théorique associé à la prédiction linéaire aboutit aux équations de Yule-walker qu'on a déjà vu à la section 2.2.3.

$$\Gamma_S(p)\left(a_1 \dots a_p\right)^T = -\left(\rho_1 \dots \rho_p\right), \tag{3.21}$$

où les ρ_i sont les coefficients normalisée de la fonction d'autocorrelation :

$$\rho_i = \frac{E\{\, s(n)\, s(n-i)\}}{E\left\{s^2(n)\right\}} \tag{3.22}$$

et $\Gamma_S(p)$ est la matrice d'autocorrelation normalisée avec

$$\Gamma_S(p) = \begin{bmatrix} 1 & \rho_1 & \cdots & \cdot & \rho_{p-1} \\ \rho_1 & \cdot & \cdots & \cdot & \cdot \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \cdot & \cdot & \cdots & \cdot & \rho_1 \\ \rho_{p-1} & \cdot & \cdots & \rho_1 & 1 \end{bmatrix} \tag{3.23}$$

Il existe plusieurs algorithmes pour la résolution de ce système d'équation comme l'algorithme de Levinson ou l'algorithme de Schur .

Cherchons à déterminer la valeur des paramètres a_i minimisant l'erreur quadratique moyenne ou maximisant le rapport signal sur bruit. Celui-ci s'écrit :

$$\frac{E\left\{s^2(n)\right\}}{E\left\{\left[s(n)-\hat{s}(n)\right]^2\right\}} = \frac{E\left\{s^2(n)\right\}}{E\left\{r^2(n)\right\}} \frac{E\left\{r^2(n)\right\}}{E\left\{\left[r(n)-\hat{r}(n)\right]^2\right\}} \tag{3.24}$$

en appelant $\hat{r}(n)$ le représentant sélectionné à l'instant n pour quantifier l'erreur de prédiction. Si l'on réalise une quantification optimale sur $r(n)$ avec une résolution de b bits par échantillon et si l'on applique la formule (3.16), on obtient :

$$\frac{E\left\{S^2(n)\right\}}{E\left\{\left[S(n)-\hat{s}(n)\right]^2\right\}} = \frac{\sigma_S^2}{\sigma_R^2} \frac{2^{2b}}{c_1} \tag{3.25}$$

Le rapport des deux puissances s'appelle le gain de prédiction. On le note :

$$G(p) = \frac{\sigma_S^2}{\sigma_R^2}, \tag{3.26}$$

ce gain est une fonction croissante de l'ordre de prédiction p et tend vers une limite $G_p(\infty)$. On a

$$\sigma_Q^2 = \sigma_{\overline{Q}}^2 = c_1 \frac{\sigma_S^2}{G_p(p)} 2^{-2b}. \tag{3.27}$$

Le gain de prédiction mesure l'amélioration des performances apportée par la quantification de l'erreur de prédiction plutôt que la quantification directe du signal (si le processus est blanc $G(p) \to 1$, et si le processus est totalement prédictible $G_p(\infty) \to +\infty$), sa valeur asymptotique s'exprime en fonction du déterminant de Γ_S avec :

$$G_p(\infty) = \lim_{p \to +\infty} \frac{1}{\left(\det \Gamma_S(p)\right)^{1/p}}. \tag{3.28}$$

La quantification scalaire prédictive peut se faire soit en boucle ouverte ou en boucle fermée

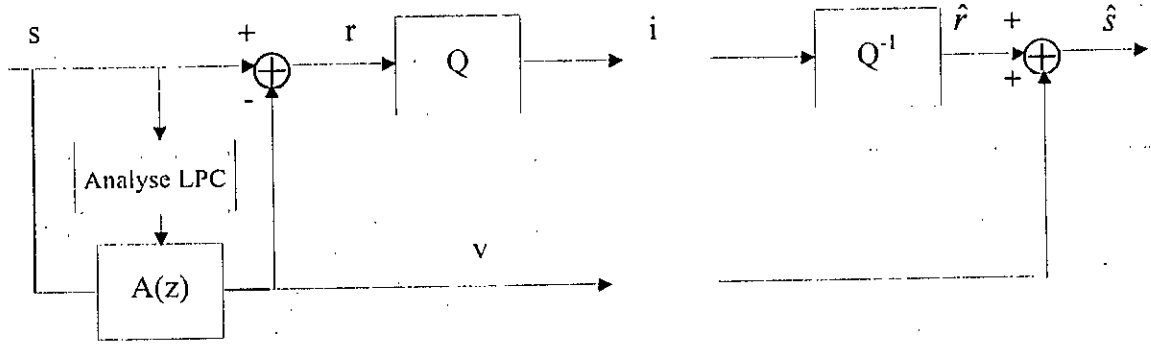


Figure 3.7 Quantificateur prédictif en boucle ouverte

Examinons la figure 3.7, sous cette forme le quantificateur exige la transmission à chaque instant non seulement du numéro i(n), résultat de la quantification de l'erreur de prédiction r(n), mais aussi d'un autre numéro qui serait associé à la quantification de la prédiction v(n) elle même. Ce schéma représente une quantification en boucle ouverte. On peut aussi réaliser une prédiction en boucle fermée comme dans la figure 3.8. Dans ce cas, il n'est plus nécessaire de transmettre v(n) au

récepteur puisque il represente maintenant la prédiction du signal reconstruit $\hat{s}(n)$. Cette prédiction peut être réalisée de façon identique au récepteur.

Les coefficients du filtre A(z) sont actualisés tous les N échantillons.

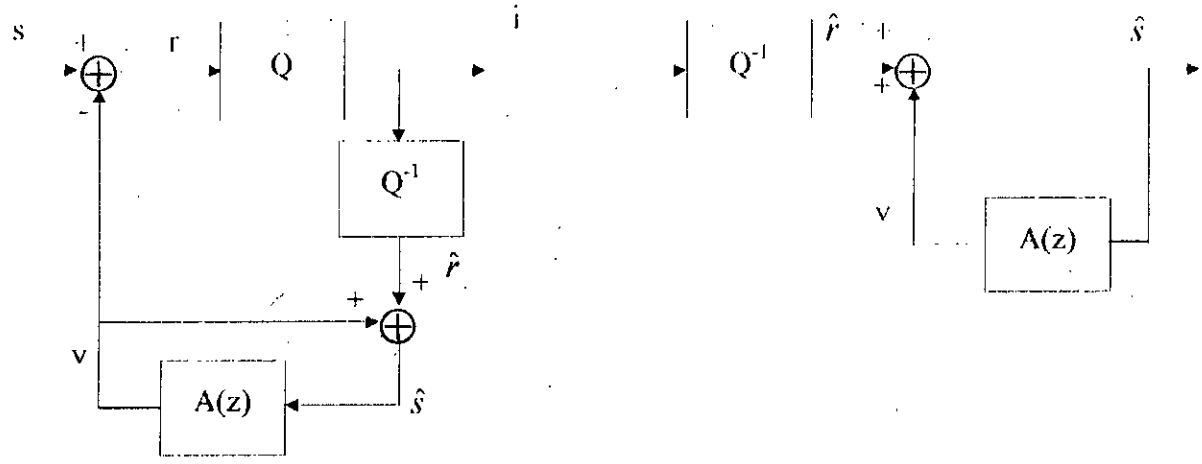


Figure 3.8 Quantification prédictif en boucle fermée

3.3 Quantification vectorielle

Lorsque le débit visé pour coder un signal de parole échantillonné à 8kHz est égal à 16, à 8 ou à 4 kbit/s, le nombre de bits disponibles par échantillon est respectivement égal à 2, 1 et ½. Pour coder un signal de parole à ces débits, il devient nécessaire de regrouper plusieurs échantillons et de chercher à quantifier l'ensemble. On parle alors de quantification vectorielle.

La quantification vectorielle n'est pas qu'une simple généralisation du cas scalaire. Elle permet de prendre en compte directement la corrélation contenue dans le signal.

La quantification vectorielle (notée **QV**) consiste alors à représenter tout vecteur x de dimension N par un autre vecteur y_i de même dimension mais ce dernier appartenant à un ensemble fini C de L vecteurs (dictionnaire).

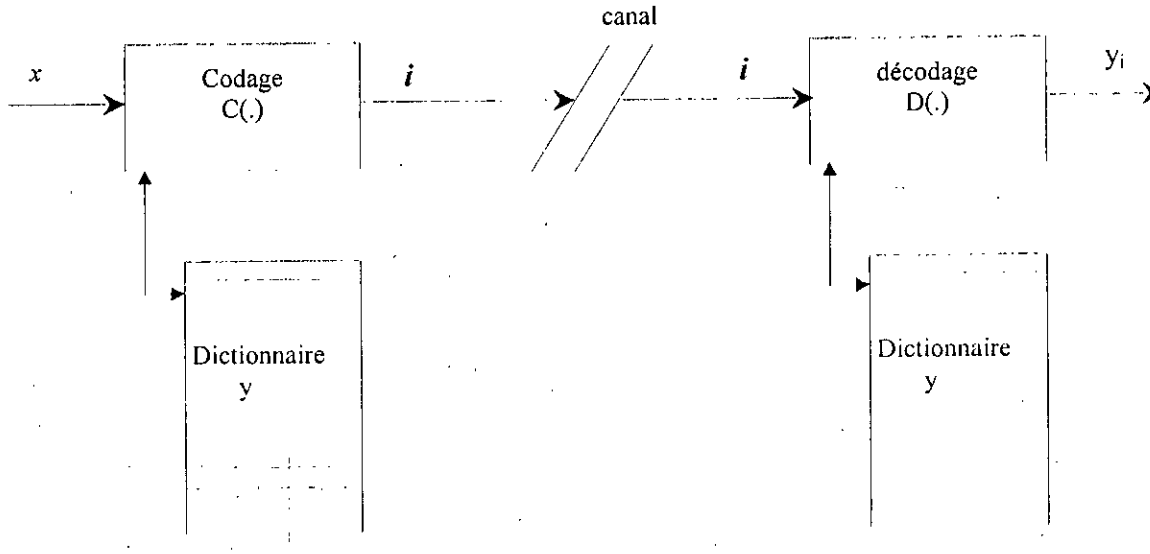


Figure 3.9 Schéma d'un quantificateur vectoriel

3.3.1 Formalisation

On appelle quantificateur vectoriel de dimension N et de taille L une application de R^N dans un ensemble fini C contenant L vecteurs de dimension N [5], [8], [11], [12], [13].

$$Q : R^N \rightarrow C \quad \text{avec} \quad C = \{y_1 \ldots y_L\} \quad \text{où} \quad y_i \in R^N \tag{3.29}$$

L'espace R^N est partitionné en L régions ou cellules (ou encore région de voronoï) définies par

$$R_i = \{x : Q(x) = y_i\} \tag{3.30}$$

On appelle C un dictionnaire (codebook), assimilable à une matrice si nécessaire, et y_i un représentant, un vecteur de sortie ou un vecteur de reproduction.

Il faut définir une mesure de distorsion d(x,y). Les deux choix habituels sont la distance euclidienne.

$$d(x,y_i) = \frac{1}{N}\|x - y_i\|^2 = \frac{1}{N}(x - y_i)^T (x - y_i), \tag{3.31}$$

ou encore une distance pondérée

$$d(x, y_i) = \frac{1}{N}(x - y_i)^T W (x - y_i), \tag{3.32}$$

où W est une matrice définie positive permettant d'introduire un caractère perceptuel comme nous l'avons déjà vu au chapitre précèdent.

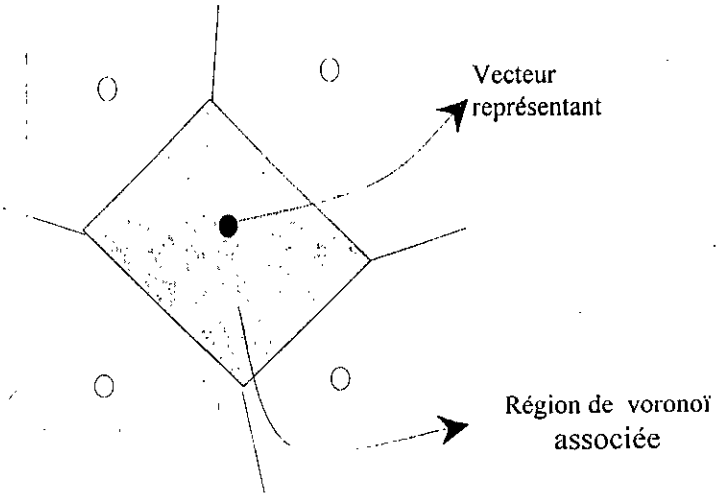


Figure 3.10 Principe de la quantification vectorielle

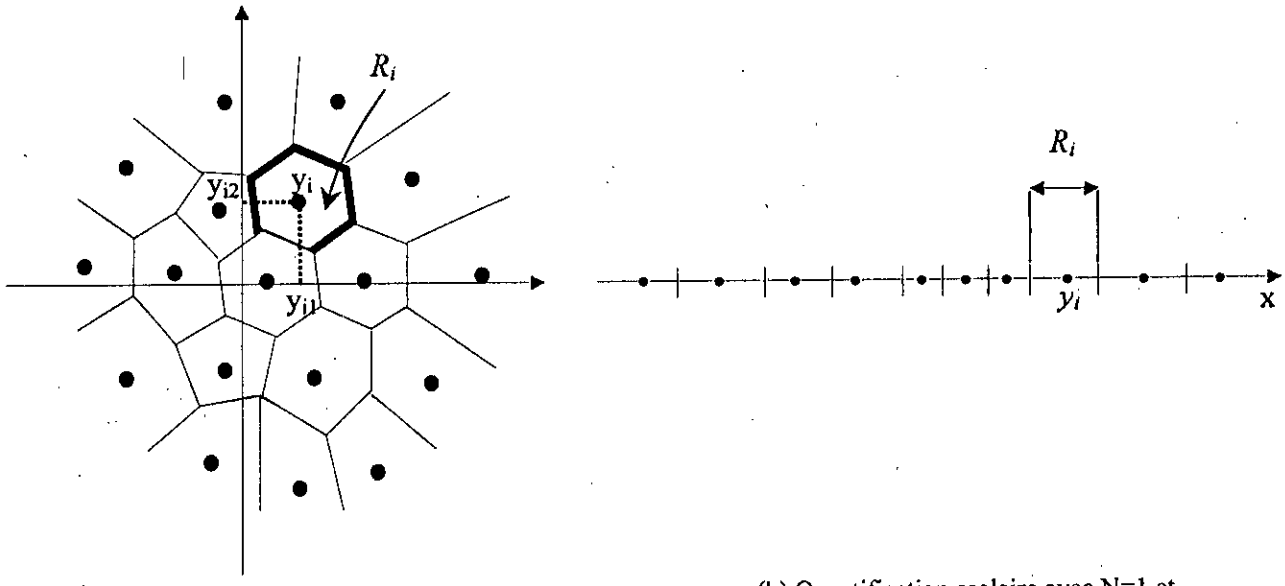
Soit x un vecteur aléatoire dans R^N, la distorsion moyenne a pour expression

$$\begin{aligned} D &= E[d(x,y)] \\ &= \sum_{i=1}^{L} p(x \in R_i) E[d(x,y_i) | x \in R_i] \\ &= \sum_{i=1}^{L} p(x \in R_i) \int_{x \in R_i} d(x,y_i) p(x) \cdot dx \end{aligned} \tag{3.33}$$

avec $p(x \in R_i)$ est la probabilité discrète que x soit dans R_i, et $p(x)$ est la fonction densité de probabilité multidimensionnelle (pdf) de x.

Lorsque l'on choisit, comme mesure de distorsion, la distance euclidienne, la distorsion moyenne devient l'erreur quadratique moyenne :

$$\sigma_Q^2 = \frac{1}{N} \int_{R^N} \|x - y\|^2 p(x) dx . \tag{3.34}$$



(a) Quantification vectorielle avec N=2 et L = 18

(b) Quantification scalaire avec N=1 et L = 10 représentants

Figure 3.11 Comparaison entre la quantification vectorielle et la quantification scalaire

La figure 3.11 section (a) montre un exemple d'une partition à deux dimensions (N=2) pour une quantification vectorielle. La région entourée avec des traits gras est la cellule R_i. N'importe quel vecteur d'entrée x se trouvant dans la cellule R_i est quantifié par y_i. Les positions des vecteurs de sortie ou de reproduction correspondants aux autres cellules sont donnés par les autres points noirs. Le total des vecteurs de sortie dans cet exemple est L=18.

Pour N=1, la quantification vectorielle devient une quantification scalaire. La figure 3.11 section (b) montre un exemple d'une partition pour une quantification scalaire. Les valeurs de sortie sont illustrées par les points. Ici aussi, n'importe quelle valeur entrée se trouvant dans l'intervalle R_i est quantifiée par y_i. Le nombre de valeurs de sortie est L=10.

Dans le cas de la quantification scalaire, bien que les cellules peuvent avoir différentes tailles, ils ont tous la même forme géométrique. En comparaison, on note que dans le cas vectoriel (figure 3.11 section (a)), les cellules à deux dimensions ont des formes géométriques différentes. Cette liberté d'avoir des formes de partitions variables dans un espace multidimensionnel donne à la

quantification vectorielle un avantage sur la quantification scalaire car la forme géométrique de la cellule intervient lors du " design " du quantificateur optimal .

3.3.2 Quantificateur optimal

Il s'agit de déterminer le dictionnaire C en choisissant N, L et les vecteurs de reproduction $\{y_1 \dots y_L\}$. On suppose, dans un premier temps, que N et L sont prédéterminés.

Comme dans le cas scalaire, il n'est pas possible de définir simultanément la meilleure partition et les meilleurs vecteurs de reproduction mais on garde les deux conditions d'optimalité qui s'expriment de façon identique.

1. Etant donné un dictionnaire $C = \{y_1 \dots y_L\}$, la meilleure partition est celle qui vérifie

$$R_i = \{x : d(x, y) \leq d(x, y_i) \quad \forall j \in \{1 \cdots L\}\} \tag{3.35}$$

c'est donc la règle du plus proche voisin.

2. Etant donné une partition, les meilleurs représentants sont obtenus par la condition du centroïde. Pour une distorsion quadratique.

$$y_i = cent\ (R_i) = E\{x / x \in R_i\} = \frac{\int_{R_i} x.p(x)dx}{\int_{R_i} p(x)dx}. \tag{3.36}$$

Pour définir le dictionnaire C, c'est-à-dire construire les vecteurs de reproduction, on est amené à utiliser une base d'apprentissage, comme dans le cas scalaire. Elle doit être composée d'un grand nombre M de vecteurs représentatifs de la source. Typiquement, on estime que chaque vecteur du dictionnaire doit être construit à partir d'une centaine de vecteurs de la base.

Le principe de l'algorithme de Lloyd-Max généralisé au cas vectoriel reste identique à celui du cas scalaire. Il faut :

1. initialiser le dictionnaire,
2. appliquer successivement la règle du plus proche voisin et la condition du centroïde,
3. itérer l'étape précédente tant que la décroissance de la distorsion moyenne reste importante.

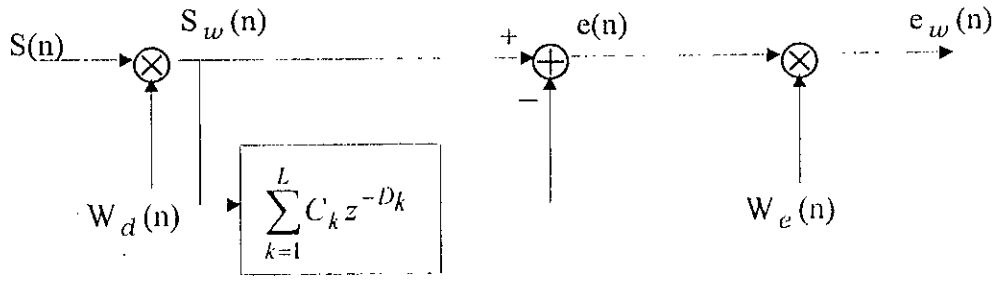


Figure 2.3 Modèle d'analyse pour les predicteur transversaux.

En se basant sur la figure 2.3 on peut écrire :

$$e_w(n) = W_e(n)\, e(n). \tag{2.17}$$

$$e_w(n) = W_e(n)\, S_w(n) - W_e(n) \sum_{k=1}^{L} C_k S_w(n - D_k) \tag{2.18}$$

où $S(n)$ est le signal d'entrée et $W_e(n)$, $W_d(n)$ sont les fenêtres de pondération. Les valeurs de D_k sont des entiers distincts arbitraires correspondant aux retards du signal d'entrée pondéré $S_w(n)$.

L'énergie de l'erreur ou l'erreur quadratique moyenne (EQM) est donnée par :

$$\varepsilon = \sum_{n=-\infty}^{+\infty} e_w^2(n). \tag{2.19}$$

Les coefficients C_k sont calculés par minimisation de ε. Ceci est obtenu en prenant la dérivée particlle de l'équation (2.19) par rapport à chaque coefficient C_k pour $k = 1, \ldots, L$, et en posant chacune des L équations résultantes à zéro selon le développement suivant :

$$e_w(n) = W_e(n)[S_w(n) - \sum_{k=1}^{L} C_k S_w(n - D_k)] \tag{2.20}$$

$$e_w(n) = W_e(n)[\sum_{k=0}^{L} C'_k S_w(n - D_k)] \tag{2.21}$$

avec $D_0 = 0$, $C'_0 = 1$ et $C'_k = -C_k$ pour $k \geq 1$, l'erreur quadratique moyenne devient donc :

L'initialisation du dictionnaire pose un problème comme dans le cas scalaire. Le choix du dictionnaire initial est capital car il conditionne les résultats finaux de l'algorithme. Plusieurs méthodes ont été proposées pour le déterminer :

- *une initialisation aléatoire* : le dictionnaire le plus simple est celui qui contient les L premiers vecteurs de la suite d'apprentissage où L vecteurs extraits aléatoirement de cette suite. Ces vecteurs peuvent bien sûr ne pas être représentatifs de la suite d'apprentissage, et conduire à des résultats très médiocres.
- *un algorithme à seuil* où au lieu de prendre L vecteurs aléatoirement, une distance minimale est fixée entre les éléments du dictionnaire initial. Cette méthode permet d'obtenir une meilleure représentativité que dans le cas précédent.
- *une méthode par dichotomie vectorielle* qui est référencée comme étant l'**algorithme LBG.** Elle combine à l'itération de Lloyd une technique dite de "Splitting" (le schéma de l'algorithme est présenté à la figure 3.12). Celle-ci consiste à découper chaque vecteur représentant y_i en 2 nouveaux vecteurs $y_i+\varepsilon$ et $y_i-\varepsilon$ (ε étant un vecteur de perturbation de faible énergie), avant d'appliquer au nouveau dictionnaire obtenu les itérations de Lloyd. L'algorithme génère ensuite une succession de dictionnaires (à chaque boucle le nombre de vecteurs de reproduction est multiplié par 2).

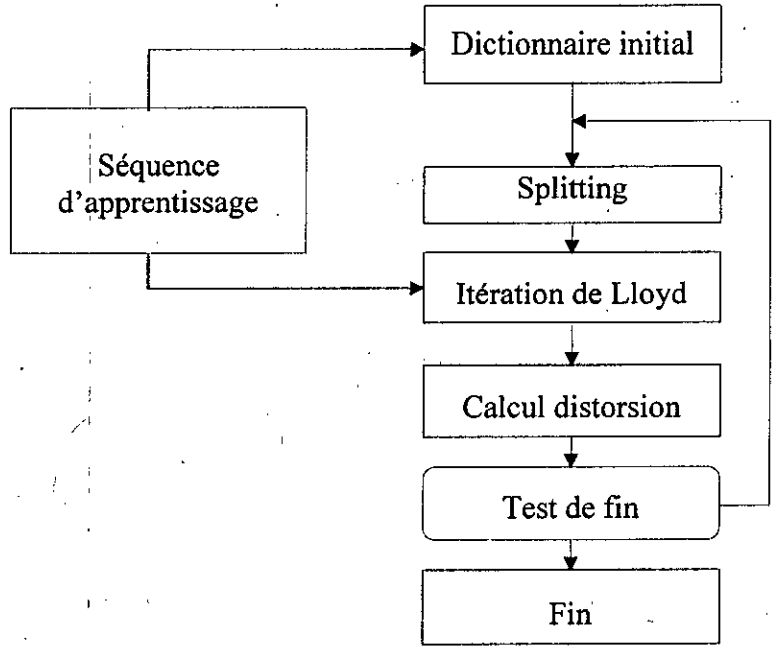


Figure 3.12 Schéma de fonctionnement de l'algorithme de LBG.

De nouveaux algorithmes basés sur des techniques de recuit simulé, par exemple, permettent d'améliorer les performances du quantificateur.

3.3.3 Modèle de quantification vectorielle

La quantification vectorielle utilise quatre propriétés interdépendantes de paramètres vectoriels et qui sont : la dépendance linéaire, la dépendance non-linéaire, la forme de la fonction densité de probabilité (pdf) et la dimension des vecteurs [5].

*1- **Dépendance** :* la compression de données est largement un processus de suppression de la redondance, il n'est pas nécessaire de gaspiller les bits pour la transmission de l'information redondante. La dépendance statistique se classe entre deux types : la dépendance linéaire et la dépendance non-linéaire.

Deux variables aléatoires qui sont corrélées sont linéairement dépendants. On dit que deux variables x_1 et x_2 de moyenne nulle sont décorrelées si :

$$E[x_1, x_2] = 0 \qquad \text{(décorrelées).} \tag{3.37}$$

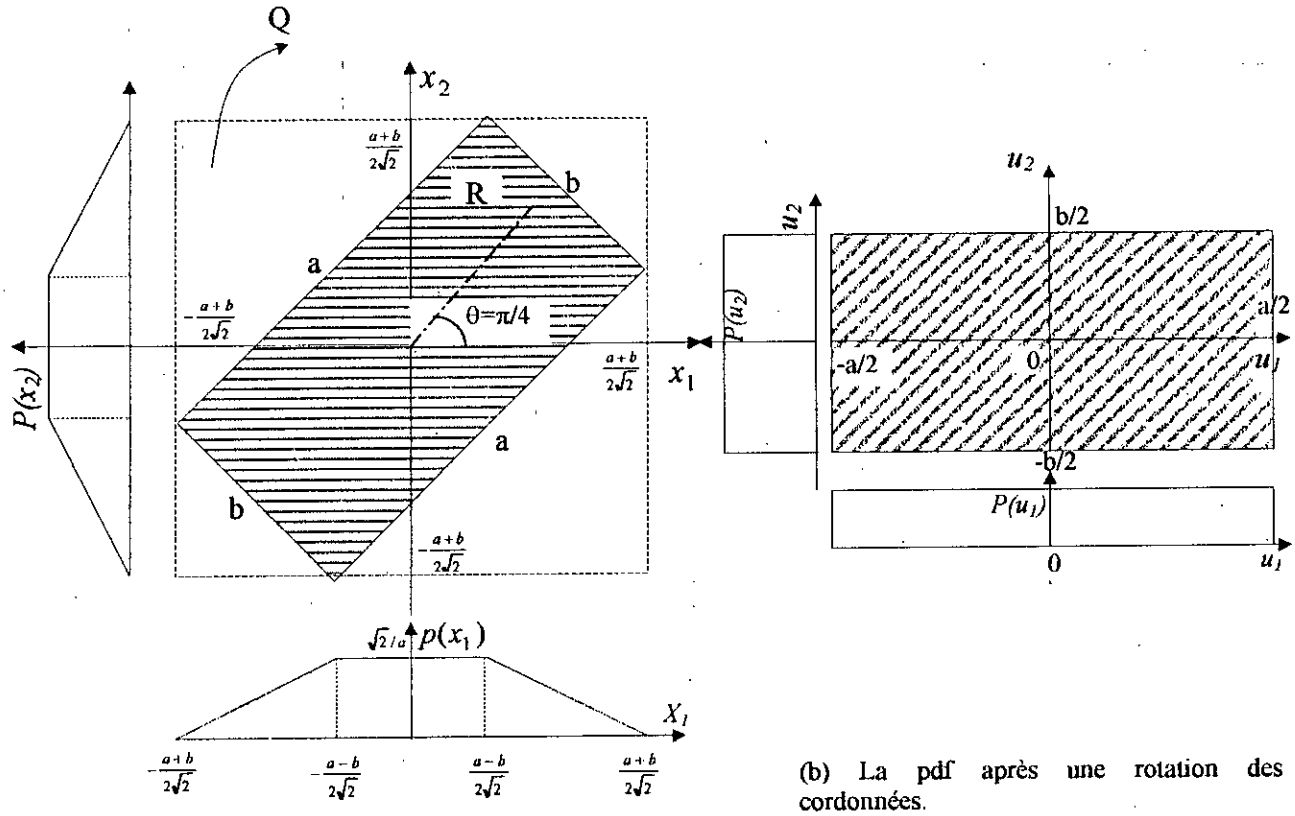
Mais x_1 et x_2 sont indépendantes si et seulement si leurs fonction densité de probabilité conjointe est égale au produit des densités marginales des variables x_1 et x_2

$$p(x_1, x_2) = p(x_1)p(x_2) \tag{3.38}$$

si x_1 et x_2 sont décorrelées mais dépendantes, cette dépendance est dite non-linéaire. Voyons maintenant comment on peut tirer profit de ces deux types de dépendances pour réduire le débit nécessaire pour la transmission.

Examinons l'exemple illustré par la figure 3.13 section (a) où x_1 *et* x_2 sont deux variables aléatoires avec une fonction densité de probabilité conjointe $p(x_1, x_2)$ [5].

$$p(x_1, x_2) = p(x) = \begin{cases} \dfrac{1}{ab} & x \in R \\ 0 & \textit{ailleurs} \end{cases} \tag{3.39}$$



(a) x_1 et x_2 dans cet exemple sont correlées

(b) La pdf après une rotation des cordonnées.
u_1 et u_2 sont maintenant décorrelées et indépendantes

Figure 3.13 Exemple de deux variables aléatoires avec une pdf uniforme à 2 dimensions

Les densités marginales $p(x_1)$ et $p(x_2)$ sont égales mais il est clair que l'équation (3.38) n'est pas vérifiée, et donc x_1 et x_2 sont dépendants.

On peut aussi montrer que x_1 et x_2 sont corrélées et par conséquent l'équation (3.37) n'est pas vérifiée non plus.

Essayons de faire une quantification scalaire uniforme de x_1 et x_2 indépendamment. Dans une QS uniforme, les intervalles de quantification R_i ont une même longueur égale Δ. Puisque x_1 et x_2 sont rangées entre $-(a+b)/2\sqrt{2}$ et $(a+b)/2\sqrt{2}$, le nombre de niveaux nécessaire pour quantifier chaque variable est de :

$$L_1 = L_2 = \frac{a+b}{\sqrt{2}\Delta}, \tag{3.40}$$

x_1 et x_2 peuvent être codées avec $R_1 = log_2 L_1$ bits *et* $R_2 = log_2 L_2$ bits respectivement. Le vecteur x peut être codé ensuite avec :

$$B_x = R_1 + R_2 = \log_2 L_1 L_2 = \log_2 \frac{(a+b)^2}{2\Delta^2} \text{ bits.} \tag{3.41}$$

Les deux quantificateurs scalaires correspondent à l'utilisation d'un quantificateur vectoriel avec L_x niveaux de sorties.

$$L_x = L_1 L_2 = \frac{(a+b)^2}{2\Delta^2}. \tag{3.42}$$

Dans ce cas la région de quantification s'étale au carré noté Q dans la figure 3.13 (a). Un tel quantificateur aura des cellules de quantification en forme de carrés d'une surface égale à Δ^2. L'utilisation d'un tel quantificateur dans cet exemple montre le gaspillage de bits puisque il assigne des bits pour des régions de probabilité zéro.

Faisons subir maintenant une rotation à l'exemple précèdent (figure 3.13 section b). Le vecteur x devient un autre vecteur u. On peut montrer que les deux nouvelles cordonnées sont décorrelées. A partir des pdf marginales on peut en déduire que :

$$p(u_1, u_2) = p(u_1)p(u_2) \qquad \forall\, u_1, u_2 \tag{3.43}$$

u_1 et u_2 sont donc d'après l'équation (3.38) indépendantes

pour un QS uniforme avec un intervalle de quantification Δ, le nombre des niveaux sera :

$$L_1 = \frac{a}{\Delta} \text{ et } L_2 = \frac{b}{\Delta} \quad L_u = L_1 L_2 = \frac{ab}{\Delta^2} \tag{3.44}$$

le nombre en bits correspondant est :

$$B_u = \log_2 \frac{ab}{\Delta^2}. \tag{3.45}$$

La différence en nombre de bits nécessaire pour coder x et u est donnée par :

$$B_x - B_u = \log_2 \frac{(a+b)^2}{2ab} \tag{3.46}$$

Pour $a = 2b$ nous aurons $B_x - B_u = 1.17 bits$, la rotation nous a permis d'économiser 1.17 bits par vecteur transmis.

L'exemple que on vient d'étudier, montre comment on peut tirer profit des avantages de la décorrélation à travers une rotation pour réduire le débit dans la quantification scalaire d'un vecteur.

Dans l'exemple suivant [5], on verra comment on peut tenir compte des avantages de la dépendance non-linéaire pour réduire le débit à travers une quantification vectorielle.

On a

$$p(u) = \begin{cases} \dfrac{8}{5ab} & u \in R \\ 0 & ailleurs \end{cases} \qquad (3.47)$$

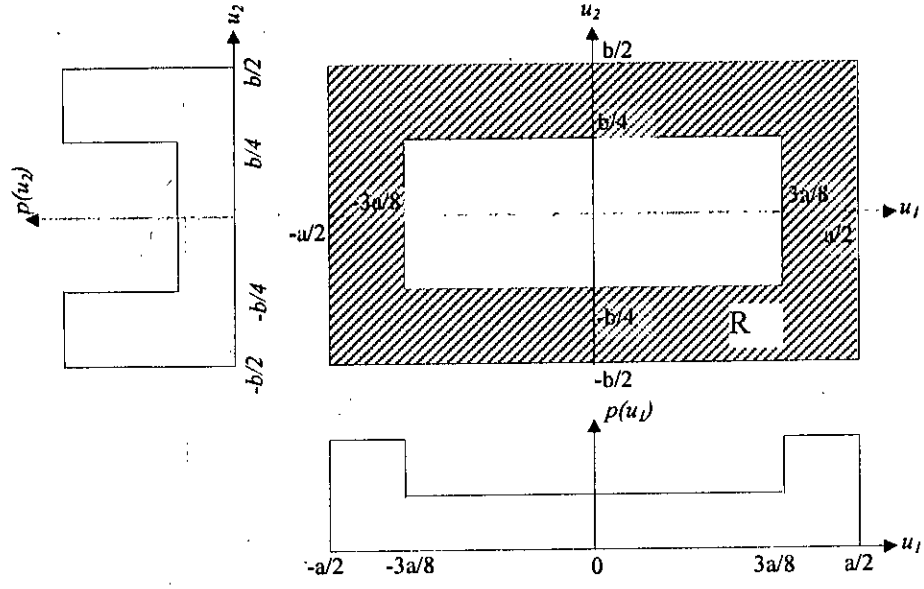


Figure 3.14 Exemple où les 2 variables u_1 et u_2 sont décorrelées mais dépendantes (dépendance non-linéaire)

Les variables u_1 et u_2 sont décorrelées mais dépendantes, on peut le vérifier à partir de ses pdf marginales dans la figure 3.14. C'est une dépendance non-linéaire. Une quantification scalaire de u_1 et de u_2 chacun, donnerait le même débit B_u que dans l'équation (3.45). Pour exploiter cette dépendance non-linéaire, on doit utiliser une quantification vectorielle où les partitions sont seulement dans la surface hachurée, donc pas de gaspillage de bits dans le petit rectangle de

probabilité zéro. La partie hachurée sera divisée en des surfaces carrées Δ^2. Le nombre de niveaux et les bits seront donnés par :

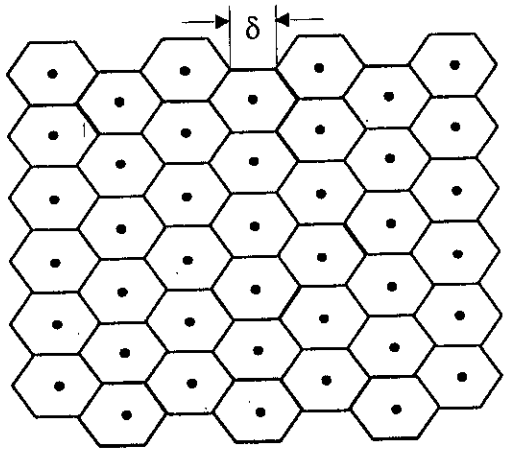
$$L'_u = \frac{5ab}{8\Delta^2} \qquad B'_u = \log_2 \frac{5ab}{8\Delta^2}. \tag{3.48}$$

La réduction en bits entre un quantificateur scalaire et un quantificateur vectoriel est dans ce cas :

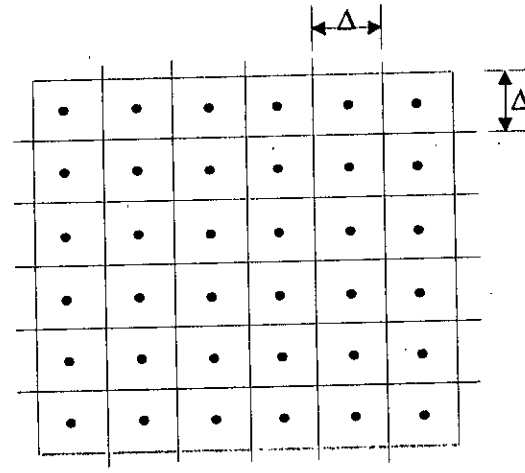
$$B_u - B'_u = \log_2 \frac{8}{5} = 0.68 \text{ bits.} \tag{3.49}$$

La dépendance non-linéaire nous permis d'économiser dans ce cas 0.68 bits/vecteur.

2- ***La Dimension :*** Dans l'exemple précèdent, on a vu que c'est la forme carrée qui a été employée pour tous les cellules. La forme du carré était inspirée à partir des quantificateurs scalaires utilisés, (produit cartésien). Mais l'une des propriétés d'un quantificateur vectoriel à des dimensions élevées, c'est qu'il a la liberté de choisir une autre forme de cellule telle que l'hexagone dans une partition à deux dimensions. Pour comparer les performances des régions de voronoï carrées avec un quantificateur dont les régions de voronoï sont hexagonales, on calcule l'erreur quadratique moyenne [5].



Quantification à deux dimensions d'une source uniforme avec des cellules hexagonales.



Quantification à deux dimensions d'une source uniforme avec des cellules carrées.

Figure 3.15 Exemple montrant l'influence de la dimension dans la forme des cellules lors de la conception du quantificateur optimal.

L'aire de l'hexagone d'après la figure 3.15 est donné par :

$$A_H = \frac{3\sqrt{3}}{2}\delta^2 . \tag{3.50}$$

En considérant que les valeurs de sorties sont localisées aux centres des cellules, on peut montrer que l'erreur quadratique moyenne est donnée par :

$$E_S = \frac{\Delta^4}{6} \qquad \text{(carré)} \tag{3.51}$$

$$E_H = \frac{5\sqrt{3}}{8}\delta^4 \qquad \text{(hexagone)} \tag{3.52}$$

l'erreur quadratique moyenne totale est obtenue en multipliant E_S et E_H par le nombre des cellules. Si nous imposons que la surface de l'hexagonale soit égale à la surface du carré $(A_S = A_H)$ et en négligeant les effets de bords, les deux quantificateurs auront le même nombre de cellules dans une même surface et ainsi le même débit.

Dans ce cas le rapport entre les deux distorsions sera égal à :

$$\frac{E_H}{E_S} = \frac{5\sqrt{3}}{9} = 0.962 . \tag{3.53}$$

Ceci veut que la forme hexagonale offre une erreur quadratique moyenne plus petite que la forme carrée. Ceci est un autre avantage de la quantification vectorielle.

3- ***La forme de la fonction densité de probabilité pdf*** *(probabilty density function) :* Dans les exemples précédents la densité de probabilité était uniforme.

L'exemple suivant (figure 3.16) montre l'influence de la densité de probabilité (gaussienne) pour la détermination des régions de voronoï qui n'auront plus la même forme et taille.

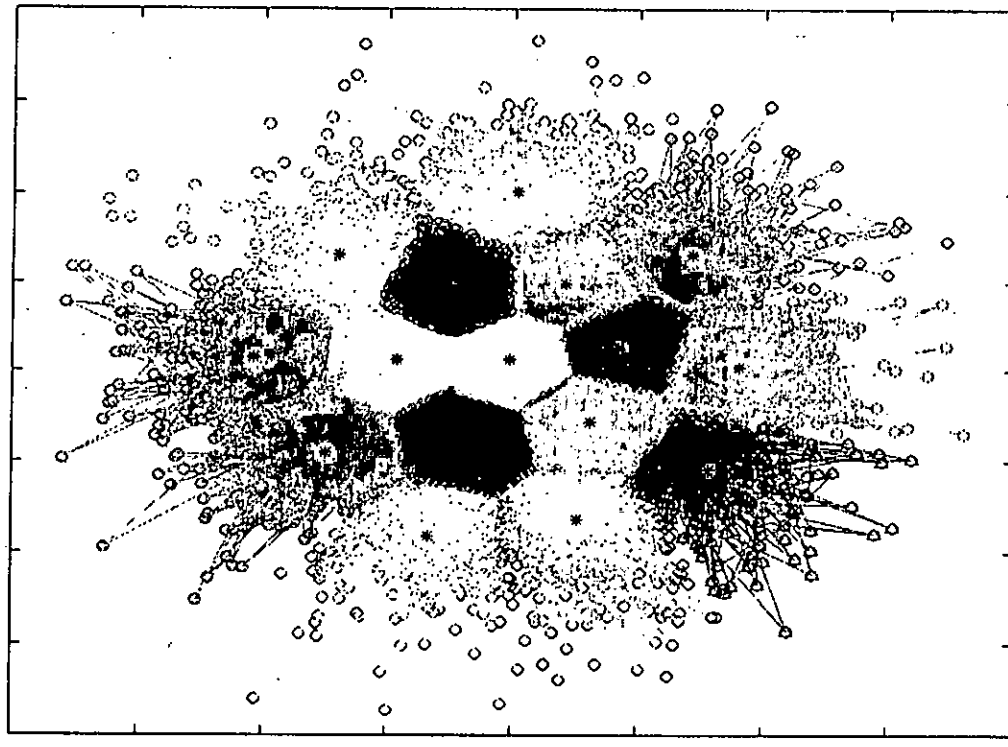


Figure 3.16 Exemple d'une QV obtenue par l'algorithme de Lloyd-Max avec N=2, L=16 et une pdf gaussienne (constellation 1,6,9).

3.4 Conclusion

Nous venons de voir les principes de la quantification scalaire et vectorielle. La connaissance de ces principes s'avère nécessaire pour une meilleure compréhension de la quantification des paramètres intervenant dans ce type de codage. En fait le codage de l'excitation est obtenu à l'aide d'un quantificateur vectoriel et le codage des paramètres des filtres prédicteurs peut être réalisé par un quantificateur scalaire ou vectoriel.

Chapitre 4

Modélisation de la Périodicité dans les codeurs CELP

4.1 Introduction

Comme nous l'avons déjà vu dans le chapitre 2, il existe une certaine périodicité dans le signal de parole qui correspond à la période de vibration des cordes vocales. Cette information est caractérisée par l'utilisation du filtre prédicteur long-terme *P(z)*. Ce filtre permet de reproduire la structure périodique du signal de parole en introduisant des segments de l'excitation passée.

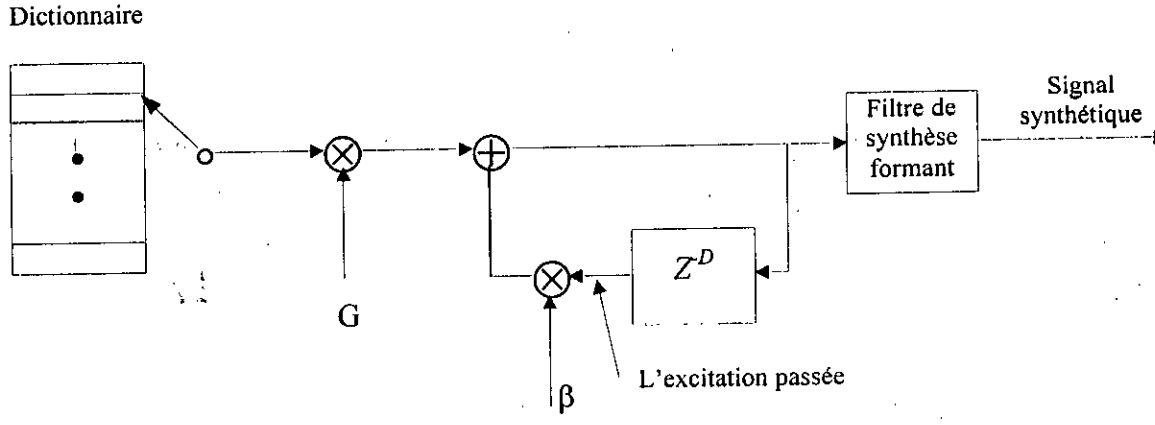


Figure 4.1 Synthèse pitch dans les codeurs CELP

Cette opération peut être vue comme une sélection d'un vecteur d'échantillons à partir d'un dictionnaire adaptatif (figure 4.2). Suivant cette interprétation le signal d'excitation peut être

considéré comme une combinaison linéaire de vecteurs issus d'un *dictionnaire adaptatif* (comportant les excitations passées) a priori connu et d'un dictionnaire (stochastique ou algébrique) à déterminer.

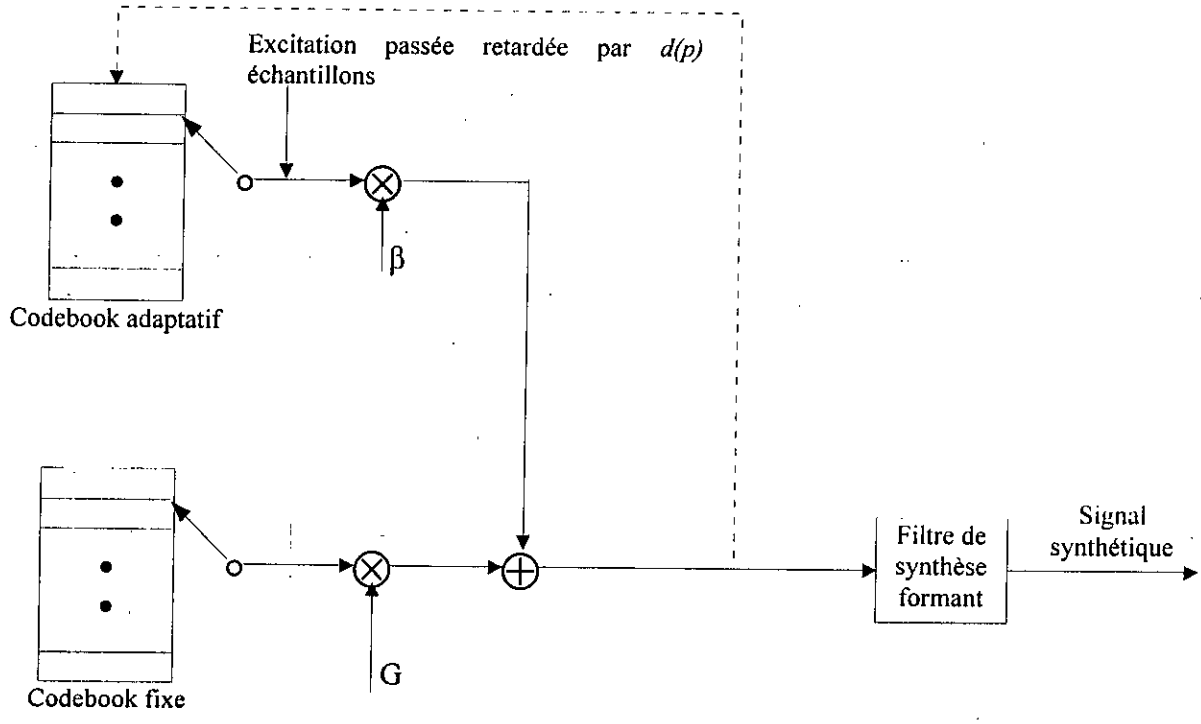


Figure 4.2 Codeur CELP avec un dictionnaire adaptatif

4.1.1 Définitions et réalisations

On appelle dictionnaire adaptatif l'ensemble des mots de code constitués à partir de la mémoire des excitations. Le vecteur *u(n)* issu du dictionnaire adaptatif, est constitué à partir du signal d'excitation $\hat{r}(n)$ lequel est défini pour $-D_{\max} \le n < 0$.

D_{max} est le retard maximum (période pitch), et correspond à 147 échantillons pour un signal échantillonné à 8 kHz. Pour $n \ge 0$, on écrit

$$u(n) = \hat{r}[n - d(p)] \qquad \text{pour } n=0,\ldots,N-1, \tag{4.1}$$

où N est la taille de la sous-trame et $d(p)$ est le délai pour le dictionnaire adaptatif.

$$\varepsilon = \sum_{n=-\infty}^{+\infty} (W_e(n) \sum_{k=0}^{L} C'_k S_w(n - D_k))^2 \tag{2.22}$$

$$= \sum_{i,j=0}^{L} C'_i C'_j \sum_{n=-\infty}^{+\infty} W_e^2(n) S_W(n - D_i) S_W(n - D_j)$$

$$\varepsilon = \sum_{i,j=0}^{L} C'_i C'_j \phi_{ss}(D_i, D_j) \tag{2.23}$$

avec $$\phi_{ss}(D_i, D_j) = \sum_{n=-\infty}^{+\infty} W_e^2(n) S_W(n - D_i) S_W(n - D_j), \tag{2.24}$$

la dérivée partielle par rapport à C'_i $(\frac{\delta \varepsilon}{\delta C'_i} = 0)$ donne :

$$\sum_{j=1}^{L} C'_j \phi_{ss}(D_i, D_j) = - \phi_{ss}(0, D_i) \tag{2.25}$$

et puisque $C'_j = - C_j$ pour j ≥ 1, on peut écrire :

$$\sum_{j=1}^{L} C_j \phi_{ss}(D_i, D_j) = \phi_{ss}(0, D_i). \tag{2.26}$$

Ecrivant l'équation (2.26) sous une forme matricielle (Φ **C** = **a**) :

$$\begin{bmatrix} \phi(D_1, D_1) & \phi(D_1, D_2) & \cdots & \phi(D_1, D_L) \\ \phi(D_2, D_1) & \phi(D_2, D_2) & \cdots & \phi(D_2, D_L) \\ \vdots & \vdots & & \vdots \\ \phi(D_L, D_1) & \phi(D_L, D_2) & \cdots & \phi(D_L, D_L) \end{bmatrix} \begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_L \end{bmatrix} = \begin{bmatrix} \phi(0, D_1) \\ \phi(0, D_2) \\ \vdots \\ \phi(0, D_L) \end{bmatrix} \tag{2.27}$$

La matrice Φ est définie symétrique et positive ($\forall$ a, $a^T.\Phi.a > 0$, et a est une matrice colonne), elle est aussi une matrice de Toeplitz si les retards d'inter-coefficients sont égaux. Selon que Φ est de Toeplitz ou non, la résolution du système d'équations peut se faire par la récursion de Levinson ou la décomposition de Cholesky. Pour un prédicteur court-terme $D_k = k$ pour $k = 1, \ldots, p$, tandis que pour un prédicteur pitch d'ordre N_p, $D_k = D + k$ pour $k = 0, \ldots, N_p - 1$. Quand $W_e(n) = 1 \; \forall n$, la formulation ci-dessus aboutit à la méthode d'autocorrélation. La méthode de covariance est

Pour une fréquence d'échantillonnage de 8kHz, $d(p) \in [D_{\min} = 20, D_{\max} = 147]$. Ceci permet de reconstituer des fréquences du fondamental allant de 54.4 Hz à 400 Hz.

Les valeurs comprises entre $\hat{r}[-147]$ et $\hat{r}[-1]$ sont la mémoire des fenêtres d'analyse précédentes (excitations passées).

Notons m le vecteur contenant toutes les valeurs de la mémoire :

$$m = \{\hat{r}[-D_{\max}] \cdots \cdots \hat{r}[-D_{\max}+N-1] \;\; \hat{r}[-D_{\max}+N] \cdots \hat{r}[-D_{\min}] \cdots \hat{r}[-1]. \quad (4.2)$$

Supposons que $d(p) = D_{max}$, alors le vecteur $u(n)$ sera :

$$u(n) = \{\hat{r}[-D_{\max}] \;\; \hat{r}[-D_{\max}+1] \cdots \hat{r}[-D_{\max}+N-1]\}. \quad (4.3)$$

Le dictionnaire adaptatif U_a, peut s'écrire sous la forme suivante :

$$U_a = \begin{bmatrix} \hat{r}[-N] & \cdots & \hat{r}[-1] \\ \vdots & \vdots & \vdots \\ \hat{r}[-2N+1] & \cdots & \hat{r}[-N] \\ \hat{r}[-2N] & \cdots & \hat{r}[-N+1] \\ \vdots & \vdots & \vdots \\ \hat{r}[-D_{\max}] & \cdots & \hat{r}[-D_{\max}+N-1] \end{bmatrix} \quad (4.4)$$

Posons $D = d(p)$, on peut constater deux cas :

- Lorsque $D \geq N$, $u(n)$ est constitué uniquement à partir des excitations passées,

$$u(n) = \{\hat{r}[-D] \cdots \hat{r}[-D+N-1]\}. \quad (4.5)$$

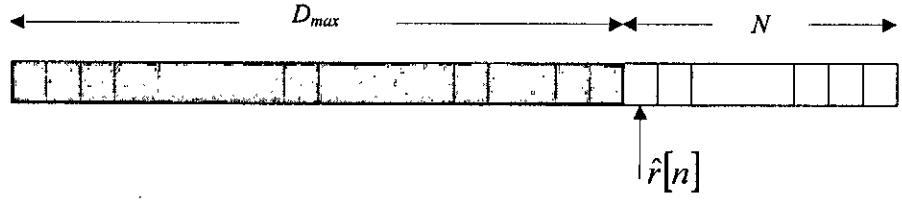
- Lorsque $D \leq N$, le vecteur issu de la mémoire des excitations passées n'est plus complet.

$$u(n) = \{\underbrace{\hat{r}[-D] \cdots \hat{r}[-1]}_{D} \; \underbrace{0 \cdots 0}_{N-D}\} \quad (4.6)$$

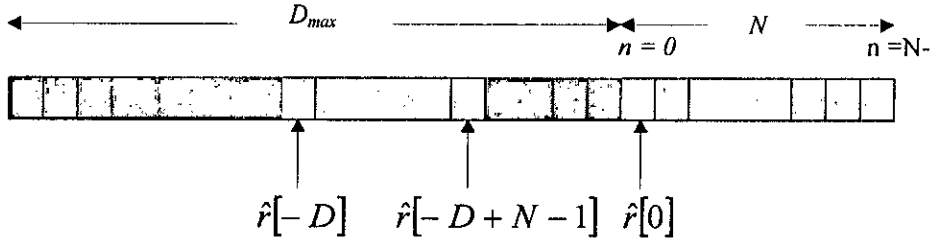
les $(N-D)$ échantillons de $u(n)$ restants correspondent à des échantillons de l'excitation courante.

Le faite de choisir $D \geq N$ reviendrait à limiter la fréquence fondamentale à une valeur maximale qui n'est pas réaliste.

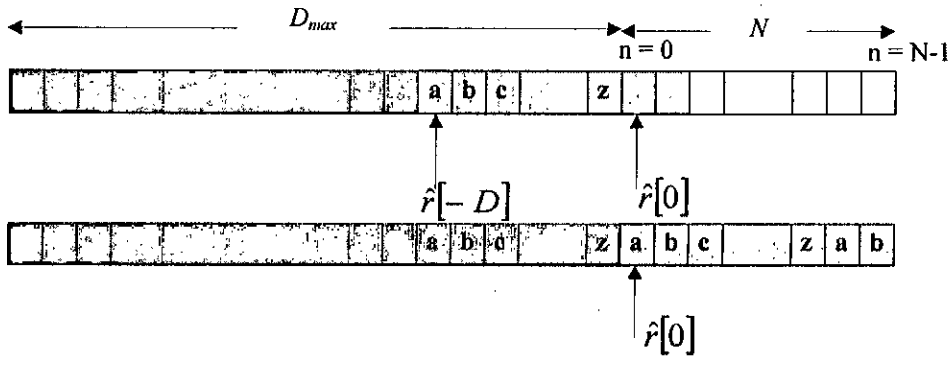
Une solution élégante à ce problème consiste à prolonger le dictionnaire des mémoires vers la droite en le construisant de la façon suivante [4], [8]. La première fois, on utilise les *N-1* échantillons disponibles $\hat{r}[-N+1]\cdots\hat{r}[-1]$ et on complète par $\hat{r}[-N+1]$, la deuxième fois on utilise les *N-2* échantillons $\hat{r}[-N+2]\cdots\hat{r}[-1]$ et les complète par $\hat{r}[-N+2]\hat{r}[-N+3]$, etc. On s'arrête à l'indice D_{min} .



(a) L'état du dictionnaire adaptatif avant la sélection de l'excitation



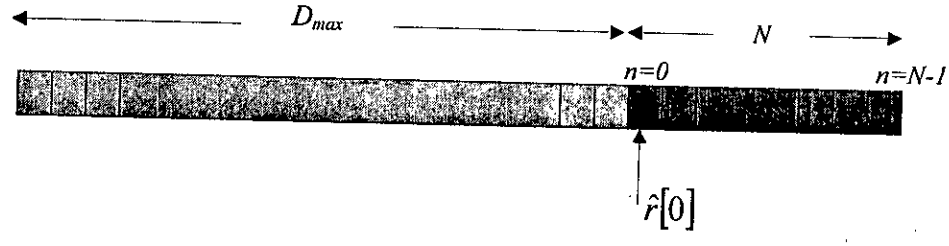
(b) Sélection d'un segment à partir de l'excitation passée quand $D \geq N$



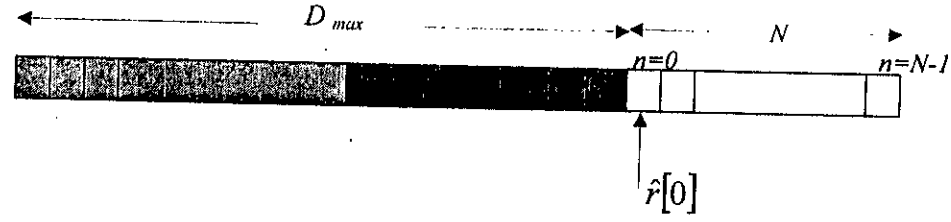
(c) Sélection d'un segment à partir de l'excitation passée lorsque $D = N\text{-}2$

Figure 4.3 Procédure de sélection de l'excitation passée

Lorsqu'on passe d'une fenêtre d'analyse à la suivante, l'ensemble du dictionnaire n'est pas remis en cause. Uniquement N vecteurs doivent être actualisés (voir figure 4.4). Les autres s'en déduisent par une translation vers la gauche.



(a) L'excitation résultante



(b) Actualisation du dictionnaire adaptatif

Figure 4.4 Procédure d'actualisation du dictionnaire adaptatif

On peut dans ce cas représenter le dictionnaire adaptatif comportant l'excitation passée comme suit :

$$U_a = \begin{bmatrix} \hat{r}[-D_{\min}] & \cdots & \hat{r}[-1] & \hat{r}[-D_{\min}] & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{r}[-N+1] & \cdots & \cdots & \hat{r}[-2] & \hat{r}[-1] & \hat{r}[-N+1] \\ \hat{r}[-N] & \hat{r}[-N+1] & \cdots & \cdots & \hat{r}[-2] & \hat{r}[-1] \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{r}[-D_{\max}] & \cdots & \cdots & \cdots & \cdots & \hat{r}[-D_{\max}+N-1] \end{bmatrix} \quad (4.7)$$

La figure 4.3 (c) donne un exemple pour cette méthode avec $D=N-2$.

4.1.2 Détermination de l'excitation

Dans cette section on va décrire les procédures utilisées pour la sélection de la meilleure excitation qui va attaquer le filtre de synthèse court-terme. On commence d'abord par rechercher le vecteur optimal du dictionnaire adaptatif (i.e. la période) et son gain β. Cette méthode n'est pas très différente de celle traiter dans la section 2.2.3, seulement ici l'excitation résultante est constituée à partir d'une contribution d'un dictionnaire adaptatif et d'un autre fixe.

Soit $S_w = [s[0],\ldots,s[N-1]]^T$ le vecteur parole et $\hat{r} = [\hat{r}[0],\ldots,\hat{r}[N-1]]^T$ le vecteur excitation résultant du dictionnaire adaptatif. Ce dernier sera noté u_p avec :

$$u_p[n] = \hat{r}[n - d(p)] = \hat{r}[n - D], \qquad n = 0,\ldots,N-1 \tag{4.8}$$

β étant le gain associé à ce vecteur.

Le vecteur résultant global sera donc donné par

$$\hat{r}_{ip} = \beta u_p + G c_i . \tag{4.9}$$

où c_i est le vecteur issu du dictionnaire fixé et G son gain associé.

Pour une performance optimale, la recherche dans les dictionnaires adaptatifs et fixe doit se faire simultanément. Cependant, vue la grande complexité requise, cette recherche est souvent réalisée séquentiellement avec le dictionnaire adaptatif en premier. Ce choix est justifié par le fait que c'est le dictionnaire adaptatif qui produit normalement la plus large contribution pour le signal d'excitation résultant dans un signal de parole.

$$\hat{r}_{opt} = \beta_{opt} u_{opt} + G_{opt} c_{opt} . \tag{4.10}$$

La sortie du filtre de synthèse formant pondéré $\hat{S}_w$ peut s'écrire sous la forme suivante

$$\hat{S}_w[n] = \sum_{i=0}^{\infty} h(i)\, \hat{r}(n-i), \qquad n=0,\ldots,N\text{-}1 \tag{4.11}$$

où h(i) est la réponse impulsionnelle du filtre de synthèse formant pondéré $(1/A(z/\gamma))$,

$$\frac{1}{A(z/\gamma)} = \frac{W(z)}{A(z)} \tag{4.12}$$

On peut décomposer l'équation 4.11 en deux termes :

$$\hat{S}_w[n] = \sum_{i=0}^{n} h(i)\hat{r}(n-i) + \sum_{i=n+1}^{\infty} h(i)\hat{r}(n-i), \tag{4.13}$$

On pose $\hat{S}_0 = \sum_{i=n+1}^{\infty} h(i)\,\hat{r}(n-i)$ ce terme correspond aux conditions initiales du filtre de synthèse $1/A(z/\gamma)$.

Ces conditions initiales sont dues aux excitations des sous-trames précédentes. Comme la séquence $\hat{S}_0$ est constante, on peut la soustraire du signal de parole original perceptualisé S_w. On aura :

$$S_t = S_w - \hat{S}_0, \tag{4.14}$$

S_t est appelé vecteur cible (target vector).

Utilisons maintenant la notation vectorielle par l'équation (4.13)

$$\hat{S}_w = H\,\hat{r}_{ip} + \hat{S}_0, \tag{4.15}$$

$$H = \begin{bmatrix} h_0 & 0 & 0 & \cdots & \cdots & 0 \\ h_1 & h_0 & 0 & \cdots & \cdots & 0 \\ h_2 & h_1 & h_0 & \ddots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ h_{N-2} & h_{N-3} & h_{N-4} & \cdots & h_0 & 0 \\ h_{N-1} & h_{N-2} & h_{N-3} & \cdots & h_1 & h_0 \end{bmatrix}$$

et $\{h_0, h_1, h_2, \ldots, h_{N-1}\}$ est la réponse impulsionnelle de $1/A(z/\gamma)$. En remplaçant (4.9) dans (4.15) on obtient

$$\hat{S}_w = \beta H u_p + G H c_i + \hat{S}_0. \tag{4.16}$$

La détermination du coefficient β et de la période D s'obtient par minimisation de l'erreur quadratique moyenne entre le signal original et le signal synthétique et ceci après avoir annuler la contribution du dictionnaire fixé ($Gc_i = 0$).

$$\varepsilon_0 = \left|S_w - \hat{S}_w\right|^2 = \left|S_w - \beta H u_p - \hat{S}_0\right|^2$$

$$= \beta^2 u_p^T H^T H u_p - 2\beta u_p^T H^T \underbrace{\left(S_w - \hat{S}_0\right)}_{S_t} + \left|S_w - \hat{S}_0\right|^2, \qquad (4.17)$$

le dernier terme est constant durant la recherche. Le coefficient β optimal est obtenu en dérivant l'erreur ε_0 par rapport à β, et en mettant ensuite le résultat égal à zéro.

$$\beta_{opt} = \frac{u_p^T H^T S_t}{u_p^T H^T H u_p} = \frac{\sum_{n=0}^{N-1} S_t[n]\ \hat{S}_d[n]}{\sum_{n=0}^{N-1} \hat{S}_d[n]\hat{S}_d[n]}, \qquad (4.18)$$

où $\hat{S}_d[n]$ est le résultat de la convolution entre l'excitation passée et la réponse impulsionnelle *h(i)*.

Substituons la valeur optimale de β dans l'équation (4.17), l'erreur devient

$$\varepsilon_0 = \left|S_t\right|^2 - \frac{\left[u_p^T H^T S_t\right]^2}{u_p^T H^T H u_p}. \qquad (4.19)$$

Posons ε_d égal à

$$\varepsilon_d = \frac{\left[u_p^T H^T S_t\right]^2}{u_p^T H^T H u_p} = \frac{\left[\sum_{n=0}^{N-1} S_t[n]\hat{S}_d[n]\right]^2}{\sum_{n=0}^{N-1} \hat{S}_d[n]\ \hat{S}_d[n]}. \qquad (4.20)$$

Pour que l'erreur ε_0 soit minimum, il faut que le second terme de l'équation (4.19) soit maximum. Le terme ε_d ne dépend que de la valeur de la période du fondamental (pitch) D.

Procédure de recherche

On cherche la valeur de *d(p)* ($d(p) \in [D_{\min}, D_{\min}+1, \ldots D_{\max}]$) qui rend ε_d maximum, cette valeur correspond à la valeur de la période *D*. Celle-ci est alors utilisée pour déterminer β grâce à la relation (4.18).

La recherche se fait de D_{min} à D_{max} et ceci par pas de 1 échantillon. Il n'est donc pas nécessaire de recalculer à chaque fois le produit de convolution entre H et u_p. On utilise généralement des relations de récurrence.

Quantification et codage

Une fois que les valeurs de β et D ont été déterminées, on doit les coder et ne transmettre au décodeur que des indices indiquant leurs positions dans un dictionnaire. Pour la période D on dispose de 128 valeurs possibles (de 20 jusqu'à 147) qu'on peut coder sur 7 bits. Pour le coefficient β, si on veut le coder par exemple sur 4 bits ($2^4 = 16$ représentants), il faut d'abord utiliser une base d'apprentissage qui est composée d'un grand nombre de valeurs représentatives de ces coefficients, et ensuite déterminer le quantificateur scalaire optimal.

4.2 Amélioration de la modélisation de la périodicité

Généralement la valeur définie pour la période D est l'entier le plus proche de la valeur réelle de la période. La numérisation du signal est la cause de cette approximation. Cette numérisation peut également engendrer une détermination d'un multiple de la période.

Nous avons jusqu'ici utilisé la méthode la plus simple et la plus économique, puisqu'elle nécessite l'évaluation de deux paramètres seulement (β et D) en utilisant un filtre prédicteur long-terme d'ordre 1 ou à un pas (one-tap pitch filter). Cependant il existe des méthodes qui peuvent améliorer davantage la détection du pitch, et améliorer ainsi sa précision et son efficacité, et par conséquent améliorer les performances des codeurs CELP.

Dans ce mémoire, nous avons utiliser deux méthodes pour attendre de meilleures performances. La première consiste à utiliser un filtre prédicteur long-terme avec un plus grand nombre de coefficients et ceci en augmentant l'ordre du filtre. Nous avons ainsi opté pour un filtre 3 pas (three-tap pitch filter). De cette méthode a découlé une autre technique qui lui est similaire (pseudo-three-tap pitch filter) mais qui nécessite moins de bits pour le codage. La deuxième méthode consiste à faire une interpolation des échantillons composant le dictionnaire. C'est la méthode du pitch fractionnaire.

4.2.1 Utilisation d'un filtre prédicteur pitch à trois pas

La fonction de transfert du prédicteur long-terme *P(z)* est définie par trois coefficients β_1, β_2, β_3, et la période *D*.

$$P_a(z) = 1 - (\beta_1 z^{-D+1} + \beta_2 z^{-D} + \beta_3 z^{-D-1}), \tag{4.21}$$

le filtre de synthèse pitch a pour expression

$$P_s(z) = \frac{1}{P_a(z)} = \frac{1}{1 - \sum_{i=-1}^{1} \beta_{i+2} z^{-(D+i)}}. \tag{4.22}$$

L'équation (4.16) devient maintenant :

$$\hat{S}_w = GHc_i + HUB + \hat{S}_0, \tag{4.23}$$

où *B* et *U* sont données par :

$$B = [\beta_1 \quad \beta_2 \quad \beta_3]^T,$$

$$U = \begin{bmatrix} \hat{r}[-D+1] & \hat{r}[-D] & \hat{r}[-D-1] \\ \hat{r}[2-D] & \hat{r}[1-D] & \hat{r}[-D] \\ \hat{r}[3-D] & \hat{r}[2-D] & \hat{r}[1-D] \\ \vdots & \vdots & \vdots \\ \hat{r}[N-D] & \hat{r}[N-1-D] & \hat{r}[N-D-2] \end{bmatrix}, \tag{4.24}$$

on calcul ensuite l'erreur quadratique moyenne ε_0 définie pour $Gc_i = 0$

$$\begin{aligned} \varepsilon_0 = \left|S_w - \hat{S}_w\right|^2 &= (S_t - HUB)^T (S_t - HUB) \\ &= |S_t|^2 - 2B^T U^T H^T S_t + B^T U^T H^T HUB, \end{aligned} \tag{4.25}$$

posons $\Phi = U^T H^T HU = \|HU\|^2$. et $g = U^T H^T S_t = [g_1 \quad g_2 \quad g_3]^T$.

La matrice Φ est symétrique, c'est-à-dire

$$\Phi = \begin{bmatrix} \phi_{11} & \phi_{21} & \phi_{31} \\ \phi_{21} & \phi_{22} & \phi_{32} \\ \phi_{31} & \phi_{32} & \phi_{33} \end{bmatrix},$$

$$\varepsilon_0 = |S_t|^2 - 2B^T g + B^T \Phi B, \tag{4.26}$$

En annulant les dérivées partielles de ε_0 par rapport aux coefficients, on aboutit au système d'équations suivant $\Phi B = g$.

$$\begin{bmatrix} \phi_{11} & \phi_{21} & \phi_{31} \\ \phi_{21} & \phi_{22} & \phi_{32} \\ \phi_{31} & \phi_{32} & \phi_{33} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} g_1 \\ g_2 \\ g_3 \end{bmatrix}. \tag{4.27}$$

Remplaçons B par $\Phi^{-1} g$ dans ε_0, on aura

$$\varepsilon_0 = |S_t|^2 - B^T g. \tag{4.28}$$

Pour résoudre le système d'équations, on utilise la décomposition de Cholesky. Cette opération est faite après que la période pitch soit déterminée (la maximisation du terme $B^T g$). La période varie toujours entre 20 et 147 pour une fréquence d'échantillonnage de 8 kHz.

4.2.2 Utilisation d'un filtre prédicteur pitch pseudo 3 pas

Le filtre prédicteur d'ordre 3 produit une meilleure qualité de parole que celui d'ordre 1. Cependant plus de bits sont nécessaires afin de coder les deux coefficients supplémentaires. Le filtre prédicteur long-terme pseudo 3 pas [2], possède trois coefficients non nuls avec un ou deux degrés de liberté. Ce filtre n'est pas aussi performant que celui du 3 pas, mais il permet d'obtenir une meilleure qualité que le filtre à 1 pas.

La méthode consiste à écrire deux coefficients en fonction du troisième

$$\beta_1 = \beta_3 = \alpha\beta, \qquad \text{avec } \beta = \beta_2, \tag{4.29}$$

où α et β sont optimisées pour une meilleure performance. On peut fixer α et avoir un seul degré de liberté.

Pour simplifier la notation les filtres pseudo trois-pas sont notés par nTmDF, n est le nombre de pas (ou l'ordre du filtre) et m le degré de liberté .

Dans le filtre à trois pas 3T3DF les coefficients β_1, β_2 et β_3 sont variables . Les filtres pseudo trois-pas sont 3T2DF (α et β variables) et 3T1DF (α est fixe et β variable).

Reprenons le système d'équations (4.27) en tenant compte de l'équation (4.29) et en posant $\gamma = \alpha\beta$ dans le cas d'un 3T2DF, on obtient

$$\begin{cases} (\phi_{11} + \phi_{33} + 2\phi_{31})\,\gamma + (\phi_{21} + \phi_{32})\,\beta = (g_1 + g_3) \\ (\phi_{21} + \phi_{32})\,\gamma + \phi_{22}\beta = g_2 \end{cases} \tag{4.30}$$

Posons $X = (\phi_{11} + \phi_{33} + 2\phi_{31})$, $Y = (\phi_{21} + \phi_{32})$, $Z = \phi_{22}$, $V = (g_1 + g_3)$, $W = g_2$.

L'équation (4.30) devient :

$$\begin{bmatrix} X & Y \\ Y & Z \end{bmatrix} \begin{bmatrix} \gamma \\ \beta \end{bmatrix} = \begin{bmatrix} V \\ W \end{bmatrix}. \tag{4.31}$$

La résolution de ce système permet d'obtenir γ_{opt} et β_{opt}

$$\begin{cases} \beta_{opt} = \dfrac{(XW - YV)}{(XZ - Y^2)} \\ \gamma_{opt} = \dfrac{(ZV - YW)}{(XZ - Y^2)} \end{cases} \tag{4.32}$$

D'une manière analogue pour α fixe et β variable c'est-à-dire 3T1DF on trouve

$$\beta_{opt} = \frac{\alpha(g_1 + g_3) + g_2}{\alpha^2(\phi_{11} + \phi_{33} + 2\phi_{31}) + \phi_{22} + 2\alpha(\phi_{32} + \phi_{21})}. \tag{4.33}$$

utilisée si $W_d(n) = 1 \ \forall n$. Comme la méthode de covariance donne des gains de prédiction élevés, elle est généralement préférée pour la prédiction pitch (prédiction long-terme).

Souvent on utilise un filtre predicteur d'ordre 1 c'est à dire à un seul retard D et un seul coefficient β. Le retard D est déterminé séparément du coefficient β en utilisant la méthode de covariance.

En prenant un prédicteur pitch d'ordre 1 ($N_p = 1$), l'équation (2.27) devient :

$$\phi(D, D)\, C_1 = \phi(0, D), \tag{2.28}$$

le coefficient optimal sera donné par :

$$\beta_{opt} = C_1 = \frac{\phi(0, D)}{\phi(D, D)}, \tag{2.29}$$

Ainsi, l'erreur quadratique moyenne devient :

$$\varepsilon = \phi(0,0) - \frac{\phi^2(0, D)}{\phi(D, D)}. \tag{2.30}$$

Minimiser ε revient à maximiser le terme $\phi^2(0,D)/\phi(D, D)$. Cette fonction est calculée pour toutes les valeurs possibles de D, et le maximum indique le meilleur choix de la période pitch D. Pour une fréquence d'échantillonnage de 8 kHz, D est compris entre 20 et 147 (128 valeurs) et ceci pour couvrir la plupart des valeurs du pitch rencontrées dans la parole humaine (54.2 Hz –400 Hz).

2.3 Technique d'analyse par synthèse

C'est le codeur CELP (code exited linear predictive coder) proposé par B.S.Atal et M.Schroeder en 1984 qui reste la base des techniques d'analyse par synthèse. Le terme d'analyse par synthèse signifie que le codage de la parole s'effectue à l'émetteur et ceci après avoir généré le signal synthétique (le décodeur fait partie intégrante du codeur). les paramètres trouvés sont ensuite transmis au récepteur pour reconstruire le signal vocal.

Dans le codage LPAS (linear prediction analysis by synthesis), le signal vocal est considéré comme stationnaire sur une durée comprise entre 10 à 30ms, ce qui correspond à un intervalle d'échantillons de 80 à 240, pour une fréquence d'échantillonnage de 8kHz.

4.2.3 Utilisation du pitch fractionnaire

Cette méthode consiste à augmenter la résolution temporelle par l'utilisation de filtres interpolateurs. Le coût de codage est très faible puisqu'il vaut un bit par sous-trame. On parle alors de *pitch fractionnaire* car le retard D n'est plus un entier mais possédant une fraction de la période d'échantillonnage [14], [15], [16], [17], [18], [19].

Essayons de généraliser l'équation (4.1) pour des valeurs fractionnaires,

$$u[n] = \beta \hat{r}[n - D], \qquad n = 0, \ldots, N-1,$$

substituer le retard entier D par une valeur réelle n'est pas possible, puisque le signal discret $\hat{r}[n]$ n'est pas défini pour les délai non-entiers. Cette difficulté est surmontée en suivant le traitement donné par [14].

Le retard fractionnaire peut être représenté comme un retard entier D à la fréquence d'échantillonnage f_e auquel s'ajoute une fraction t/I, $t = 0,1,\ldots,I-1$, où I est la résolution temporelle, spécifiée comme étant un multiple de la fréquence d'échantillonnage f_e,

$$D' = D + \frac{t}{I}, \qquad t = 0,1,\ldots,I-1. \tag{4.34}$$

Le retard fractionnaire t/I à la fréquence d'échantillonnage f_e correspond à un délai t à la fréquence d'échantillonnage $I f_e$. En d'autres termes, pour obtenir un délai de t/I échantillons, la fréquence devra être multipliée par un facteur I (en insérant $I-1$ échantillons nuls entre chaque échantillon du signal $\hat{r}[n]$). Le signal obtenu est ensuite filtré par un filtre passe-bas afin d'avoir la version interpolée. Le signal interpolé est ensuite retardé de t échantillons puis sous-échantillonné par un facteur de I (en sélectionnant chaque I-ième échantillon). Le signal résultant (échantillonné de nouveau à f_e) , est le signal original retardé par un retard non-entier t/I. Il est nécessaire de souligner qu'un retard constant supplémentaire est introduit en raison du retard dû au filtre d'interpolation passe-bas.

Le filtre d'interpolation $b[n]$, $n = 0,1,\ldots,M-1$ est un filtre à phase linéaire de M échantillons. Le délai de ce filtre à la fréquence d 'échantillonnage $I f_e$ est de $(M-1)/2$ échantillons. On choisira M de façon à compenser ce délai à la fréquence f_e tel que $(M-1)/2$ soit un multiple de I,

$$M = 2\tau I + I, \tag{4.35}$$

où τ est le retard entier introduit par le filtre interpolateur à la fréquence f_e.

Les délais fractionnaires peuvent aussi être obtenus en utilisant les filtres polyphases $p_t[n]$. Ces filtres sont obtenus à partir des coefficients du filtre passe-bas $b[n]$.

$$p_t[n] = b[nI - t], \qquad \text{pour} \quad t = 0,1,\ldots,I-1$$
$$\text{et} \quad n = 0,1,\ldots,K-1, \tag{4.36}$$

avec $b[n] = 0$ pour $n < 0$, et $p_t[0] = 0$ pour $t > 0$. Le nombre de coefficients du filtre polyphase p_t est donné par

$$K = 2\tau + 1. \tag{4.37}$$

Si l'on désire retarder un signal d'entrée $x(n)$ de t/I, le filtre polyphase correspondant est utilisé et le signal de sortie $y(n)$ peut s'écrire sous la forme suivante :

$$y(n) = \sum_{m=0}^{k-1} P_t(m)x(n-m), \qquad n = 0,\ldots,N-1. \tag{4.38}$$

Prenons en compte le retard τ du filtre passe bas, la contribution du dictionnaire adaptatif pour un retard fractionnaire $D + t/I$ est donnée par

$$\hat{r}[n] = \beta\, u_{opt} = \beta \sum_{m=0}^{N-1} p_t[m]\ \hat{r}[n - D + \tau - m], \tag{4.39}$$

La détermination du gain et de la période se fait d'une manière analogue à celle déjà vue dans 4.1.2, seulement ici la période pitch peut avoir des valeurs non-entières [1].

$$\beta_{opt} = \frac{\sum_{n=0}^{N-1} S_t[n]\, y[n]}{\sum_{n=0}^{N-1} y[n]\, y[n]}, \tag{4.40}$$

$y[n]$ est le vecteur d'excitation filtré

$$y[n] = \sum_{n=0}^{N-1} u_{opt}[\,i]h[n-i] \qquad \text{i} = 0,\ldots,\text{N} - 1. \tag{4.41}$$

[1] β_{opt} est généralement limité entre 0 et 1.4.

4.3 Conclusion

Nous venons de voir comment la modélisation de la périodicité est réalisée dans les codeurs CELP en introduisant des segments de l'excitation passée à partie d'un dictionnaire adaptatif. L'estimation de la période peut être améliorée en utilisant des techniques appropriées et qui consistent à augmenter l'ordre du filtre ou à augmenter la résolution temporelle. Toutefois, Cette amélioration peut entraîner une augmentation du débit de codage ou un accroissement en complexité.

Dans le chapitre qui va suivre on comparera les résultats obtenus par l'utilisation des différentes méthodes de détection du pitch.

- Filtre prédicteur d'ordre 1.
- Filtre prédicteur d'ordre 3.
- Filtre prédicteur d'ordre 3 (pseudo).
- Filtre prédicteur fractionnaire.

Chapitre 5

Simulation et Interprétation des Résultats

Les techniques exposées dans le chapitre précédent ont pour rôle d'améliorer la reproduction de la périodicité par le codeur, ce qui va engendrer une amélioration de la qualité de la parole synthétique. De plus l'amélioration de la qualité ne doit pas se faire au détriment du débit.

Le codeur CELP utilisé pour notre simulation est un codeur à 4.8 kbps réalisé au CDTA (équivalent au FS 1016) et qui possède les caractéristiques suivantes :

- Une analyse LPC à l'ordre p = 10 est réalisée sur des trames de 30 ms (240 échantillons).
- Codage des paramètres LSP sur 34 bits (par trame).
- Un dictionnaire algébrique de 12 bits (4096 vecteurs) avec le gain quantifié sur 5 bits (4 bits +signe).
- Un dictionnaire adaptatif de 7 bits (128 vecteurs). Le gain associé β est quantifié sur 3 bits.

Les paramètres des 2 dictionnaires sont transmis à chaque sous-trame de 7.5 ms (60 échantillons). L'allocation binaire pour notre codeur est résumée dans le tableau suivant.

paramètres		bits par trame de 30 ms	Nombre de bits total	Débit
Prédiction court-terme	10 *LSP*	4,4,4,4,3,3,3,3,3,3	34	1.133 kbps
Prédiction long-terme	*période D*	7*4	28	0.933 kbps
	β	3*4	12	0.4 kbps
Dictionnaire algébrique (codebook)	*index i*	12*4	48	1.6 kbps
	gain G	4*4	16	0.533 kbps
	bit signe	1*4	4	0.133 kbps
Total			142	4.732 kbps

Tableau 5.1 Allocation binaire du codeur CELP à 4.8 kbps

Les phrases tests utilisées correspondent à des locuteurs féminins et masculins et échantillonnées à 8 kHz. En tout nous avons utiliser 28 phrases phonétiquement équilibrées et dont nous pouvons citer quelques-unes.

Phrase 01 : « Il se garantira du froid avec un bon capuchon ».

Phrase 02 : « Annie s'ennuie loin de mes parents ».

Phrase 03 et 19 : « Les deux camions se sont heurtés de face ».

Phrase 04 : « Dés que le tambour bat, les gens accourent ».

Phrase 05 et 12 et 07 : « La vaisselle propre est mise sur l'évier » .

Phrase 06 : « Vous poussez, des cris de colère ».

Phrase 08 et 20 : « Un loup s'est jeté immédiatement sur la petite chèvre ».

Phrase 09 : « Je ne peux atteindre les bocaux de confiture ».

Phrase 10 : « Dans cette crémerie on vend du fromage fort ».

Phrase 11 : « La voiture s'est arrêtée au feu rouge ».

Phrase 13 : « La pirogue se met au travers du courant ».

Phrase 14 : « Elle a perdue tout contacte avec la Pologne ».

Phrase 15 : « Quand il s'est réveillé il était trop tard ».

Phrase 16 : « huit satellites ont été mobilisés».

Phrase 17 : « Là bas il y a de mauvaises vagues très hautes ».

Phrase 18 : « C'est la question que tout le monde se pose ».

Phrase 22 : « I must have reread that article three times before I realized what was bothering me ».

Phrase 25 : « The other memorable event in that conference was the worst presentation I have ever heard ».

5.1 Prédicteur long-terme à 3 pas et pseudo 3 pas

Comme nous l'avons déjà vu dans le chapitre 4 la fonction de transfert du prédicteur long-terme *P(z)* d'ordre trois est donnée par l'expression suivante :

$$P_a(z) = 1 - (\beta_1 z^{-D+1} + \beta_2 z^{-D} + \beta_3 z^{-D-1}). \qquad (5.1)$$

La procédure consiste à déterminer les trois coefficients $\beta_1, \beta_2, \beta_3$ et la période pitch D qui minimisent l'erreur quadratique moyenne..

La période sera toujours codée sur 7 bits par sous-trame, mais par contre le codage des coefficients β_1, β_2, β_3 contrairement à un seul coefficient β, nécessite plus de bits.

paramètres		bits par trame de 30 ms	Nombre de bits total	Débit
Prédiction court-terme	10 *LSP*	4,4,4,4,3,3,3,3,3,3	34	1.133 kbps
Prédiction long-terme	*période D*	7*4	28	0.933 kbps
	$\beta_1, \beta_2, \beta_3$	(4,3,4)*4	44	1.466 kbps
Dictionnaire algébrique (codebook)	*index i*	12*4	48	1.6 kbps
	gain G	4*4	16	0.533 kbps
	bit signe	1*4	4	0.133 kbps
Total			174	5.8 kbps

Tableau 5.2 Allocation binaire du codeur CELP avec un predicteur pitch d'ordre 3

La figure 5.1 représente les périodes trouvées dans certaines régions après utilisation des filtres prédicteurs d'ordre 1 et 3.

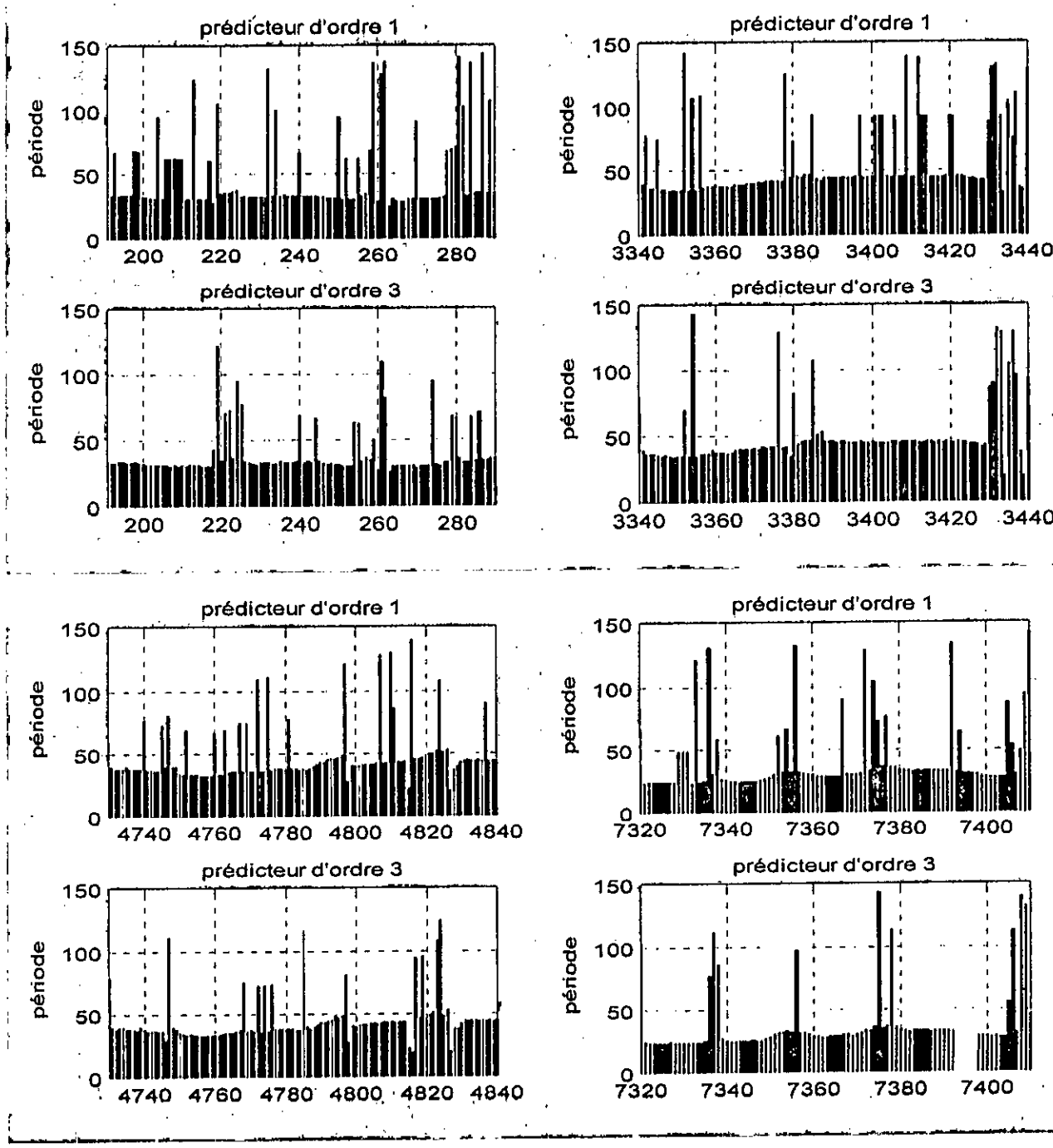


Figure 5.1 Les périodes trouvées dans quelques régions avec le prédicteur long-terme d'ordre 1et 3.

Cette figure représente les périodes obtenues dans certaines régions des signaux synthétiques selon les phrases tests utilisées. Les pics de période correspondent aux multiples de la valeur réelle de la période (double ou triple). Nous remarquons une diminution considérable de ces pics avec le filtre prédicteur d'ordre 3.

La figure 5.2 illustre les périodes déterminées pour une séquence plus grande dans les deux méthodes. Là aussi nous remarquons la diminution des doubles et des triples de la période.

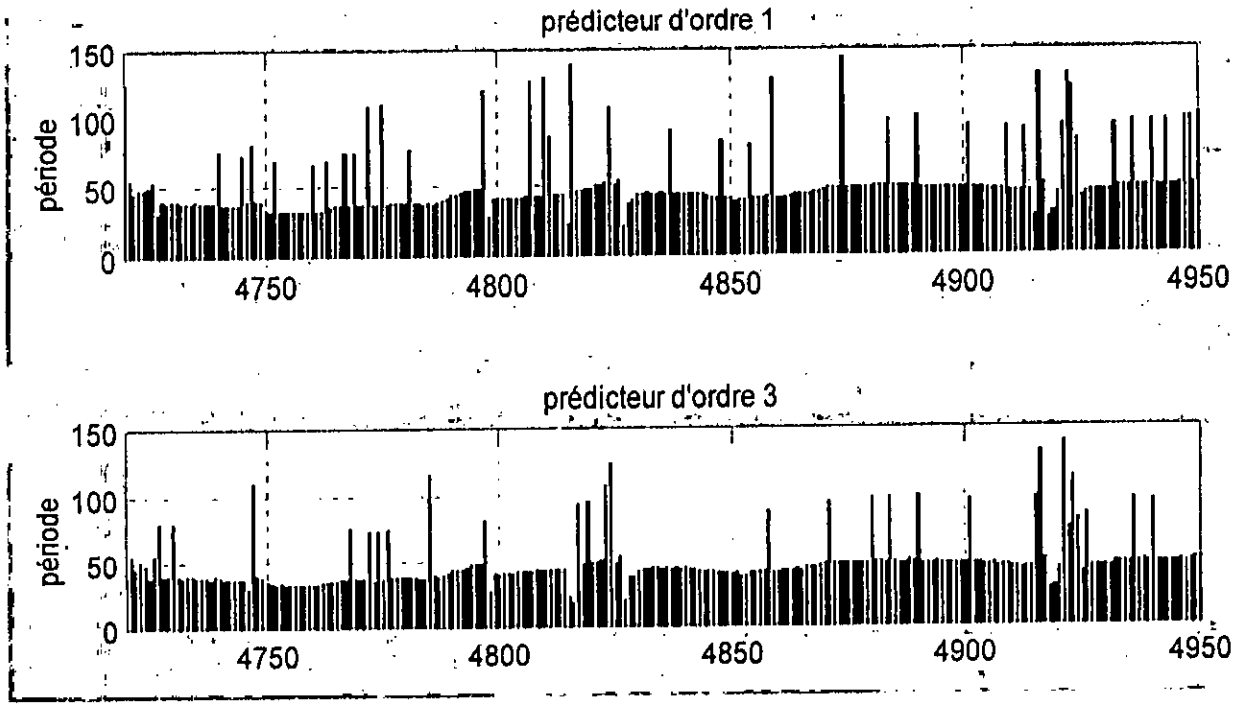


Figure 5.2 Evolution du délai optimal du prédicteur long-terme avec l'utilisation des prédicteurs d'ordre 1 et 3.

Cette diminution implique une meilleure reproduction de la périodicité, et elle se traduit par un signal synthétique d'une meilleure qualité. Le test objectif utilisé et qui consiste à calculer le rapport signal sur bruit segmental[1] (en dB) montre l'amélioration recherchée surtout pour les locuteurs féminins ou la dégradation est plus importante lors d'une prédiction long-terme ordinaire (ordre 1).

[1] Voir annexe A

		SNR_{segm}	
Les phrases tests	**Sexe**	**Prédicteur d'ordre 1**	**Prédicteur d'ordre 3**
PHRASE 01	M	9.09	9.41
PHRASE 02	M	11.40	12.32
PHRASE 03	M	9.60	9.77
PHRASE 04	M	10.16	11.03
PHRASE 05	F	8.32	8.89
PHRASE 06	M	7.69	8.93
PHRASE 07	M	9.09	9.78
PHRASE 08	F	8.62	8.69
PHRASE 09	M	8.93	9.66
PHRASE 10	M	8.48	8.37
PHRASE 11	M	9.67	9.85
PHRASE 12	M	9.36	9.51
PHRASE 13	F	8.32	8.46
PHRASE 14	F	8.67	9.04
PHRASE 15	F	6.95	7.49
PHRASE 16	F	8.45	9.01
PHRASE 17	F	8.37	8.78
PHRASE 18	F	9.03	9.69
PHRASE 19	M	8.26	8.41
PHRASE 20	M	7.39	7.54

Tableau 5.3 Comparaison du rapport signal/bruit segmental en utilisant les prédicteur long-terme d'ordre 1 et 3

Récapitulons et interprétons les résultats trouvés précédemment. L'augmentation de l'ordre du prédicteur long-terme permet une amélioration de la qualité mais au prix d'une importante augmentation du débit. En effet la quantification scalaire des deux coefficients supplémentaires sur 4 bits a engendrée une augmentation du débit de (8*4/30 ms) = 1.06 kbps, atteignant ainsi 5.8 kbps. Ce qui n'est pas vraiment notre but. L'utilisation d'une quantification vectorielle dans ce cas sur l'ensemble des coefficients (β_1, β_2 et β_3) en 7 ou 8 bits diminuera sensiblement le débit, mais le codage reste toujours moins efficace. Il existe quand même une autre solution qui consiste imposer un rapport fixé entre les coefficients. On aura ainsi à coder qu'un seul coefficient et donc le débit reste inchangé. C'est la méthode qu'on a vue dans la section 4.2.2 et qui est appelée pseudo-three-tap filter ou filtre prédicteur pseudo trois-pas et qui suppose ceci

$$\beta_1 = \beta_3 = \alpha\beta, \qquad \text{avec } \beta = \beta_2, \tag{5.3}$$

PHRASE 02									
α	SNR_{seg}	α	SNR_{seg}	α	SNR_{seg}	α	SNR_{seg}	α	SNR_{seg}
0.01	11.38	0.05	11.60	0.09	11.58	0.13	12.08	0.17	11.47
0.02	11.12	0.06	11.66	0.10	11.52	0.14	11.37	0.18	11.41
0.03	11.54	0.07	12.01	0.11	11.80	0.15	11.61	0.19	10.80
0.04	11.59	0.08	11.51	0.12	11.52	0.16	11.54	0.20	10.97
PHRASE 04									
α	SNR_{seg}	α	SNR_{seg}	α	SNR_{seg}	α	SNR_{seg}	α	SNR_{seg}
0.01	10.16	0.05	9.99	0.09	9.92	0.13	10.16	0.17	10.45
0.02	10.23	0.06	10.19	0.10	10.01	0.14	10.38	0.18	10.11
0.03	10.13	0.07	10.16	0.11	10.29	0.15	10.48	0.19	10.19
0.04	10.24	0.08	10.06	0.12	9.80	0.16	10.16	0.20	10.40

Tableau 5.4 Les $SNR_{seg}(dB)$ des phrases 2 et 4 obtenus avec le prédicteur d'ordre pseudo-3 pas

pour la simulation nous avons utilisé le 3T1DF (α est fixe et β variable). Nous cherchons la période et le gain optimal avec une valeur de α fixé qui va nous permettre d'avoir un résultat meilleur que

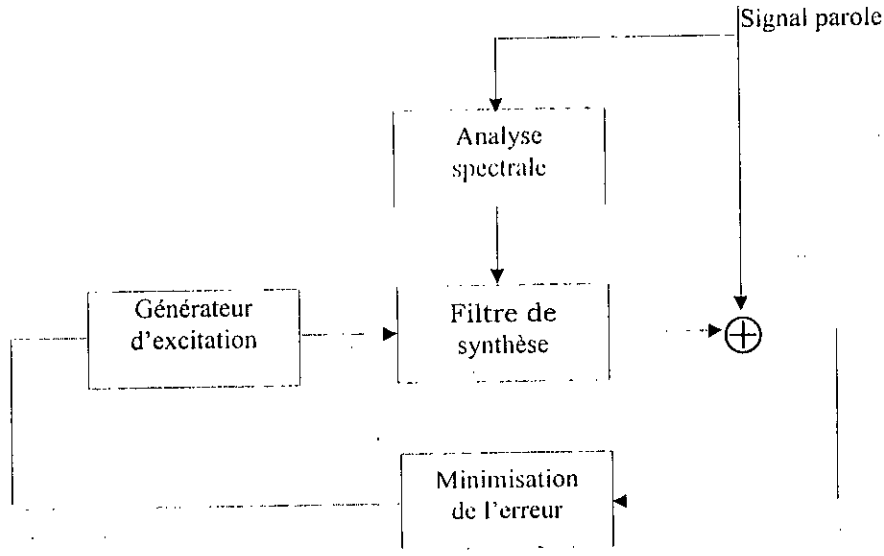


Figure 2.4 Codeur LPAS

Le signal de parole est segmenté de façon régulière en supposant que le principe de stationnarité est respecté ; C'est pour cela que sur chaque segment, les filtres de prédiction linéaire court-terme et long-terme sont supposés invariants. On désigne par trame ou fenêtre d'analyse chacun de ces segments. Pour chaque trame, on commence par déterminer à partir du signal vocal original les coefficients des prédicteurs court-terme et long-terme par une analyse LP (linear prediction analysis). Le signal d'excitation est filtré séquentiellement à travers les filtres de synthèse pour reproduire le signal reconstruit.

Le signal original est ensuite soustrait du signal reconstruit et l'erreur résultante est minimisée selon le critère de l'erreur quadratique moyenne. Le signal d'excitation qui correspond au minimum de l'erreur est sélectionné et les paramètres correspondant à ce signal (index) sont transmis au récepteur. Le récepteur utilise la même structure de synthèse pour reconstruire le signal de parole.

Dans la technique d'analyse par synthèse on peut introduire certaines caractéristiques de la perception auditive. Il est connu que l'oreille est incapable de distinguer deux fréquences proches si l'une d'elles possède une intensité élevée. Ce phénomène s'appelle le masquage.

Les fonctions de coût quadratiques se prêtent bien aux calculs ; elles possèdent la bonne propriété de fournir un système linéaire lorsque l'on dérive ce critère par rapport aux paramètres inconnus. Par contre, ce critère n'est pas forcément bien adapté à notre système auditif. Une correction perceptive est très largement utilisée pour palier cet inconvénient. On rajoute une fonction de

celui où nous utilisons un prédicteur d'ordre 1. En utilisant les phrases tests 02 et 04, nous avons calculé le rapport signal/bruit segmental pour différentes valeurs de α (tableau 5.4).
En comparant ces résultats avec le tableau 5.3 pour les phrases 02 et 04, on peut choisir de prendre $\alpha = 0.15$ et obtenir ainsi le tableau suivant.

	$SNR_{seg}\,(dB)$		
PHRASES	**Prédicteur long-terme d'ordre 1**	**Prédicteur long-terme d'ordre pseudo-3 pas avec $\alpha = 0.15$**	**Prédicteur long-terme d'ordre 3**
PHRASE 02	11.40	11.61	12.32
PHRASE 04	10.16	10.48	11.03

Tableau 5.5 Comparaison entre les différents prédicteurs pour $\alpha = 0.15$.

5.2 Méthode du pitch fractionnaire

L'interpolation est réalisée en utilisant un filtre FIR basé sur la fonction sinc pondérée par une fenêtre Hamming $w(n)$ et le facteur d'interpolation est égal à 3.

$$w(n) = 0.54 - 0.46 * \cos(2\pi n / N - 1) \quad \text{pour } n = 0, \ldots, N-1, \tag{5.4}$$

N est la longueur du filtre.
Nous cherchons la période comprise dans un ensemble de 256 délais entiers et fractionnaires de 20 à 143. En premier temps pour notre simulation, nous choisissons les 256 valeurs possibles de la période, avec 124 valeurs correspondantes à des délais entiers de 20 à 143 et 132 valeurs fractionnaires de $(19\frac{1}{3})$ jusqu'à $(84\frac{2}{3})$.

L'estimation de la période et le gain est toujours basée sur la minimisation de l'erreur quadratique moyenne ce qui revient à maximiser le terme ε_d comme dans (4.20). La période qui rend ε_d maximum est sélectionnée avec son gain associé.

Approche pour accélérer la recherche

La recherche de la période peut se faire sur l'ensemble des 256 valeurs possibles, mais dans ce cas beaucoup de calculs seront faits inutilement. Pour remédier à cet inconvénient, il faut réduire le nombre de délais à examiner. Pour cela nous allons procéder en deux étapes. La première étape consiste à rechercher la période dans la partie comportant les délais entiers seulement (une valeur parmi 124). La seconde étape consiste à chercher quelques valeurs dans la partie fractionnaire.. Selon le meilleur délai entier trouvé, la recherche se fera aux fractions se trouvant au voisinage des nombres qui peuvent être les valeurs réelles de la période. Si l'entier trouvé est $d = 58$ qui peut être un double, on cherche donc autour de 29 et 58 en ajoutant les fractions suivants $\left(-\frac{2}{3}, -\frac{1}{3}, 0, \frac{1}{3}, \frac{2}{3}\right)$.

On n'aura donc pas à examiner toutes les valeurs fractionnaires.

Pour mieux éclaircir la procédure, considérons les intervalles [20,39] [40,79] [80,143]. Selon la position du délai entier dans ces régions, nous pouvons en déduire les valeurs autours desquelles nous allons examiner et chercher la période fractionnaire qui se situe dans l'intervalle $\left[19\frac{1}{3}, 84\frac{2}{3}\right]$.

Le tableau 5.6 expose les différentes régions de recherche.

délais	**Résolution**	**Régions de recherche**	
		Délais entiers $[20, 143]$	**Délais fractionnaires** $\left[19\frac{1}{3}, 84\frac{2}{3}\right]$
$\left[19\frac{1}{3}, 84\frac{2}{3}\right]$	1/3	si $d \in [20, 39]$ $\rightarrow$	d
		si $d \in [40, 79]$ $\rightarrow$	$d, d/2$
$[85, 143]$	délais entiers seulement	si $d \in [80, 143]$ $\rightarrow$	$d, d/2, d/3$

Tableau 5.6 Les intervalles à examiner pour les délais entiers et fractionnaires

Ainsi, cette procédure accélère la recherche et réduit les calculs.

Le choix de distribuer et de concentrer les délais fractionnaires uniquement entre $\left[19\frac{1}{3}, 84\frac{2}{3}\right]$ est volontaire et il est dû au faite que le problème des multiples de la période se pose surtout pour les

locuteurs féminins où la fréquence est haute (atteignant 400 Hz). L'amélioration recherchée doit se faire dans cette région.

La figure 5.3 montre clairement l'amélioration apportée en comparant le filtre prédicteur d'ordre 1 avec celui utilisant un délai fractionnaire.

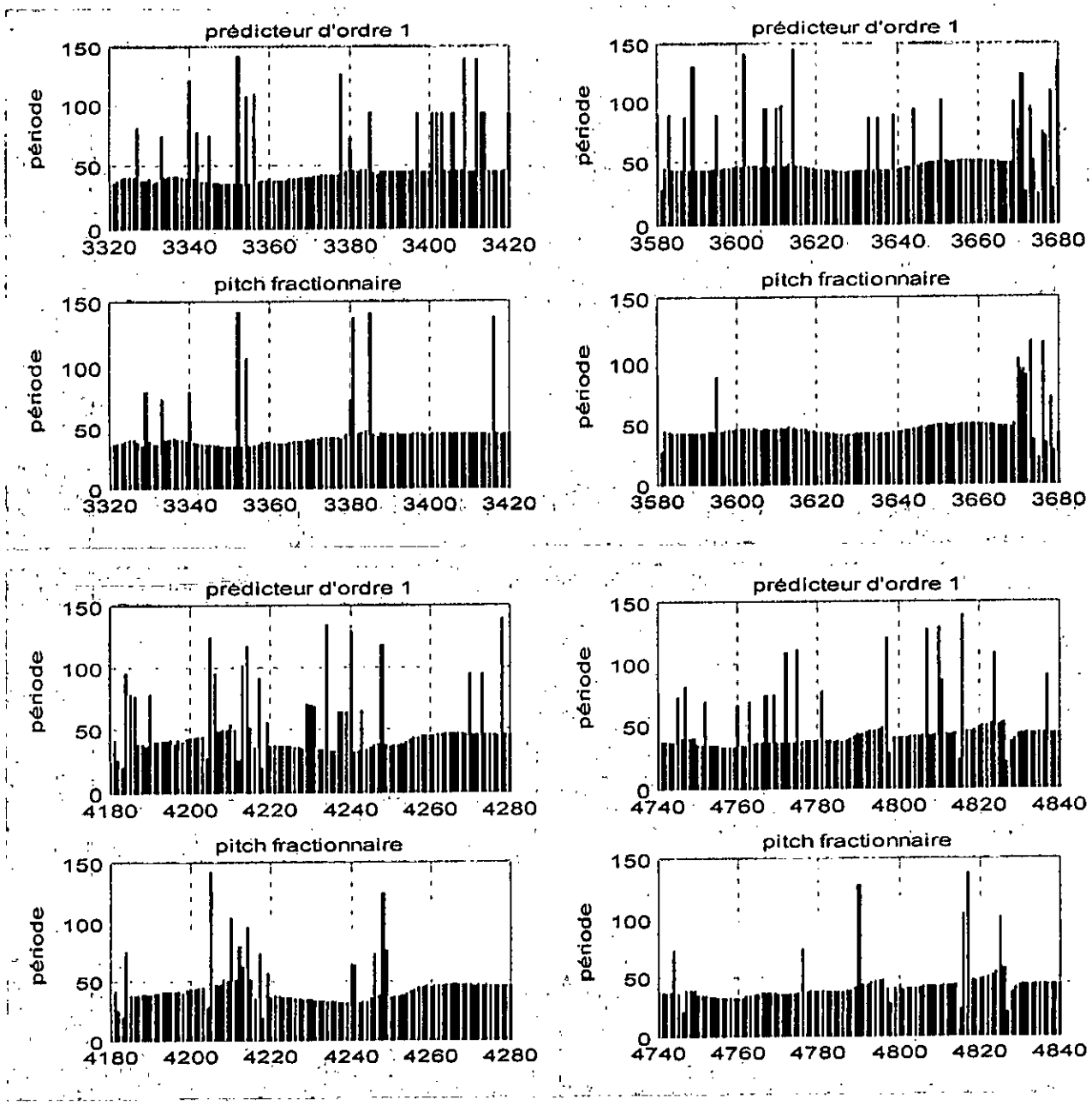


Figure 5.3 Les périodes trouvées dans quelques régions avec le prédicteur long-terme d'ordre 1 et le pitch fractionnaire.

		SNR_{seg}	
Les phrases tests	**Sexe**	**Prédicteur d'ordre 1**	**Pitch fractionnaire**
PHRASE 01	M	9.09	9.22
PHRASE 02	M	11.40	11.91
PHRASE 03	M	9.60	10.14
PHRASE 04	M	10.16	10.34
PHRASE 05	F	8.32	8.79
PHRASE 06	M	7.69	8.38
PHRASE 07	M	9.09	9.80
PHRASE 08	F	8.62	9.00
PHRASE 09	M	8.93	9.19
PHRASE 10	M	8.48	8.60
PHRASE 11	M	9.67	10.11
PHRASE 12	M	9.36	9.55
PHRASE 13	F	8.32	8.56
PHRASE 14	F	8.67	9.12
PHRASE 15	F	6.95	7.67
PHRASE 16	F	8.45	9.25
PHRASE 17	F	8.37	9.36
PHRASE 18	F	9.03	9.33
PHRASE 19	M	8.26	8.28
PHRASES 01 à 16		9.09	9.28

Tableau 5.7 Comparaison entre les rapports signal/bruit segmental calculés en utilisant le prédicteur long-terme d'ordre 1 et le pitch fractionnaire

Les sauts de période observés dans les figures sont moins nombreux lorsqu'on utilise un délai fractionnaire. Le calcul du SNR_{seg} pour plusieurs phrases tests confirme l'amélioration obtenue (tableau 5.7).

Contrairement au prédicteur long-terme d'ordre 3, l'amélioration apportée par le pitch fractionnaire ne coûte qu'un seule bit supplémentaire par sous-trame au codeur. La période du prédicteur long-terme est donc codée sur 8 bits. Ceci donne un avantage à la méthode du pitch fractionnaire. On peut songer à comparer les périodes optimales trouvées par ces deux méthodes. La figure 5.4 illustre cette comparaison et nous remarquons que parmi ces trois prédicteurs, le pitch fractionnaire est relativement le meilleur puisqu'il présente le minimum de sauts qui ne sont rien d'autres que les multiples des délais réels. Les mesures subjectifs confirment l'amélioration de la qualité des signaux de parole synthétiques.

Il faut aussi souligner qu'on peut enlever le bit supplémentaire nécessaire au codage de la période du nombre de bits alloué au dictionnaire.

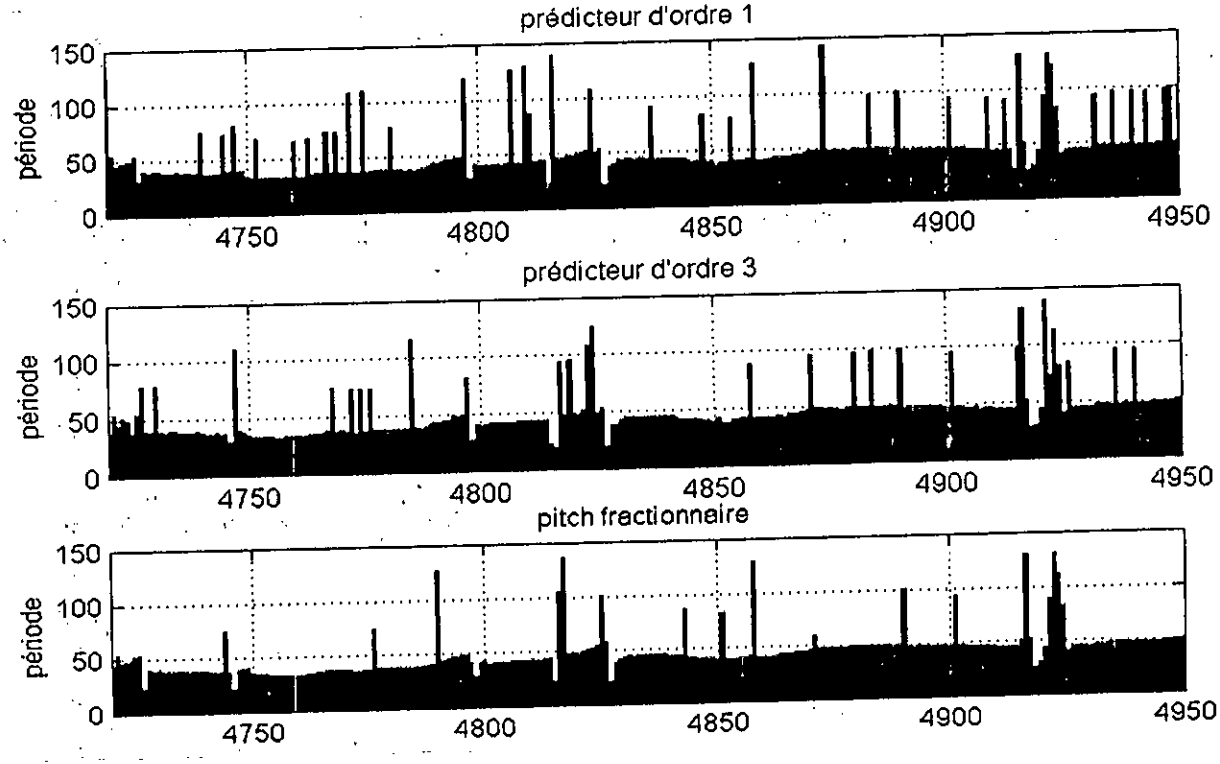


Figure 5.4 Evolution du délai optimal dans le prédicteur long-terme avec les prédicteur d'ordre 1 et 3 et le pitch fractionnaire.

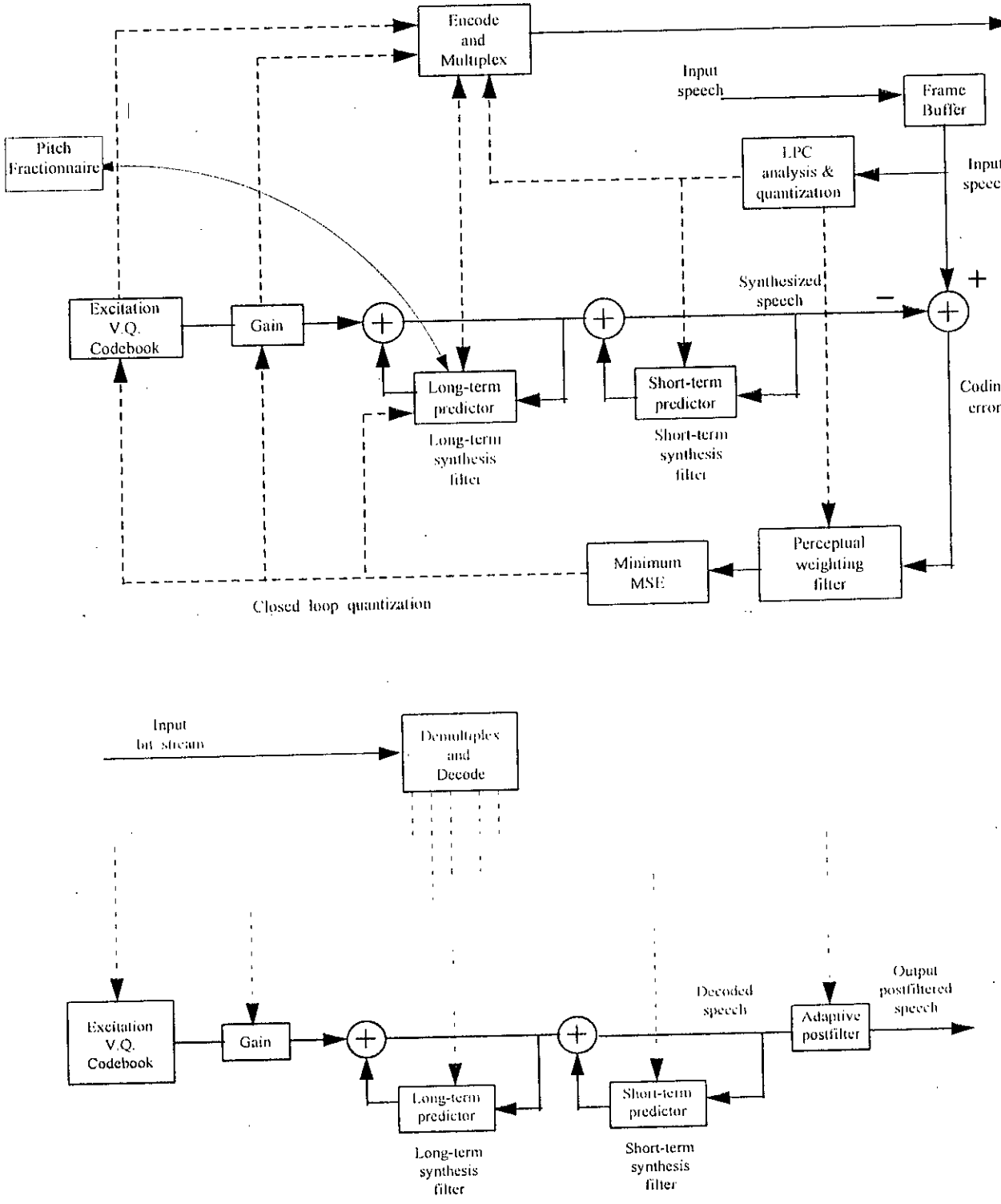


Figure 5.5 Schéma du codeur/décodeur CELP

Conclusions Générales

Le travail effectué dans ce mémoire a été réalisé en utilisant un codeur de parole de type CELP à 4.8 kbps. Ce type de codage comme nous l'avons déjà vu, met en œuvre la méthode d'analyse par synthèse du signal vocal en associant à la fois des techniques de prédiction linéaire et de quantification vectorielle. Deux types de prédicteurs sont présents dans la structure CELP ; le premier dit court-terme modélise le conduit vocal, le second, appelé long-terme, modélise le caractère périodique du signal d'excitation. Ils permettent de réduire les deux formes de redondance présentes dans le signal de parole pour aboutir à un signal d'erreur de prédiction qui, dans le cas du codeur CELP, est quantifié vectoriellement. Les performances des codeurs basés sur la prédiction linéaire et les techniques d'analyse par synthèse, sont fortement liés à la prédiction pitch. Les études ont montré que la mauvaise reproduction de la périodicité du signal d'excitation provoque une dégradation de la qualité de parole délivrée par le codeur surtout pour les locuteurs féminins.

Le but de ce mémoire était l'amélioration des performances des codeurs CELP par une meilleure reproduction de la périodicité. Notre étude s'est donc focalisée sur la modélisation de la périodicité.

Notre simulation basée sur un codeur CELP à 4.8 kbps a porté sur l'utilisation de prédicteurs long-terme avec des délais entiers et non entiers. La méthode utilisée doit tenir compte du compromis qualité/débit/complexité du codeur. Dans cette optique, l'utilisation de la méthode du pitch fractionnaire au niveau du prédicteur long-terme a permis d'obtenir des résultats très satisfaisants. En fait, les prédicteurs pitch avec les délais non entiers peuvent offrir de meilleures performances que les prédicteurs pitch d'ordre légèrement élevé (3 ou 5) et utilisant des délais entiers. Les mesures objectives ou subjectives réalisées révèlent une amélioration de la qualité de parole synthétique par rapport à la méthode ordinaire avec un prédicteur d'ordre 1 et ne comportant que des délais entiers seulement. Le fait de rechercher le pitch sur des délais entiers et ensuite

trouver le pitch fractionnaire autour des sous-multiples de la valeur entière diminue considérablement la complexité. De plus le codage du pitch fractionnaire nécessite un seul bit supplémentaire qui peut être retranché du dictionnaire d'excitation ce qui a pour effet de garder le même débit. La résolution temporelle du dictionnaire adaptatif a été augmentée pour les périodes se situant entre 19 et 85. Il en a résulté une meilleure estimation du pitch pour les hautes fréquences. Les tests d'écoute ont confirmé les résultats objectifs (SNR_{seg}).

Pour récapituler nous pouvons dire que l'utilisation de la méthode du pitch fractionnaire permet d'avoir une meilleure qualité de la parole et avec une légère augmentation de la complexité.

Ce mémoire n'est qu'une initiation qui nous a permis de nous familiariser avec les codeurs de type CELP, la quantification, la technique d'analyse par synthèse .....etc. Le monde du codage et de la compression de la parole est un vaste domaine qui nécessite des diverses connaissances. Nous souhaitons que ce mémoire puisse servir de référence pour tout chercheur désirant s'initier à ce domaine.

Annexe A

Techniques d'évaluation de la Qualité de la parole

A.1 Introduction

Le bruit introduit dans un système de codage du signal de parole peut provenir de diverses sources :

- La quantification : la transformation des échantillons en valeurs entières (nombre fini de niveaux du quantificateur) crée une dégradation irréversible du signal.
- La saturation : écrétage éventuel du signal d'entrée au cours de l'opération de quantification.
- Le système de codage lui-même : aussi cherche-t-on à réduire le bruit inhérent aux techniques employées grâce au codage efficace, et au principe du masquage spectral.

Pour estimer le niveau de qualité du signal délivré par un codeur, on dispose de deux types de critères: objectif et subjectif. Les uns s'appuient exclusivement sur des mesures physiques du signal, et les autres prennent en compte des avis humains, basés sur des écoutes. La qualité d'un signal est difficile à évaluer de manière systématique car nous ne connaissant pas encore de relations mathématiques susceptibles de simuler le système auditif humain. Aussi est-il préférable d'utiliser conjointement les critères objectif et subjectif.

A.2 Critères objectifs

Le critère objectif le plus couramment utilisé est le rapport signal sur bruit (SNR). Il est défini comme le rapport entre la puissance du signal original et la puissance du signal d'erreur. Ce dernier représente la différence entre le signal original $S[n]$ et le signal reconstruit $\hat{S}[n]$. L'unité utilisée est le décibel (dB). L'expression du RSB ou le SNR (Signal to Noise Ratio)en anglais est donnée par :

$$SNR = 10\log \frac{\sum_{n=1}^{n=N} S[n]^2}{\sum_{n=1}^{n=N} \left(\hat{S}[n] - S[n]\right)^2}$$

N représente le nombre d'échantillons considérés pour le calcul. Dans le cas où N est égal à la quantité d'échantillons représentant le signal complet, nous parlons de ***rapport signal/bruit global***. Il caractérise la qualité du signal reconstruit dans son ensemble.

Quand on détermine le RSB global, on calcul la puissance globale des deux signaux, original et bruit de codage, sans tenir compte de la répartition mutuelle des énergies dans le domaine du temps. De ce fait, on peut rencontrer des zones de faibles énergies dans lesquelles le bruit de codage est important par rapport au signal à coder. La faible valeur des énergies respectives des deux signaux n'affectent pas le rapport signal/bruit, cependant la dégradation du signal est audible. Pour pallier cet inconvénient, on défini le ***rapport signal/bruit segmental.*** Il est défini comme étant la moyenne arithmétique des rapports signal/bruit (en dB) calculés sur des intervalles de durée fixe. Il prend davantage en compte les zones de faible énergie du signal que le rapport S/B global. Il est défini par la relation :

$$SNR_{seg} = \frac{1}{L}\sum_{l=1}^{L} 10\log \frac{\sum_{n=1}^{n=N} S[n]^2}{\sum_{n=1}^{n=N} \left(\hat{S}[n] - S[n]\right)^2}$$

où L est la quantité d'intervalles pris en compte et N le nombre d'échantillons de chaque intervalle. Il ne faut surtout pas considéré ces mesures comme des critères suffisamment représentatifs de la qualité d'un codeur. Elles utilisent des relations mathématiques qui ne tiennent pas compte des

pondération, sous la forme d'un filtre de fonction de transfert W(z) avant le critère de minimisation comme le montre la figure 2.5.

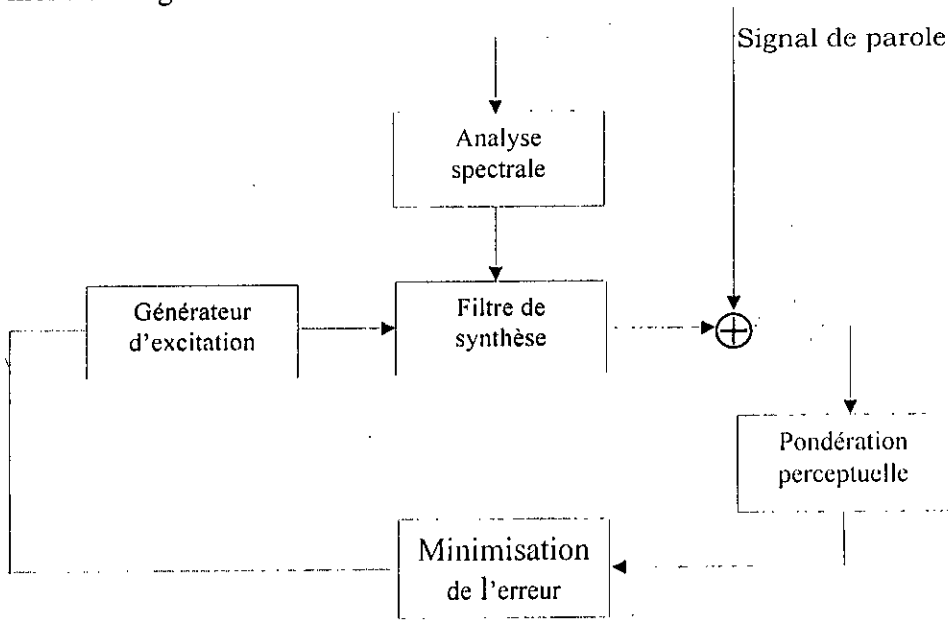


Figure 2.5 Introduction d'une fonction de pondération

Le bruit dû à la quantification est simplement moins perceptible lorsque le signal a beaucoup d'énergie. on dit que le signal masque le bruit. Il n'est pas possible de jouer sur la puissance totale du bruit. On cherche donc une fonction de pondération qui attribue moins d'importance aux zones fréquentielles énergétiques c'est à dire aux zones formantiques. La fonction de transfert W(z) joue ce rôle avec :

$$W(z) = \frac{A(z)}{A(z/\gamma)} \qquad 0<\gamma<1. \tag{2.31}$$

Ainsi la réponse fréquentielle de ce filtre varie au rythme du calcul des coefficients de prédiction à court-terme, il est donc adaptatif. En effet, si on note :

$$A(z) = 1 + a_1 z^{-1} + \ldots\ldots\ldots + a_p z^{-p} = \prod_{i=1}^{p} (1 - p_i z^{-1}). \tag{2.32}$$

où p_i spécifie la i^{eme} racine du polynôme A(z), on remarque que :

$$A(z/\gamma) = 1 + a_1 \gamma z^{-1} + \ldots\ldots\ldots + a_p \gamma^p z^{-p} = \prod_{i=1}^{p} (1 - \gamma\, p_i z^{-1}). \tag{2.33}$$

propriétés de l'audition humaine. Néanmoins des tests d'écoute ont ainsi montré qu'une amélioration du rapport signal/bruit entraînait nécessairement un accroissement de la qualité du signal reconstruit sur le plan de perception auditive. De toute manière ces mesures de RSB doivent être accompagnées par des techniques de mesure subjective de la qualité.

A.3 Critères subjectifs

Les essais d'écoute sont nécessaires car la qualité d'un système de codage de la parole ne vaut que par le jugement humain.

La qualité subjective d'un système de codage s'évalue principalement selon trois points :

- Diagnostic Rythme Test (DRT) qui mesure l'intelligibilité sur un grand nombre de mots.
- Diagnostic Acceptabilty Mesure (DAM) qui mesure le naturel perçu de la parole.
- Mean Opinion Score (MOS) où l'auditeur évalue un codeur sur une échelle absolue allant de 1 à 5 avec

5 Excellent

4 Bon

3 Passable

2 Médiocre

1 Mauvais.

Annexe B

Eléments de la Théorie de l'information pour le codage de la source

B.1 Entropie

Considérons une source à temps discret et ergodique $\{x(n)\}, n = 0, \pm 1, \pm 2, \ldots$ le flux des **Symboles** $x\{n\}$ forme une séquence aléatoire $\{X(n)\}, n = 0, \pm 1, \pm 2, \ldots$ dont les réalisations x_i appartiennent à l'**alphabet** de la source $A = \{x_i / i = 1, 2, \ldots, K\}$. Si K est fini la source est dite à **amplitude discrète**, si K est infini la source est dite à **amplitude continue**. La source est **sans mémoire** si les échantillons successifs sont statistiquement indépendants.

B.1.1 Source à amplitude discrète et sans mémoire

Soit p_i la probabilité d'occurrence du symbole x_i : $p_i = \Pr\{X(n) = x_i\}$. Une appréciation numérique de l'information propre de l'événement est alors donnée par : $I(x_i) = -\log_2 p_i$ [en bit/symbole] et l'information propre moyenne ou **entropie** (ou entropie d'ordre 0), qui représente la limite fondamentale pour représenter sans distorsion la source d'information, est [en bpp] :

$$H(X) = E(I(X)) = -\sum_{i=1}^{K} p_i . \log_2 p_i$$

Nous avons : $0 \le H(X) \le \log_2 K$, *si* $H(X) = 0$. La source est totalement **prédictible**, et si $H(X) = \log_2 K$, la source est **non prédictible** (les symboles sont équiprobables).

Log_2 K mesure la **capacité** de l'alphabet. les **redondances** de la source sont appréciées en calculant la différence : $(\log_2 K - H(X))$.

B.1.2 Source à amplitude discrète avec mémoire

Il existe alors des dépendances statistiques entre les échantillons successifs de la source. Soit $x = x(n) = (x(n), x(n+1),...,x(n+k-1))^T$, un vecteur constitué de k échantillons successifs de la source. Ce vecteur est caractérisé par sa probabilité conjointe $p_X(x)$ qui est indépendante du temps si nous considérons une source stationnaire, x est une réalisation du vecteur aléatoire $X=(X(n),X(n+1),...,X(n+k-1))^T$. Soit l'**entropie conjointe** des vecteurs aléatoires (ou entropie d'ordre k) [en bpp] :

$$H_k(X) = \frac{1}{k} E(-\log_2 p_X(X)) = -\frac{1}{k} \sum_x \sum \cdots \sum p_X(x) . \log_2 p_X(x)$$

et

$$H(X) = \lim_{k \to \infty} H_k(X)$$

nous considérons aussi l'**entropie conditionnelle** d'ordre k d'un symbole à l'instant n étant donnés les (k - 1) symboles précédents : $H(X(n)/X(n-1),X(n-2),...,X(n-k+1))$.

Il est démontré que :

$$H(X(n)/X(n-1),X(n-2),...,X(n-k+1)) \leq H_k(X)$$

et que :

$$\lim_{k \to \infty} H(X(n)/X(n-1),X(n-2),...,X(n-k+1)) = H(X)$$

les deux fonctions $H_k(X)$ et $H(X(n)/X(n-1),X(n-2),...,X(n-k+1))$ sont décroissantes avec k. de plus, pour deux source ayant deux alphabets identiques et même probabilité pour les symboles, il est prouvé que :

$$(H(X)/source\ avec\ mémoire) \leq (H(X)/source\ sans\ mémoire).$$

Ces résultats annoncent l'intérêt de la quantification vectorielle car pour atteindre le débit entropique il faut regrouper les échantillons de la source pour en exploiter la mémoire.

B.1.3 Codage sans perte d'une source à amplitude discrète

La théorie annonce qu'un codage sans perte d'information avec un débit binaire proche de l'entropie est réalisable pour une source à amplitude discrète, cependant elle n'explique pas comment construire ce code. Ce codage est dit **entropique** ou à **longueur de code variable** (ex : codage Huffman). Pour présenter de façon intuitive ce code, nous pouvons expliquer que les statistiques du signal à coder sont exploitées en affectant les mots de code les plus courts aux représentants les plus fréquents. Ce code est efficace (c-à-d le débit proche de l'entropie) si les symboles de la source ont des probabilités qui puissent être approchées par des puissances négatives de deux. Pour avoir un tel résultat, il est souvent nécessaire de regrouper les échantillons et de considérer les probabilités des vecteurs obtenus, les procédures de codage deviennent alors plus sophistiquées.

B.1.4 Source à amplitude continue et sans mémoire

Soit {x(n)} une source stationnaire et sans mémoire, et à moyenne nulle. Soient $p_X(x)$ sa fonction de densité de probabilité, $\sigma_X^2 = E(X^2(n))$ sa variance et $R_X = \sigma_x^2.\delta(u)$ sa fonction d'autocorrélation ($\delta(u)$ étant le symbole de Kronecker). Ce signal a une entropie absolue infinie (en effet $p_i = 0$ donc $H(X) = +\infty$), c'est pourquoi l'entropie différentielle est introduite (elle peut être positive, négative ou nulle en fonction de l'amplitude de la source) :

$$h(X) = E(-\log_2 p_X(X)) = -\int_{-\infty}^{+\infty} p_X(x).\log_2 p_X(x)\,dx$$

il est démontré que l'entropie différentielle est maximale si la source suit une loi gaussienne $N(0, \sigma_X^2)$, dans ce cas :

$$p_X(x) = (2.\pi.\sigma_X^2)^{-1}.\exp(\frac{-x^2}{2.\sigma_X^2}) \quad et \quad h(X)_G = \log_2\sqrt{2.\pi.e.\sigma_X^2}$$

il est aussi pratique de définir la puissance entropique qui traduit la répartition de l'information par unité de variance du signal :

$$P = \frac{1}{2.\pi.e} 2^{2.h(X)}$$

pour une source gaussienne $P = \sigma_X^2$, pour une non-gaussienne $P < \sigma_X^2$.

B.1.5 Source à amplitude continue avec mémoire

Nous considérons cette fois un vecteur x constitué de k échantillons successifs de la source, x est décrit par sa fonction de densité de probabilité conjointe $p_X(x)$ et l'entropie différentielle est définie par :

$$h(X) = \lim_{k \to \infty} h_k(X)$$

avec : $h_k(X) \quad \frac{1}{k} E(-\log_2 p_X(X)) \quad \frac{1}{k} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} p_X(x) . \log_2 p_X(x) \, dx$

de manière générale :

$$(h(X) / \textit{source avec mémoire}) < (h(X) / \textit{source sans mémoire}) \leq \frac{1}{2} . \log_2 (2.\pi.e.\sigma_X^2)$$

la borne supérieure est atteinte dans le cas d'une source gaussienne. Les entropies inférieures a cette valeur sont dues à :

1. source non gaussienne.
2. Source qui présente une mémoire (spectre de puissance n'est pas "plat" ou "blanc").

La puissance entropique d'une source gaussienne "colorée" (i.e avec mémoire) est : $P = \gamma_X^2 . \sigma_X^2$ ou γ_X^2 est la mesure d'étalement du spectre. Nous avons $0 \leq \gamma_X^2 \leq 1$, si $\gamma_X^2 = 1$ nous retrouvons le cas d'une source gaussienne sans mémoire, si $\gamma_X^2 < 1$ la source présente une mémoire (le signal est plus ou moins prédictible). L'inégalité suivante est alors prouvée :

$$(P / \textit{source avec mémoire}) < (P / \textit{source sans mémoire}) \leq \sigma_X^2$$

B.2 Fonction débit-distortion

B.2.1 Introduction

En pratique la source à coder est plus souvent à amplitude continue. Il est souhaité que le système codeur-décodeur assure la transmission de l'information avec un débit R [en bpp] adapté au canal pour une erreur moyenne de reconstruction D' minimale. En faisant varier R une courbe $D'(R)$ est obtenue. La théorie de l'information annonce qu'il existe une fonction débit-distortion

$D(R)$qui fournit une borne aux performances du système de codage. Cette fonction $D(R)$ indique la distorsion minimale théorique pour un codeur avec un débit R: $D(R) \leq D'(R)$

Dans le cas d'une source à amplitude discrète (pour laquelle une transmission sans erreur est possible), la courbe $D'(R)$ est souvent utilisée. Là encore, la théorie fournie la courbe distorsion-débit $R(D)$qui est l'inverse de la fonction $D(R)$et telle que : $R'(D) \geq R(D)$.

B.2.2 Source à amplitude discrète

Un codage entropique est donc réalisable. Le débit minimum pour transmettre, sans perte, l'information est : $\min\{R\} = R(0) = H(X)$. $R'(0) = h(X)$ est atteint à l'aide d'un code entropique ou à longueur variable. Si la source est avec mémoire le calcul du code est plus aisé. La figure B.1 donne un exemple d'une telle courbe *D(R)* qui est donc une fonction monotone décroissante avec R. la source étant de variance finie, alors $D(0) = \sigma_X^2$. En effet nous pouvons considérer que, pour R = 0 le codeur n'émet que des 0, les erreurs de reconstruction au décodeur sont alors égales aux échantillons des vecteurs source et la variance de l'erreur équivaut à celle de la source.

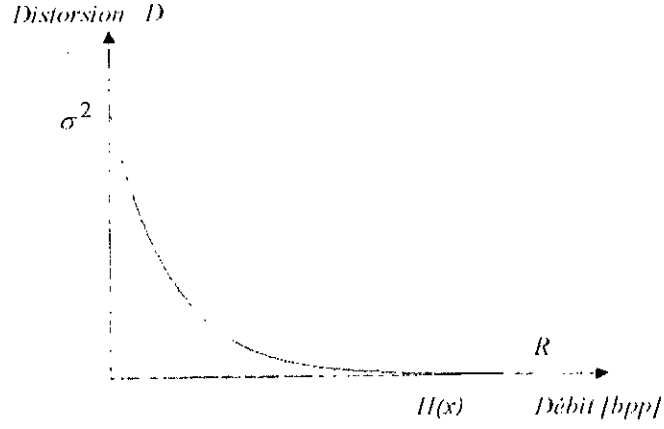


Figure B.1 Exemple de la courbe débit-distorsion d'une source à amplitude discrète

B.2.3 Source à amplitude continue

Distorsion

Nous considérons $\{x(n)\}, n = 0, \pm 1, \pm 2, \ldots$ une source stationnaire, à amplitude continue. Soit x un vecteur de k échantillons de la source, x est une réalisation du vecteur aléatoire X, ce dernier est caractérisé par sa fonction de densité de probabilité conjointe $p_X(x)$. les y sont les vecteurs de reproduction qui arrivent au récepteur, En général le vecteur reproduction y, est différent du x émis. Nous choisissons donc d'apprécier la distorsion moyenne par symbole à l'aide d'une mesure d'erreur quadratique moyenne :

$$E[d(X,Y)] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\ldots\int_{-\infty}^{\infty+} d^2(x,y).p_{XY}(x,y)\,dx.dy$$

ou $d(x,y) = p(y/x) = \frac{1}{N}\sum_{k=1}^{N}(x(k)-y(k))^2$ est la distance euclidienne.

Nous avons $P_{XY}(x,y) = P_{Y/X}(y/x)P_X(x)$. la distorsion moyenne, sujette à une contrainte sur le débit, dépend donc de la statistique de la source $p_X(x)$ et de la probabilité des transitions $P_{Y/X}(y/x)$. pour la minimiser il faut choisir une modélisation appropriée entre la source et les vecteurs de reconstruction.

Information mutuelle

Le calcul de la courbe D(R) repose sur le concept d'information mutuelle [par symbole] I(X,Y) qui est la mesure capable de décrire le flux d'information entre le codeur et le décodeur. Plus précisément elle apprécie la quantité moyenne d'information qu'implique la "réception" des valeurs par rapport a ceux émis.

$$I(X;Y) = \lim_{k\to\infty} I_k(X;Y)$$

avec :

$$I_k(X;Y) = \frac{1}{k}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\ldots\int_{-\infty}^{+\infty} p_{XY}(x,y).\log_2(p_{Y/X}(y/x).p_X(x))\,dx.dy$$

$$= \frac{1}{k}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\ldots\int_{-\infty}^{+\infty} p_{XY}(x,y).\log_2(p_{X/Y}(x/y).p_Y(y))\,dx.dy$$

I(X;Y) dépend donc également de la statistique de la source et de celle des transitions. L'adjectif "mutuel" vient de l'égalité : $I(X;Y) = I(Y;X)$. l'information mutuelle calcule le débit minimal R nécessaire pour avoir une fidélité de reconstruction D. en effet pour une densité $p_X(x)$ donnée considérons S, l'ensemble de tous les schémas de codage ayant une fonction de densité de probabilité de transition pour laquelle l'information mutuelle par symbole est inférieure à un débit donné :

$$S = \{P_{Y/X}(y/x) : I_k(X;Y) \leq R\}$$

Chacun de ces schémas réalise une erreur moyenne E[d(X,Y)]. Nous recherchons alors celui assurant le minimum de distorsion :

$$D_k(R) = \min_{P_{Y/X}(y/x) \in S} E[d(X,Y)]$$

le codeur réalisant $D_k(R)$ est optimal, il assure la distorsion moyenne minimale sous une contrainte de débit inférieur à R. nous remarquons qu'à ce stade $D_k(R)$ est une fonction monotone décroissante avec R. la théorie montre que ce schéma de codage optimal doit être tel que ses vecteurs source soient statistiquement indépendants. Alors **la fonction débit-distorsion** est définie par :

$$D(R) = \lim_{k \to \infty} D_k(R).$$

La fonction D(R) fournit pour un débit donné R, une borne minimale de la distorsion de tous codeurs (un exemple d'une telle fonction est donné à la figure B.2). En pratique aucun schéma de codage ne peut atteindre une telle performance. Cependant ces équations annoncent que de meilleurs résultats seront obtenues en utilisant des quantificateurs vectoriels.

La courbe inverse de $D(R)$ est $R(D)$. elle correspond au débit minimal nécessaire pour que le signal reconstruit au récepteur ait une distorsion D. dans le cas d'une source à amplitude continue $R(0) = +\infty$. Pour D = 0, $P_{Y/X}(y/x)$ = (1 si y = x ; 0 sinon) et $H(X) = H(Y) = I(X;Y) = +\infty$. Dans ce cas il n'y a donc pas de codage, le cas étudié est plutôt $D \neq 0$ et $I(X;Y) < +\infty$. Le codage de la source est donc une opération irréversible causant une distorsion D pour un débit R.

Source sans mémoire : cas d'une source gaussienne

Nous considérons la source obéissant à la loi normale $N(0, \sigma_X^2)$ alors :

$$R(D)_G = \max\left\{0, \frac{1}{2}\log_2 \frac{\sigma_X^2}{D}\right\} = \begin{cases} \frac{1}{2}\log_2 \frac{\sigma_X^2}{D} & si\ 0 \le D \le \sigma_X^2 \\ 0 & si\ D \ge \sigma_X^2 \end{cases}$$

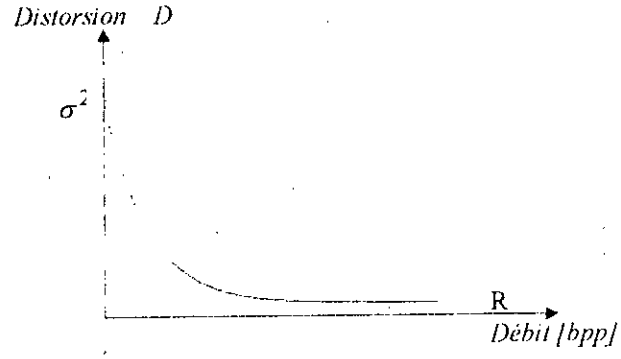


Figure B.2 Exemple de la courbe débit –distorsion d'une source à amplitude continue

et $$D(R)_G = 2^{-2.R}.\sigma_X^2.$$

Comme dans le cas de la source à amplitude discrète, il est évident que pour avoir $D \quad \sigma_X^2$ il n'y a pas d'information à transmettre (e.g. tous les vecteurs de reproduction ont leurs composantes nulles). Nous remarquons que l'erreur quadratique moyenne est réduite d'un facteur quatre (augmentation du SNR de 6.02 dB) pour chaque bit ajouté à la transmission.

Il est intéressant de modéliser les liens entrée / sortie du système global de codage afin d'en connaître la configuration optimale et d'analyser les effet de la quantification.

Ainsi le système réalisant $R(D)_G$ peut être décomposé en :

$$p_{Y/X}(y/x) = \frac{1}{\sqrt{2.\pi.\beta.D}}.\exp\left(\frac{-(y-\beta.x)^2}{2.\beta.D}\right) \quad avec \quad \beta = 1 - \frac{D}{\sigma_X^2}.$$

La sortie obéit à une loi normale $N(\beta.x, \beta.D)$. Les erreurs de quantification suivent donc également une loi gaussienne et elles sont indépendantes vis à vis à l'entrée.

enfin : $$R(D)_G = \frac{1}{2}.(\log_2 \sigma_X^2 - \log_2 D)$$

le débit nécessaire pour reproduire la source avec la distorsion D est la différence en entropie entre la source et le bruit de quantification qui sont deux variable aléatoires normales de variances σ_X^2 et D.

Source sans mémoire : cas d'une source non-gaussienne

Il n'existe pas de fonction débit-distorsion explicite mais des bornes de la forme :

$$^{L}D(R) < D(R) < D(R)_G$$

la borne supérieure est la fonction $D(R)_G$ correspondant à la source gaussienne sans mémoire. La borne inférieure correspond à la borne de Shannon :

$$^{L}D(R) = \frac{1}{2.\pi.e}.2^{-2.(R-h(X))} = 2^{(-2.R.P)} \quad ou \quad ^{L}R(D) = h(X) - \frac{1}{2}.\log_2(2.\pi.e.D).$$

La puissance entropique P correspond pour une source non-gaussienne à la variance de la gaussienne qui aurait la même entropie différentielle. Pour une gaussienne $P = \sigma_X^2$ et on trouve $D(R) = {}^{L}D(R)$. En pratique, pour une large classe de distributions et à haut débit, $^{L}D(R)$ tend vers la fonction débit-distorsion du système $D(R)$. Pour les débit inférieurs (de 1 à 3 bits/échantillon), $^{L}D(R)$ est une borne trop optimiste, $D(R)$ est alors calculée numériquement via l'algorithme de Blahut.

Source avec mémoire : cas d'une source gaussienne

La théorie annonce qu'une plus grande compression est possible pour les sources non-gaussienne avec mémoire. Cependant pour atteindre un tel résultat (traduit par la courbe $D(R)$), le système de codage nécessite plus d'information sur la source que celle uniquement fournit par son spectre de puissance $(S_X(e^{j.w}))$ et sa fonction d'autocorrélation.

La présentation théorique qui suit est introduite afin de mieux appréhender les problèmes liés à la quantification (la localisation des erreurs introduites par le codage).

La fonction $D(R)$ d'une source gaussienne colorée (c.a.d avec mémoire, par opposition à la source sans mémoire dont le spectre de puissance constant est dit blanc) est donnée sous une forme paramétrique (le paramètre est ϕ) :

Le module de la réponse en fréquence du filtre $1/A(z/\gamma)$ présente des pics moins accentués que celui du filtre $1/A(z)$ puisque les pôles du filtre $1/A(z/\gamma)$ sont ramenés vers le centre du cercle unité par rapport à ceux du filtre $1/A(z)$. Le module de la réponse en fréquence du filtre de pondération $W(z) = A(z)/A(z/\gamma)$ a donc la forme souhaitée.

Le choix de la valeur numérique du facteur perceptuel γ permet de moduler la fonction de pondération à sa convenance. Pour $\gamma = 1$, tout se passe comme si on n'utilisait pas de fonction de pondération; on effectue une modélisation du signal original. Pour $\gamma = 0$, on réalise une modélisation du signal résiduel. La valeur généralement choisie (d'après les tests d'écoute) est $\gamma = 0.8$.

Le signal d'excitation peut prendre différentes formes dans les systèmes LPAS. On peut citer à titre d'exemples les codages MPE (multi-pulse excitation), et RPE (regular-pulse excitation) qui produisent un signal parole de bonne qualité d'écoute avec une complexité raisonnable à un débit autour de 10 kbps. Dans le codage MPE, l'excitation est représentée comme une suite d'impulsions. On remplace l'excitation exacte par une excitation formée d'impulsions isolées, dont la position et l'amplitude sont calculées de manière à générer un signal synthétique aussi proche que possible de l'original. Le codage RPE utilise la même idée mais les espacements entre les impulsions sont fixés. On détermine seulement la position de la première impulsion et les amplitudes.

Dans le codage CELP, le codeur utilise la quantification vectorielle pour coder le résiduel. Les divers vecteurs d'excitation constituent un dictionnaire. Chaque vecteur est filtré par un filtre long-terme et un filtre court-terme (filtre LPC) pour produire un signal de parole synthétique. On choisit de garder l'excitation qui produit le signal de parole synthétique le plus proche du signal original au sens d'un critère perceptuel. Dans le cas où le dictionnaire contient $2^{10} = 1024$ vecteurs de longueur 40, chaque excitation peut être codée sur 10 bits soit une quantification de 10/40=0.25 bits/échantillon.

2.4 Codeur/décodeur CELP

Le dictionnaire peut être stochastique (valeurs aléatoires) ou algébrique (valeurs binaires ou ternaires,...).

$$D(\phi)_G = \frac{1}{2.\pi}.\int_{-\pi}^{+\pi} \min\left\{\phi, S_X(e^{j.w}\right\} dw$$

$$D(\phi)_G = \frac{1}{2.\pi}.\int_{-\pi}^{+\pi} \min\left\{0, \frac{1}{2}\log_2 \frac{S_X(e^{j.w})}{\phi}\right\} dw$$

Cas de petites distorsions

Nous parlons de "petites distorsions" lorsque $\phi \leq \min_w\left\{S_X(e^{jw})\right\}$ (c'est le cas avec l'hypothèse haute résolution). Alors une forme simple de la fonction débit-distorsion est obtenue avec : $D(R)_G = \gamma_X^2.2^{-2.R}.\sigma_X^2$ ou γ_X^2 est la mesure d'étalement du spectre S_X :

$$\gamma_X^2 = \frac{\exp(\frac{1}{2}.\int_{-\pi}^{+\pi}\log_e S_X(e^{j.w})dw)}{\sigma_X^2}$$

pour un débit donné, nous avons :

$$(D(R)_G \text{ / source avec mémoire}) = \gamma_X^2.(D(R)_G \text{ / source sans mémoire})$$

en exploitant la mémoire la distorsion peut donc être réduite d'un facteur γ_X^2 et ceci dans la zone de "petites distorsions".

Fonction débit-distorsion en considérant des vecteurs de taille k

Nous avons déjà introduit $D_K(R)$, la fonction débit distorsion correspond à une source constituée de blocs de taille k, ces blocs étant indépendants statistiquement entre eux. Dans le cas d'une telle source gaussienne nous avons toujours : $D(R) = \lim_{k\to\infty} D_k(R)$ et il est montré que:

$$D_k(R) = \frac{1}{k}\sum_{u=0}^{k-1}\min\left\{\phi, \lambda_u\right\} \quad et \quad R_k(\phi) = \frac{1}{k}\sum_{u=0}^{k-1}\max\left\{0, \frac{1}{2}.\log_2\frac{\lambda_u}{\phi}\right\}$$

λ_u étant la u° valeur propre de la matrice d'autocorrélation Γx d'ordre k du processus {X(n)}. Nous remarquons que R_k apparaît comme la moyenne de k débits $R_u = \max\left\{0, \frac{1}{2}.\log_2\frac{\lambda_u}{\phi}\right\}$ (ou chaque R_u résulte du codage de sources gaussiennes sans mémoire de variance λ_u) et que D_k apparaît aussi comme la moyenne de k distorsions optimales $D_u = \min\left\{\phi, \lambda_u\right\}$. Toutes les variables

aléatoires dont les variances sont supérieures au paramètre ϕ contribuent de la même façon à la distorsion globale du système. Ces variables qui n'apportent aucune information n'ont donc pas besoin d'être transmise, alors $R_u = 0$ pour $\phi < \lambda_u$

Fonction débit-distorsion en considérant des vecteurs de taille k et de petites distorsions

C'est le cas si $\phi \leq min.\ \{\lambda_u\}$ avec $u=0, 1, \ldots, k-1$, alors $D_k(R)=\phi$ et $D_k(R)=2^{-2R}\left(\prod_{u=0}^{k-1}\lambda_u\right)^{\frac{1}{k}}$

finalement nous obtenons :

$$D_k(R) \leq (D(R)_G \text{ / sans mémoire}) = 2^{-2.R}.\sigma_X^2$$

car $\sigma_X^2 = \frac{1}{k}\sum_{u=0}^{k-1}\lambda_u \geq (\prod_{u=0}^{k-1}\lambda_u)^{\frac{1}{k}}$, il y a égalité si et seulement si les λ_u sont tous égaux (la source est alors blanche).

Les sources avec mémoire peuvent donc être transmises avec des distorsions inférieures aux sources sans mémoire).

En utilisant le résultat connu (où det (Γx(k) est le déterminant de la matrice d'autocorrélation à l'ordre k): $\det(\Gamma_X(k))=\prod_{u=0}^{k-1}\lambda_u$. Nous pouvons définir la puissance entropique par $P_K = (\det(\Gamma x(k)))^{1/k}$. nous obtenons alors :

$$D_k(R) = 2^{-2.R}\,P_k = {}^{L}D(R)$$

la fonction débit-distorsion est donc égale à la borne de Shannon pour les petites distorsions.

Les théorèmes précédent sont aussi vrais si l'on considère des vecteurs successifs qui ne sont pas indépendants, alors il faut prendre une valeur élevée de k.

Source avec mémoire : cas d'une source non gaussienne

Nous retrouvons :

$${}^{L}D(R) \leq D(R) \leq D(R)_G$$

avec ${}^{L}D(R)$ la borne de Shannon. Là encore, en considérant un second moment fixé et la métrique euclidienne, une source gaussienne est moins compressible qu'une non gaussienne.

Bibliographie

[1] R.P.Ramachandran, P.Kabal, "Pitch prediction filters in speech coding," *IEEE, trans Acoustics, Speech, Signal Processing*, vol 37, pp. 467-477, April 1989.

[2] Q.Yasheng, P.Kabal, "Pseudo-three-tap pitch prediction filters," *Proc IEEE trans Acoustics, Speech, Signal Processing*, pp. 523-526, 1993

[3] R.Boite et M.Kunt, "Traitement de la parole, "*presses polytechniques Romandes* 1987.

[4] M.Mauc, "Réduction de la complexité des algorithmes de codage de la parole de type CELP," Thèse de doctorat de l'université de PARIS-XI, Novembre 1993.

[5] J.Makhoul, S.Roucos, "Vector quantization in speech coding," *Proceedings of the IEEE*, vol 73, N°11, November 1985

[6] B.S.Atal, S.L.Hanauer, "Speech analysis and synthesis by linear prediction of the Wave," *J. Acoustical Society of America*, vol. 50, pp. 637-655, Aug. 1971

[7] C.Papacostantinou, "Improved pitch modelling for low bit rate speech coders," Thèse Master, Université de McGill, Montreal, Canada, Août 1997.

[8] N.Moreau, "Techniques de compression des signaux," *collection technique et scientifique des télécommunications* MASSON 1995.

[9] E.Panos.Papamichalis, "Practical approaches to speech coding ," *Prentice-Hall, Inc* Englewood Cliffs, New Jersey 1987.

[10] A.Gersho and R.M.Gray, "Vector quantization and signal compression," *Kluwer Academic Publishers*, 1992.

[11] V.Ricordel, C.Labit, "Etude de schémas de quantification vectorielle algébrique et arborescente. Application à la compression de séquence d'image numérique. " Thèse de doctorat de l'université de Rennes I Décembre 1996

[12] V.Ricordel, C.Labit, "Quantification vectorielle par emboîtage d'une hiérarchie de réseaux réguliers de points," *I.R.I.S.A, publication interne* N° 945, juillet 1995

[13] R.M.Gray, "Vector quantization," *IEEE ASSP Magazine*, pp 4-29, April 1984.

[14] J.S.Marques, I.M.Trancoso, J.M.Tribolet, L.B.Almeida "Improved pitch prediction with fractional delays in CELP coding," *Proc IEEE Int. Conf on Acoustics, Speech, Signal Processing* (Albuquerque), pp 665-668, Apr 1990.

[15] C.O'Neil , B.Murray, A.D.Fagan ,"An efficient algorithm for pitch prediction using fractional delays," *Signal processing Theories and Applications, Elsevier Science Publishers B.V.* 1992

[16] P.Kroon, B.S.Atal, "Pitch predictors with high temporal resolution, " *Proc IEEE, Int., Conf on Acoustics, Speech and Signal Processing* (Albuquerque), pp. 661-664, Apr. 1990.

[17] P.Kroon, B.S.Atal, " On Improving the performance of Pitch predictors in speech coding systems, " *Advances in speech coding*, Edition. V.Cuperman, A.Gersho, B.S.Atal. pp 321.

[18] M.Yong, A.Gersho, "Efficient encoding of the long-term predictor in vector excitation coders, " *Advances in speech coding*, Edition. V.Cuperman, A.Gersho, B.S.Atal. pp 329.

[19] Yu-Hung Kao, "Low complexity CELP speech coding at 4,8 kbps ," Master of Science, University of Maryland, 1990.

[20] M.Ouled-cheikh, "Conception et realisation d'un codeur/décodeur de la parole à large bande (13 kbits/s), Thèse de magister, Département d'Electronique, ENP Alger Juin 1999.

[21] B.Pucci, "Compression numérique de la parole au moyen de codeurs de type CELP," thèse de doctorat, université de Nice-Sophia Antipolis, Juin 1992.

[22] R.M.Gray, D.L.Neuhoff, "Quantization, "*IEEE Transactions on Information Theory* vol 44, N°6, Octobre 1998.

[23] M.Bellanger, "Traitement numérique du signal, " Collection technique et scientifique des télécommunications MASSON 1987.

[24] Site internet.

L'étape du décodage peut se décomposer en quatre étapes :

- Sélection de la séquence d'excitation à partir de l'index k_0.
- Ajustement en énergie par multiplication du gain associé G_0.
- Convolution du signal par le filtre 1/B(z), ce qui a pour effet d'introduire la composante périodique.
- Convolution par le filtre 1/A(z), ce qui a pour effet de reconstituer l'enveloppe formantique.

Pour chaque fenêtre d'analyse, on transmet au récepteur les paramètres des filtres 1/A(z) et 1/B(z), l'index du vecteur candidat sélectionné et son gain associé.

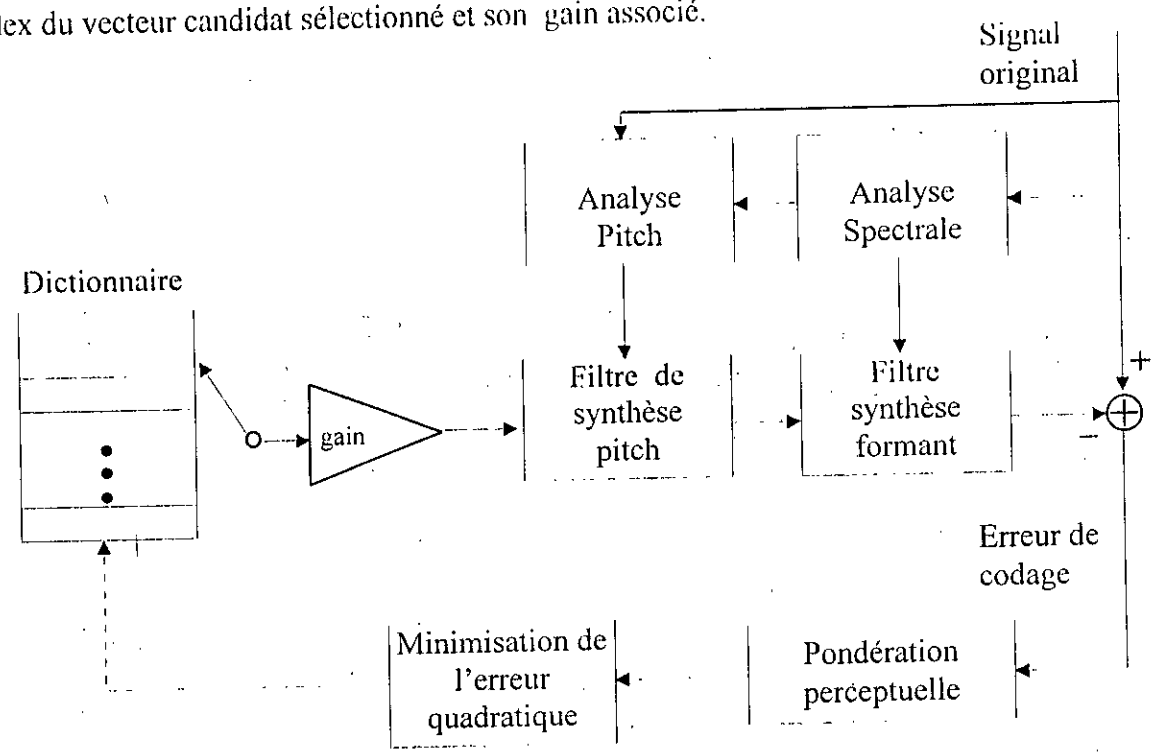


Figure 2.6 Modèle d'un codeur CELP.

Les paramètres du predicteur court-terme sont calculés et transmis à une cadence qui correspond à une trame d'échantillons, tandis que les autres paramètres qui sont l'index de l'excitation et son gain associé, la période pitch et son coefficient sont déterminés et envoyés à chaque sous-trame d'échantillons. [2]

[2] Généralement on divise la trame en quatre sous-trames, ce qui correspond à 7.5 ms pour une trame de 30 ms