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

ECOLE NATIONALE POLYTECHNIQUE

المدرسة الوطنية المتعددة التقنيات
Ecole Nationale Polytechnique

DEPARTEMENT D'ELECTRONIQUE

Laboratoire : Signal et Communications

MEMOIRE DE MAGISTER

Option : Signal et Communications

Présenté par : Mr. SADLI Ahmed

Ingénieur d'état en électronique (Université de BLIDA)

Thème :

Utilisation des méthodes d'interpolation pour la compression des signaux pseudo-périodiques

Devant le jury composé de

Présidente du jury :
Mme L.HAMAMI Professeur à L'ENP
Rapporteur :
Mr D. BERKANI Professeur à L'ENP
Examinateurs :
Mlle M.GUERTI Professeur à L'ENP
Mr M.S. AIT CHEIKH Maitre de Conférences à L'ENP

Décembre 2011

Remerciements

Je tiens à remercier et exprimer ma sincère reconnaissance, premièrement au professeur Daoud Berkani, qui m'a proposé ce projet et qui m'a accepté dans son groupe. Je le remercie aussi de m'avoir guidé dans mes recherches par sa rigueur et sa motivation.

Je remercie également les membres de jury, Mme L.HAMAMI, Pr M.GUERTI, Mr M.S.AIT-CHEIKH d'avoir accepté de juger ce travail.

Un grand remerciement à tous mes professeurs. A l'Université de BLIDA, et à L'ECOLE POLYTECHNIQUE, pour la formation solide qu'ils nous ont procurée pendant les années d'études.

Je remercie aussi mes parents ainsi que toute ma famille qui ont, durant toutes ces années, toujours encouragé mon goût de l'aventure et sans qui tout cela n'aurait sans aucun doute pas été possible.

Je remercie mes amis et tous ceux qui ont contribué à ce travail de prés ou de loin.

الملخص:

إن محدودية سرعة أجهزة الإعلام الآلي ،خاصة قنوات الإتصال تستوجب عملا مكثفا لتطوير برامج ضغط المعطيات والبيانات .يعتبر WI (Waveform Interpolation) برنامج التشفير الذي يعتمد على مبدأ إستكمال شكل الموجة من أنجع الطرق الحديثة المطبقة على الإشارة الصوتية، حيث يعمل بتدفق منخفض ويعطي بذلك نوعية صوت جيدة.

عملنا يتمثل في بسط ما وصل إليه هذا المشفر من تطور فيما يخص تكميم المركبة REW(Rapidly Evolving Waveform)، ومن ثم إضافة تحسين يتمثل في التخلص من التكرار الكامن في المركبة SEW (Slowly Evolving Waveform) أثناء الضغط، وإعادة خلقه بواسطة الإستكمال الخطي أثناء فك الضغط.

النتائج المحصل عليها ذات أهمية بالغة لأن الإجراء الجديد أنقص من التدفق محافظا على نوعية الإشارة الصوتية المضغوطة.

كلمات مفتاحية : إستكمال شكل الموجة (WI)، تشفير الكلام، المركبة المتغيرة ببطئ (SEW)، المركبة المتغيرة بسرعة (REW)، التكميم، الإشارة الصوتية.

Résumé :

La limitation de vitesse du matériel informatique, notamment les canaux de transmission requiert un travail intensif afin de développer des algorithmes de compression des données. Le codeur WI (Waveform Interpolation) qui s'articule sur le principe de l'interpolation de la forme d'onde, est considéré parmi les nouvelles méthodes les plus efficaces, qui sont appliquées sur les signaux vocaux. Ces méthodes fonctionnent avec un taux de compression réduit et donnent ainsi une meilleure qualité du son.

Notre travail consiste à détailler le développement apporté sur ce codage concernant la quantification de la composante REW (Rapidly Evolving Waveform), ainsi à apporter une amélioration par l'élimination de la redondance au niveau de la composante SEW (Slowly Evolving Waveform) lors de la compression, et de la recréer par l'interpolation linéaire durant la décompression.

En effet, les résultats donnés sont d'une grande importance, en révélant l'efficacité d'une nouvelle procédure qui réduit le taux de compression et préserve la qualité du signal vocal compressé.

Mot clés : interpolation de la forme d'onde (WI), codage de la parole, composante à évolution lente (SEW), composante à évolution rapide (REW), quantification, signal parole.

Abstract:

Computers speed limitation, especially channels of transmission requires an intensive work, in order to develop algorithms of data compression. The algorithm WI (Waveform Interpolation) which consists of the principle of the wave form interpolation is among the most effective modern methods used to compress the sound signals; it works on a ground of a low rate that provides a best quality of sound.

Our work consists of the development extension provided by this coder, especially the quantization of the formulation REW (Rapidly Evolving Waveform) and adding an improvement by the elimination of the redundancy from the formulation SEW (Slowly Evolving Waveform) during the compression and recreates it by the liner interpolation during the decompression.

In fact, the given results are very important, they reveal the effectiveness of the new method which reduces the flow compression and maintains the compressed voice sound quality.

Key words: Waveform Interpolation (WI), speech coding, Slowly Evolving Waveform (SEW), Rapidly Evolving Waveform (REW), quantization, speech signal.

Liste des Abréviations :

ACELP	Algebraic Code Excited Linear Predictive.
ADM	Adaptive Delta Modulation.
ADPCM	Adaptive Differential Pulse Code Modulation.
AR	Auto Regressive.
CELP	Code-Excited Linear Prediction.
CODEC	Encoder and Decoder.
CS-ACELP	Conjugate Structure Algebraic CELP.
DPCM	Differential PCM.
CW	Characteristic Waveform.
DCVQ	Dimension Conversion Vector Quantization.
DCT	Discrete Cosine Transform.
DFT	Discrete Fourier Transform.
DSP	Digital Signal Processing.
DTFS	Discrete Time Fourier Series.
EVRC	Enhanced Variable Rate Codec.
EWI	Enhanced Waveform Interpolation
FFT	Fast Fourier Transform.
FS	Federal Standard (U.S).
GLA	Generalized Lloyd Algorithm.
GSM	Global System for Mobile.
HSX	Harmonic Stochastic Excitation Coders.
IMBE	Improved Multi-Band Excitation.
IP	Internet Protocol.
ITU	International Telecommunication Union.
ITU-T	ITU - Telecommunication standardization sector.
LBG	Linde Buzo and Gray.
LBG-VQ	Linde Buzo and Gray -Vector Quantization.
LD-CELP	Low-Delay Code Excited Linear Prediction.
LP	Linear Prediction.
LPC	Linear Predictive Coding.
LSF	Line Spectral Frequency.
LSP	Line Spectral Pair.
LTP	Long Term Predictor.

MBE	Multi-Band Excitation.
MELP	Mixed Excitation Linear Prediction.
MIPS	Million Instructions Per Second.
MOS	Mean Opinion Score.
MSE	Mean Square Error.
MSVQ	Multi Stage Vector Quantization.
PCM	Pulse Code Modulation.
PESQ	Perceptual Evaluation of Speech Quality.
PWI	Prototype Waveform Interpolation.
QOS	Quality Of Service.
REW	Rapidly Evolving Waveform.
SEGSNR	Segmental SNR.
SEW	Slowly Evolving Waveform.
SNR	Signal to Noise Ratio (RSB).
SO 68	Standards For Service Options 68.
STC	Sinusoïdal Transform Coders.
SVQ	Split Vector Quantization.
TIMIT	Texas Instruments and Massachusetts Institute of Technology.
UIT-T	l'Union Internationale des Télécommunications.
UMTS	Universal Mobile Telecommunication System.
V/UV	Voiced/Unvoiced.
VBR	Variable Bit Rate.
VDVQ	Variable Dimension Vector Quantization.
VQ	Vector Quantization.
WI	Waveform Interpolation.

Liste des figures :

Liste des tableaux:

Table Des Matières

Chapitre 2 : Généralités sur la parole.

Chapitre 3 : Codeur WI.

Chapitre 4 : Compression des SEW avec sous échantillonnage et interpolation.

Introduction générale :

Le codage de la parole est, essentiellement, parmi les méthodes permettant l'obtention d'un usage plus efficace des réseaux de télécommunications numériques, en particulier les réseaux cellulaires, il permet aussi de réduire la mémoire nécessaire dans les systèmes de stockage de la parole. La volonté d'avoir une représentation numérique de la parole à faible débit, n'est pas souvent compatible avec la demande d'une reconstruction de la parole de haute qualité.

Le codage de la parole à des taux autour de 4 kbps est largement utilisé dans des applications comme, la téléphonie visuelle et les communications mobiles. Ce qui a conduit à développer un codeur de parole basé sur de nouvelles approches, contrairement à celles utilisées dans les codeurs classiques, avec comme objectif, une reconstruction fidèle de la parole à des débits inferieurs à 4 kbps.

Le défi d'actualité est de chercher des codeurs de parole qualifiés avec un taux autours de 4 kbps. C'est bien connu que la qualité de la parole à base d'algorithmes CELP (Code-Excited Linear Prediction), comme le G.729, se détériore rapidement pour des taux inférieurs à 4 kbps, par conséquent, il y a la nécessité d'une nouvelle génération de codeurs. Un des candidats les plus prometteurs dans la prochaine standardisation de l'ITU pour les 4 kbps est le codeur à interpolation de forme d'onde, abrégé WI pour (Waveform Interpolation).

Dans le codage WI, l'amélioration de la qualité est concentrée sur le codage efficace des segments de la parole, sans toutefois modifier le format de base du codeur. La parole est représentée par une somme des SEW (Slowly Evolving Waveforms) pour les ondes qui ont une évolution lente et REW (Rapidly Evolving Waveforms) pour les ondes qui ont une évolution rapide.

Comme l'efficacité du codage dans la WI se base essentiellement sur la méthode du codage de ces deux composantes, il est primordial d'avoir une technique efficace pour la quantification de ces deux dernières.

Le codeur WI marque une évolution très importante pendant ces deux dernières décennies, en développant des techniques de quantification efficace des paramètres de codeur WI, surtout ceux qui sont appliquées sur les composantes REW.

Notre travail consiste à utiliser la duale (décimation – interpolation) pour construire un codeur WI avec un taux autour de 3 kbps sans toucher à la qualité du signal.

Chapitre 1 : Compression des données

1.1. introduction

L'augmentation rapide du volume des données transmises ou stockées soulève le problème de la compression de ces données. La compression des données a d'importantes conséquences économiques comme :

- L'utilisation plus efficace des canaux de communication grâce à la compression de la bande de fréquence occupée par le signal.
- La réduction de la capacité des systèmes de stockage de données.

Dans ce chapitre nous allons présenter brièvement les notions de base de la théorie de l'information relatives au codage source et les techniques du codage de la parole. Le domaine est très vaste, nous ne tenterons pas de récapituler tous les détails. Nous allons nous borner aux utiles les plus répandues. On conclura le chapitre par les différents critères couramment utilisés pour juger et classer les méthodes du codage.

1.2. Systèmes de communication

La théorie des communications s'intéresse aux moyens de transmettre une information depuis une source jusqu'à un utilisateur (Figure 1.1). La nature de la source peut-être très variée. Il peut s'agir par exemple d'une voix, d'un signal électromagnétique ou d'une séquence de symboles binaires. Le canal peut être une ligne téléphonique, une liaison radio ou encore un support magnétique ou optique : bande magnétique ou disque compact. Le canal sera généralement perturbé par un bruit qui dépendra de l'environnement et de la nature du canal : perturbations électriques, rayures, ...

Dans les années 40, C. E. Shannon a développé une théorie mathématique appelée théorie de l'information qui décrit les aspects les plus fondamentaux des systèmes de communication [1]. Cette théorie s'intéresse à la construction et à l'étude des modèles mathématiques à l'aide essentiellement de la théorie des probabilités. Depuis ce premier exposé, la théorie de l'information s'est faite de plus en plus précise et est devenue aujourd'hui incontournable dans la conception de tout système de communication.

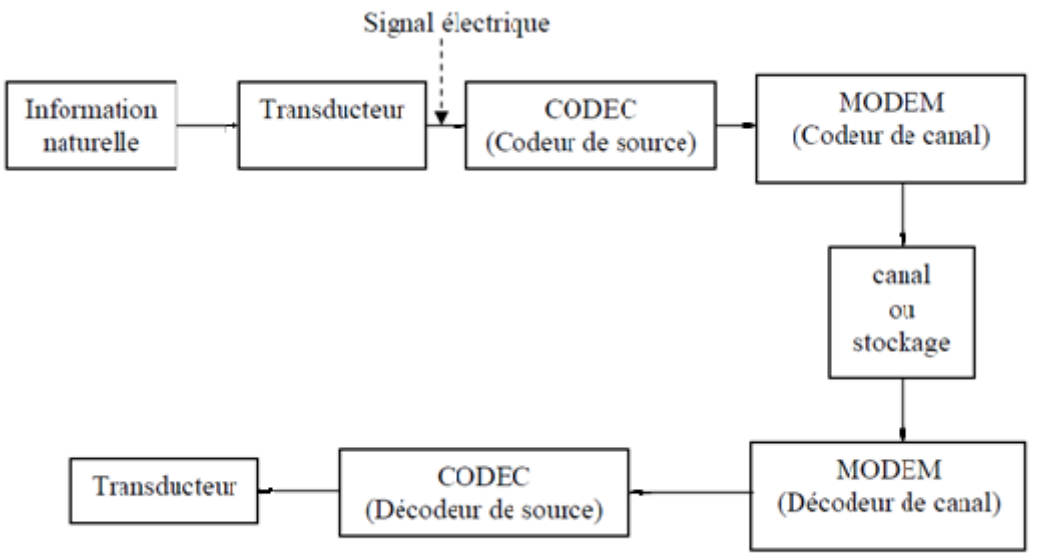

Fig.1.1 : Schéma d'un système de communications.

Le but du codeur de source est de représenter la sortie de la source, ou information, en une séquence binaire, et cela de la façon la plus économique possible. Le but du codeur du canal est l'adaptation de l'information au canal et assurer sa protection contre les bruits.

Shannon a formulé la théorie de l'information [1] il a démontré qu'il y a une limite fondamentale à la compression de données sans distorsion. Cette limite appelée entropie est dénotée par H, sa valeur exacte dépend de la nature statistique de la source. Il est possible de compresser la source sans distorsion, avec un taux de compression près de H.

Shannon a également développé la théorie de compression avec distorsion. Ceci est connu sous le nom de théorie de la distorsion [1]. Dans la compression de données avec distorsion, une certaine distorsion du signal original est tolérée durant la compression. Shannon a montré que pour une source donnée (dont toutes ses propriétés statistiques sont connues) et un taux de distorsion donné, il existe une fonction R(D) appelée fonction de distorsion. La théorie indique que si D est la distorsion tolérable, alors R(D) est le meilleur taux de compression.

1.3. Entropie

Il apparaît qu'il existe un lien entre l'information fournie par une source et la distribution de probabilité de la sortie de cette source. Plus l'évènement donné par la source est probable, moins la quantité d'information correspondante est grande. Plus précisément, si une lettre x_i a pour probabilité $p(x_i)$ d'être tirée, son information propre sera $I(x_i) = -\log_2 P(x_i)$. Cette définition paraît conforme à l'idée intuitive que l'on peut se faire de l'information, et en particulier

on a $I(x_i) = 0$ si $P(x_i) = 1$, c'est-à-dire que l'occurrence d'un évènement certain ne peut fournir aucune information.

La valeur moyenne de l'information propre calculée sur l'ensemble de l'alphabet revêt une grande importance. Elle est appelée *entropie* de la source et vaut :

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \qquad (1.1)$$

L'entropie d'une source est le nombre moyen minimal de symboles binaires par lettre nécessaires pour représenter la source.

Par exemple, si un alphabet contient 2^L lettres équiprobables, il est immédiat que l'entropie de la source correspondante vaut L. Or il est bien clair que pour représenter 2^L lettres distinctes, L symboles binaires sont nécessaires.

L'entropie d'une source est parfois donnée en bits par seconde (débit d'information), si l'entropie d'une source discrète est H, et si les lettres sont émises toutes les τ_s secondes, son entropie en bits/s sera H/τ_s.

L'entropie dans ce cas a la propriété suivante:

$H(x)$ est maximal si tous les symboles $\{x_i\ , x_i\ , \ldots\ldots\ldots,\ x_N\ \}$ de la source X sont équiprobables. Dans ce cas, l'entropie est égale à l'information associée à chaque message pris individuellement.

Alors l'inégalité suivante est évidente :

$$0 < H(X) \leq \log_2 N$$

1.4. Information mutuelle – Information propre

Nous considérons un espace probabilisé joint XY, avec $X\ = \{x_1, \ldots, x_N\}$ et $Y\ = \{y_1, \ldots, y_M\}$.

Les variables aléatoires x et y sont associées respectivement aux espaces X et Y.

1.4.1. Information propre

L'information propre de l'évènement $x\ =\ x_i$ est définie par

$$I(x_i)\ =\ -\log_2 P(x_i) \qquad (1.2)$$

L'information propre s'interprète comme la "quantité d'information fournie par la réalisation d'un événement".

Notons que l'information propre est toujours positive ou nulle, et que plus un évènement est improbable, plus son information propre est grande. À l'inverse, lorsque $P(x_i) = 1$, on a $I(x_i) = 0$, c'est-à-dire que la réalisation d'un évènement certain n'apporte aucune information, ce qui semble conforme à l'intuition.

1.4.2. Information mutuelle

L'information mutuelle entre les évènements $x = x_i$ *et* $y = y_j$ est définie par

$$I(x_i ; y_j) = \log_2 \frac{P(x_i | y_j)}{P(x_i)} \tag{1.3}$$

Remarquons que cette définiton est symétrique, en effet, on a par définition de la probabilité conditionnelle, $P(x_i ; y_j) = P(x_i | y_j)P(y_j) = P(x_i | y_j)P(x_i)$, et donc

$$I(x_i ; y_j) = I(y_j ; x_i) = \log_2 \frac{P(x_i ; y_j)}{P(x_i)P(y_j)} \tag{1.4}$$

1.4.3. L'information propre conditionnelle

On peut également définir dans l'espace probabilisé joint XY l'information propre conditionnelle qui est égale à la quantité d'information fournie par un évènement $x = x_i$ sachant que l'évènement $y = y_j$ s'est réalisé.

L'information propre conditionnelle de l'évènement $x = x_i$, sachant que $y = y_j$ est définie par

$$I(x_i \mid y_j) = -log_2 P(x_i \mid y_j) \tag{1.5}$$

Cette dernière définition nous permet de donner une nouvelle interprétation de l'information mutuelle entre deux évènements. En effet d'après la relation (1.3)

$$I(x; y) = I(x) - I(x \mid y) \tag{1.6}$$

1.5. Information mutuelle moyenne – Entropie

1.5.1. Information mutuelle moyenne

L'information mutuelle moyenne de *X* et *Y* dans l'espace probabilisé joint *XY* est définie par

$$I(X;\ Y\ ) = \sum_{i=1}^{N}\sum_{j=1}^{M} P(x_i\ ;\ yj) I(x_i\ ;\ yj)\ = \sum_{i=1}^{N}\sum_{j=1}^{M} P(x_i\ ;\ yj) \log_2 \frac{P(x_i\ ;\ yj)}{P(x_i)P(y_j)} \qquad (1.7)$$

On peut également définir la moyenne de l'information propre d'un espace probabilisé

$X\ =\ \{x_1\ ,\ldots,x_N\ \}$, cette moyenne porte le nom d'entropie.

$$H(X) = \sum_{K=1}^{K} P(ak) I(ak)\ = \sum_{K=1}^{K} -P(ak)\ log2\ P(ak) \qquad (1.8)$$

Enfin, l'information propre conditionnelle est également une variable aléatoire réelle et nous pouvons définir sa moyenne.

1.5.2. L'entropie conditionnelle

L'entropie conditionnelle de *X* sachant *Y* dans l'espace probabilisé joint *XY* est définie par

$$H(X;\ Y\ ) = \sum_{i=1}^{N}\sum_{j=1}^{M} -P(x_i\ ;\ y_j) \log_2 P(x_i\ |\ y_j) \qquad (1.9)$$

L'équation (1.6) peut se réécrire en moyenne

$$I(X;\ Y\ )\ =\ H(X)\ -\ H(X\ |\ Y\ ) \qquad (1.10)$$

1.6. La fonction taux de distorsion

Les performances du codeur de source sont évaluées par sa courbe débit-distorsion R(D), pour une classe donnée de signaux $\{x[n]\}$. La valeur R(D) est le nombre minimal de bits nécessaires pour représenter la sortie y avec une distorsion D.

Cette opération de compression, de "diminution d'entropie", ou encore de diminution de redondance, est schématisée par la forme du bloc du codeur de source sur le schéma.

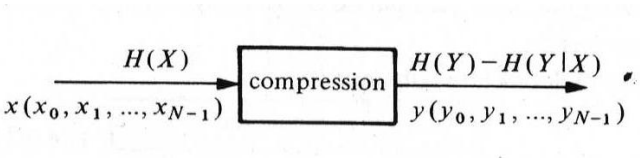

Fig.1.2 Quantificateur type, x est la grandeur d'entrée à quantifier et y est la valeur de sortie.

La fonction débit distorsion R(D) d'une source X donnée est l'information mutuelle I(X, Y) minimale entre cette source X et sa reconstitution Y, avec une distorsion donnée D.

La fonction appelée « taux de distorsion [1] » caractérise la relation entre les distorsions admises dans une transformation avec distorsion et la valeur minimum du débit d'information de la source effective. En effet, en réduisant le débit d'information, respectivement en augmentant la compression, les distorsions augmentent.

Soit $x(t) = x$ un vecteur *N*- dimensionnel qui représente, dans un espace métrique, le signal généré par la source et $y(t) = y$ le vecteur qui représente le signal comprimé transmis au destinataire. La distorsion introduite par la compression peut être mesurée en introduisant une « distance » $d(x, y)$ entre le signal x et le signal y.

Soit $x_0, x_1, \dots, x_{N-1}$ et $y_0,\ y_1, \dots, y_{N-1}$ les composantes (échantillons) du signal avant compression, respectivement après compression (Figure .1.2), on peut définir une mesure $d(x_i, y_i)$ de la distorsion subie par l'échantillon x_i , par exemple par la relation :

$$d(x_i, y_i) = E\{(x_i - y_i)^2\} \tag{1.11}$$

L'emploi de la moyenne statistique est justifié par le fait que x_i et y_i sont des variables aléatoires.

La distance entre les signaux x et y est

$$d(x, y) = \sum_{i=0}^{N-1} d(x_i, y_i) \tag{1.12}$$

Et elle mesure la distorsion introduite par la compression.

La réduction de l'entropie dans le processus de compression est équivalente à l'effet des perturbations dans un canal de transmission (Figure .1.2) le circuit de compression peut être donc considéré comme un canal affecté de perturbations et on peut définir la fonction « taux de distorsions » par la relation :

$$R(D) = min\, I(X;Y) = min[H(X) - H(X \mid Y)] \qquad (1.14)$$

Où le minimum est pris par rapport à l'ensemble des probabilités $p(x_i/y_i)$ et il est soumis à la condition que les distorsions $d(x,y)$ ne dépassent pas une valeur D admise

$$d(x,y) \leq D$$

La fonction $R(D)$ représente le débit d'information minimum à la sortie du circuit de compression (respectivement de la source effective) qui garantit que les distorsions admises D ne seront pas dépassées.

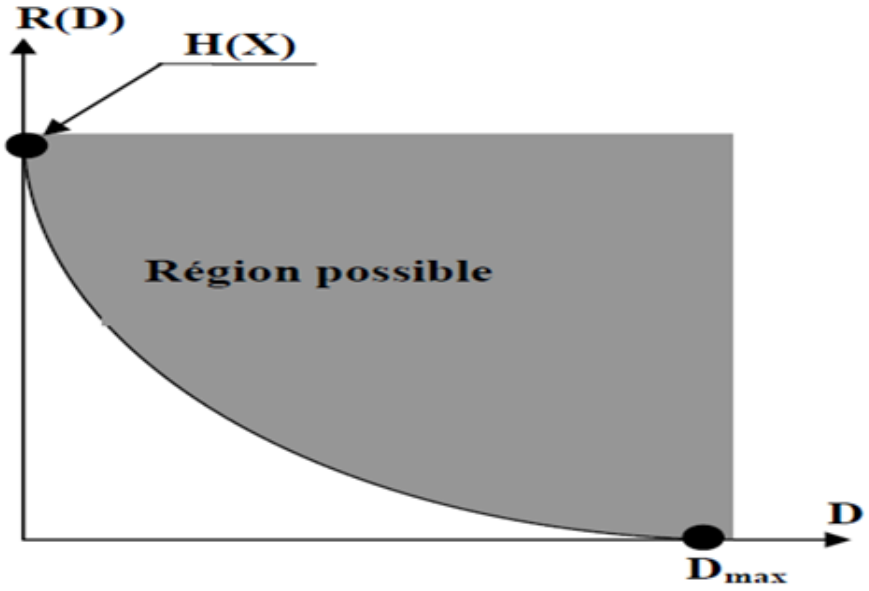

Fig.1.3 : Allure d'entropie en fonction de la distorsion d'une source discrète.

En conclusion. Une compression effectuée par une transformation avec distorsion, transforme la source de débit d'information $H(X)$ bits/s en une nouvelle source de débit $R(D)$ bits/s avec $R(D) < H(X)$ et de manière que les distorsions ne dépassent pas une valeur admise D. Par conséquent, on peut définir un rapport de compression

$$C = \frac{H(X)}{R(D)} \qquad (1.15)$$

Qui représente en fait une limite supérieure. Cette limite ne peut être dépassée par aucun procédé de compression sans augmenter la distorsion au-delà de la valeur admise D.

Si l'entrée dans le circuit de compression est constituée par les échantillons quantifiés du message, représenté par les nombres binaires, le rapport de compression est :

$$C = \frac{n_e}{n_s} \qquad (1.16)$$

Où n_e est le nombre de bits dans la suite d'entrée et n_s est le nombre de bits dans la suite de sortie.

Le calcul de $R(D)$ étant compliqué, le rapport de compression en général est calculé à l'aide de la relation (1.16) au lieu de (1.15)**.**

1.7. Quantification

1.7.1 Quantification scalaire

1.7.1.1 Définition

Au cours du traitement numérique du signal de parole, toutes les données sont représentées sur un certain nombre d'éléments binaires, avec une précision finie. La quantification est une opération qui consiste à arrondir une grandeur d'entrée en lui associant une valeur choisie dans un ensemble fini de valeurs prédéterminées. La quantification est le cœur de la numérisation des signaux. Outre la nécessité de la quantification pour numériser les données, elle est aussi un moyen de compression.

Un quantificateur scalaire Q à un taux R et N-niveaux de sortie, est une application de la ligne réelle $\Re$ à un ensemble d'éléments fini B.

$Q: \Re \rightarrow B$

$$x \rightarrow Q(x) = yi \qquad (1.17)$$

Où $B = \{y1, y2\ldots\ldots,,\ldots\ldots\ldots, yN\} \subset \Re$

L'ensemble B est appelé le dictionnaire et les éléments du dictionnaire y_i sont appelés les niveaux de reproduction ou les mots code.

La ligne réelle $\Re$ est divisée en N régions, Si; i=1,2,....., N. définie par

$S_i = \{x \in \Re : Q(x) = y_i\} \equiv Q^{-1}(y_i)$, i =1, 2, ..., N $\qquad (1.18)$

Les propriétés importantes de ces régions sont:

$$\bigcup_{i=1}^{N} Si = \Re \qquad (1.19)$$

Et $$S_i \cap S_j = \emptyset \text{ Pour } i \neq j. \qquad (1.20)$$

Un paramètre important d'un quantificateur scalaire est le taux de codage R, défini comme :

$$R = log_2 N \qquad (1.21)$$

Et il représente le nombre de bits requis pour identifier une sortie spécifique quantifiée.

L'application est exécutée comme suit: x étant l'entrée au quantificateur Q, alors La sortie $Q(x)$ du quantificateur est obtenue comme:

$$Q(x) = y_i \text{ si } x \in S_i, \ i = 1,2,\ldots\ldots\ldots, N. \qquad (1.22)$$

Par conséquent un quantificateur scalaire sans mémoire peut être complètement défini par le dictionnaire B et l'ensemble des régions $\{S_i\}$

Soit Q un quantificateur défini par un dictionnaire et par une partition de l'espace euclidien $\Re$. Un exemple de partition dans l'espace $\Re$ est représenté par la figure 1.4.

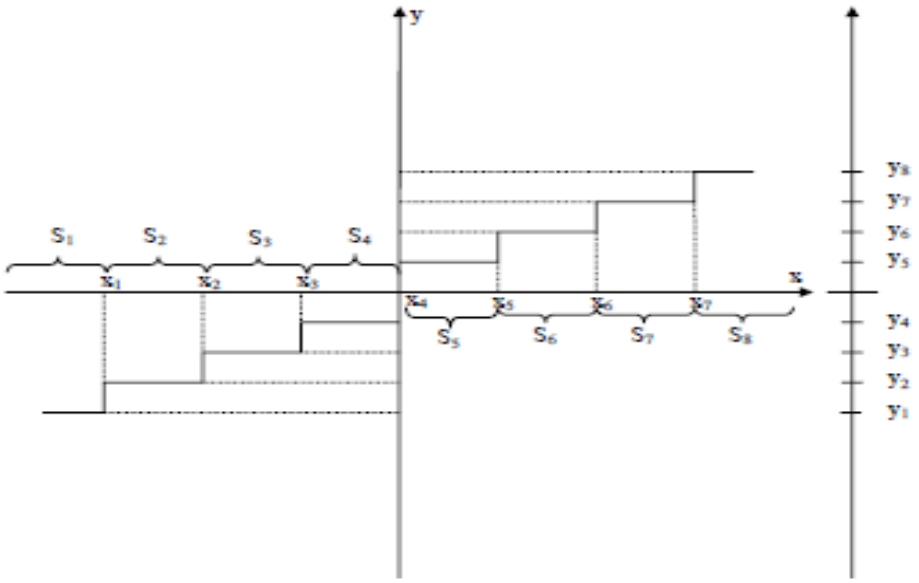

Fig.1.4 : caractéristique d'un quantificateur scalaire à 8 niveaux (N=8).

1.7.1.2 Mesure de la performance d'un quantificateur

Pour évaluer la performance du quantificateur nous choisissons comme mesure de distorsion l'Erreur Quadratique Moyenne:

$$d\,(x,y) = \,(x-y)^2 \tag{1.23}$$

Considérons la variable aléatoire x être l'entrée au quantificateur scalaire avec une fonction de densité de probabilité connu $p\,(x)$. La distorsion D entre l'entrée x et les sorties du quantificateur $Q\,(x)$ est définie comme:

$$D = E\big[d\big(x,Q(x)\big)\big] = \sum_{i=1}^{N} \int_{Si} (x-y_i)^2\, p(x)dx \tag{1.24}$$

Quand la source d'entrée de quantificateur est stationnaire et ergodique,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} d(\,x_i, Q(x_i)\,) = \,D \tag{1.25}$$

1.7.1.3 La conception d'un Quantificateur optimal

D'un point de vue engineering, l'amélioration de la performance est primordiale. Puisque la fonction de taux *R* est fixe pour un quantificateur scalaire, la distorsion est le foyer principal. La fonction de distorsion est une expression simple indicative de la dégradation de signal due à la quantification. Puisque l'entrée est inconnue, on suppose que la source est une variable aléatoire, habituellement décrite par sa fonction densité de probabilité. Le but de la conception d'un quantificateur scalaire est de réduire au minimum la distorsion *D*, pour un nombre fixe de niveaux de reproduction *N* et d'une fonction densité de probabilité particulière de source. Le quantificateur optimal est espéré pour être réalisé en sélectionnant convenablement les niveaux de reproduction $y_i$et les cellules de partition S_i. En général, il n'y a aucune solution connue pour trouver le quantificateur optimal, mais il y a des conditions nécessaires pour l'optimalité.

Un problème classique de quantification peut souvent être décrit comme suit. Pour une source donné, identiquement distribuée c-à-d avec une distribution connu p(x), une mesure de distorsion donné, $d\,(x,Q(x)) \;=\; (x-Q(x))^2$ et un nombre donné de niveaux de sortie du quantificateur N (taux fixe), nous souhaitons trouver le dictionnaire B et l'ensemble de partition $\{s_i\}$ tels que la distorsion définie dans l'équation (1.24) est minimisée.

1.7.1.4 Conditions d'optimalité

Le procédé de conception suivant qui fournit la solution au problème ci-dessus est le quantificateur Lloyd:

Considérons un quantificateur avec des régions de quantification de forme

$$S_i = \begin{cases} (x_{i-1}, x_i);\ i = 1,2,\ldots\ldots\ldots\ldots\ldots, N-1 \\ (x_{N-1}, x_N);\ i = N \end{cases} \quad (1.26)$$

Ici nous adoptons la convention $-\infty = x_0 < x_1 < x_2 < \ldots\ldots\ldots\ldots < x_N = \infty$ la distorsion moyenne peut être écrite comme :

$$D = \sum_{i-1}^{N} \int_{x_{i-1}}^{x_i} (x - y_i)^2\, p(x)\, dx \quad (1.27)$$

Si nous souhaitons minimiser *D* pour un taux fixe *R*, nous pouvons dériver les conditions nécessaires en différenciant *D* en ce qui concerne x_i et y_i et obtenir deux ensembles d'équations

$$D = \sum_{i=1}^{N} \int_{S_i} (x - y_i)^2\, p(x) dx \quad (1.28)$$

$$D = \int_{-\infty}^{x_1} (x - y_1)^2\, p(x)\, dx + \sum_{i=1}^{N-2} \int_{x_i}^{x_{i+1}} (x - y_{i+1})^2\, p(x) dx + \int_{x_{N-1}}^{\infty} (x - y_N)^2\, p(x)\, dx \quad (1.29)$$

On dérive D en ce qui concerne x_i

$$\frac{\partial}{\partial x_i} D = p(x_i)\{(x_i - y_i)^2 - (x_{i-} y_{i+1})^2\} = 0 \quad (1.30)$$

La solution de cette équation est :

$$x_i = \frac{y_i + y_{i+1}}{2} \quad i = 1, \ldots\ldots\ldots\ldots\ldots\ldots N-1 \quad (1.31)$$

On dérive D en ce qui concerne *yi*

$$\frac{\partial}{\partial y_i} D = \int_{x_{i-1}}^{x_i} 2(x - y_i) p(x) dx = 0$$

La solution de cette équation est

$$y_i = E[x/x \in S_i] = \frac{\int_{x_{i-1}}^{x_i} x\, p(x)dx}{\int_{x_{i-1}}^{x_i} p(x)dx}, \qquad i = 1, \ldots\ldots\ldots\ldots\ldots\ldots N-1 \qquad (1.32)$$

L'équation (1.31) implique que le niveau de seuil doive être l'intermédiaire entre deux niveaux adjacents de reconstruction. L'équation (1.32) suggère que le niveau de reconstruction optimal y_i est le centre entre x_{i-1}et x_i. Ainsi le problème de conception d'un quantificateur "optimal" peut être divisé en deux problèmes conceptuellement indépendants : (I) étant donné le dictionnaire B, trouver la meilleure partition de $\Re$; (II) étant donné la partition de la ligne réelle $\Re$, trouver le dictionnaire optimal B tels que la distorsion moyenne est minimisée. Les équations (1.31) et (1.32) peuvent être résolu itérativement pour x_i et y_i avec l'hypothèse initiale de y_1. Après chaque étape d'itération, la distorsion moyenne *D* est calculée. L'algorithme continue jusqu'à ce que la diminution relative de *D* de deux itérations consécutives soit moins qu'un seuil prédéfini. Le quantificateur résultant est appelé le quantificateur Lloyd. Il devrait souligner que le résultat du quantificateur obtenu est seulement une solution localement optimale selon les conditions initiales.

Nous pouvons prolonger les résultats dans les équations (1.31) et (1.32) à des mesures de distorsion plus générales *d*(.,.). Ces deux conditions nécessaires générales sont connues comme la condition du plus proche voisin et de Centroïde généralisé

Condition du plus proche voisin

Pour le dictionnaire donné B avec *N* niveaux de reproduction, la partition optimale satisfait:

$$S_i = \{x : d(x, yi) \leq d(x, yj);\ i \neq j\}\ ,\forall i. \qquad (1.33)$$

Condition du Centroïde

Pour une partition donnée $S = \{Si,\ i = 1,.....,.....\ N\}$ et un variable aléatoire d'entrée, les niveaux de sortie optimaux satisfait :

$$y_i = argmin\ E[di\ (x; y)\ /x \in S]\ \forall i \in \{1,2,.....,N\} \qquad (1.34)$$

Ces deux conditions généralisées peuvent être employées itérativement pour obtenir une solution localement optimale. L'opération d'espérance dans la condition du Centroïde généralisé implique que la distribution de la densité de probabilité de la source est connue. Dans la pratique, des séquences d'entraînement sont employées pour obtenir une distribution empirique de la source.

1.7.2 Quantification vectorielle

1.7.2.1 Définition

La quantification vectorielle est l'extension de la quantification scalaire à des dimensions plus élevées. Elle peut offrir divers d'avantages par rapport à la quantification scalaire comme elle exploite la redondance statistique entre les échantillons de la source.

La quantification vectorielle est une application de compression de données qui arrondi une séquence de vecteurs continus ou discrets par un nombre fini de vecteurs prédéterminés appelés niveaux de reproduction. Elle a été employée avec beaucoup de succès dans le codage d'image et de parole.

Un quantificateur k-dimensionnel à N niveaux est défini par le dictionnaire, $Bi = \{yi,\ i = 1,2\ldots\ldots\ldots, N\}$ avec N vecteurs de reproduction et l'ensemble de partition, $S = \{Si,\ i = 1,\ldots\ldots\ldots,\ldots\ldots\ldots\ N\}$ constitue les sous espaces de l'espace euclidien k^{ieme} dimension $\Re^K$ (Un vecteur $(x = x1, x2,\ldots\ldots\ldots\ldots\ldots, xk)$ constitué de k échantillons source représente un élément de l'espace euclidien $\Re^K$).

L'opération de la quantification vectorielle fait correspondre à tout vecteur $x = (x1, x2,\ldots\ldots\ldots\ldots\ldots, xk)$ un vecteur $y = (y_{i,1}\ ;\ y_{i,2}\ ;\ldots\ldots\ldots\ldots\ ;\ y_{i,k})$, après que le vecteur source soit arrondi par un des niveaux de reconstruction, l'indice du sous-espace auquel le vecteur x d'entrée appartient est envoyée (à travers le canal) au décodeur qui choisit alors la séquence correspondante à cette valeur de l'indice.

La QV est essentiellement une application:

$$Q(x) = y_i \quad si\ x \in Si \tag{1.35}$$

Où $y = \left(y_{i,1}\ ;\ y_{i,2}\ ;\ldots\ldots\ldots\ldots\ ;\ y_{i,k}\right)$

Et la partition S satisfait

$$\cup_{i=1}^{N} S_i = \Re^k \quad \text{et}\ \ S_i \cap S_j = \emptyset \quad \forall\ \ i \neq j \tag{1.36}$$

Le taux du quantificateur est défini comme

$$R = \frac{1}{K} log_2 N\ \text{ Bits / échantillon} \tag{1.37}$$

Un quantificateur vectoriel souvent utilisé dans les systèmes de communication numérique peut être décomposé on deux opérations conceptuelle, le Codeur et le Décodeur :

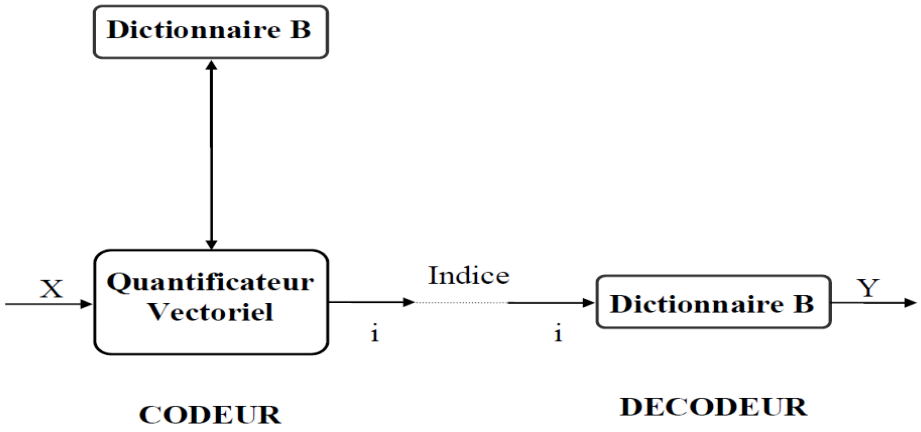

Fig.1.5 : Schéma d'un Codeur Décodeur Quantificateur.

1.7.2.2 Le Codeur

Le rôle du codeur consiste, pour tout vecteur x_i du signal en entrée à rechercher dans le dictionnaire B le code vecteur y_i le plus proche du vecteur source x_i. C'est uniquement l'adresse du code vecteur y_i ainsi sélectionnée qui sera transmise ou stockée. C'est à ce niveau donc que s'effectue la compression.

1.7.2.3 Le Décodeur

Il dispose d'une réplique du dictionnaire et consulte celui-ci pour fournir le code vecteur d'indice correspondant à l'adresse reçue. Le décodeur réalise l'opération de décompression.

1.7.2.4 Mesure de la performance d'un quantificateur

Pour évaluer la performance d'un quantificateur vectoriel, une mesure de distorsion doit être définie. La mesure de distorsion la plus commode et largement la plus répandue entre un vecteur *x* d'entrée et un vecteur quantifié *y* est l'erreur quadratique, qui est définie comme:

$$d(x,y) = || x - y ||^2 = \sum_{i=1}^{k} (x_i - y_i)^2 \qquad (1.38)$$

Pour mesurer la performance du QV, nous définissons la distorsion par échantillon dans le sens d'erreur quadratique moyenne, Considérons la variable aléatoire *x* être l'entrée au quantificateur avec une fonction de densité de probabilité connu *p* (*x*). La distorsion D entre l'entrée *x* et les sorties du quantificateur *Q (x)* est définie comme:

$$D = \frac{1}{k}\sum_{i=1}^{N} p(x)d(x, y_i)dx \qquad (1.39)$$

1.7.2.5 La conception d'un Quantificateur optimal

Le problème de conception d'un Quantificateur Vectoriel peut être énoncé comme suit: étant donné un vecteur source avec une connaissance des propriétés statistiques, une mesure de distorsion, le nombre des vecteurs code et sa dimension, trouver le dictionnaire optimal B et la meilleure partition S; tels que la distorsion moyenne est minimale.

Ainsi, donné le nombre de niveaux de reproduction *N*, et la dimension du bloc d'échantillons source *k* le but est de trouver le dictionnaire B^* et l'ensemble de partition S^* tels que la distorsion définie dans l'équation (1.39) est minimisé.

On a proposé la première fois la solution au problème ci-dessus par Linde, Buzo et Gray [47]. Les auteurs fournissent un algorithme itératif pour concevoir un quantificateur vectoriel localement optimal et le QV résultant s'appelle souvent LBG-VQ.

1.7.2.6 Conditions d'optimalité

Condition du centroïde

Pour un ensemble de partition donnée S, les conditions nécessaires pour minimiser l'équation (1.38) sans considérer des erreurs de canal sont:

$$y_i = \frac{\int_{S_i} xp(x)dx}{\int_{S_i} p(x)dx}, \qquad i = 1, \ldots\ldots\ldots\ldots\ldots\ldots N-1 \qquad (1.40)$$

Condition du plus proche voisin

Si nous supposons que le dictionnaire B est donné, quand les erreurs de canal ne sont pas considérées, alors l'ensemble de partition S qui réduit au minimum l'équation (1.39) est donnée par

$$S_i = \{x \,|d(x, yi) \leq d(x, yl) \quad \forall\, i \neq l\,\} \qquad (1.41)$$

L'équation (1.40) est la généralisation de la condition du centroïde. Lors que k=1 se réduit à la condition de la quantification scalaire donnée par l'équation (1.32) la région définie dans l'équation (1.41) est appelée région de Voronoi. Les équations (1.40) et (1.41) sont utilisées itérativement pour la mise à jour du dictionnaire et de l'ensemble de partition. Noté qu'après chaque itération, la distorsion par échantillon est diminue; donc, en général l'algorithme LBG converge à une solution localement optimale. Le choix du dictionnaire initial joue un rôle très important dans l'algorithme LBG.

1.7.3 Quantification vectorielle à dimension variable

La VDVQ (Variable Dimension Vector Quantization) [33], est basée sur la supposition que la génération d'un vecteur à dimension variable, est le résultat d'un échantillonnage uniforme d'un autre vecteur à dimension fixe et large. Cette technique fonctionne comme suit :

Avant la formation (training) du dictionnaire, chaque spectre dans la séquence d'entraînement est, d'abord, interpolé à bande limitée en un vecteur à dimension fixe *L*. Le choix naturel de cette dimension est le nombre maximal d'harmoniques dans le spectre à coder. Une fois que tous les vecteurs d'entraînement sont convertis à la même dimension, on applique la technique GLA conventionnelle pour former le dictionnaire. Par conséquent, le dictionnaire résultant aura la dimension uniforme *L*.

La quantification d'un vecteur de longueur *L'* consiste à le sur-échantillonné, pour avoir la longueur *L*, ensuite on choisi dans le dictionnaire le vecteur qui minimise l'erreur quadratique moyenne MSE.

Après la dé-quantification, le spectre de longueur *L* est sous-échantillonné ainsi pour avoir sa longueur initial *L'*.

1.8. Evaluation de qualité de signal

Il est évident dans tous les procédés de codage de juger la qualité du signal reconstruit. Parallèlement à la recherche d'un débit de transmission ou d'un taux de compression minimale.

Deux types de mesures, objective et subjective [5], peuvent permettre l'évaluation de la qualité de parole.

1.8.1. Tests subjectifs

Ce type est basé sur le jugement des participants à qui on demande de tester un système de télécommunications dans différentes conditions et de noter sur une échelle d'aptitude la qualité de signal de ce système. Ce test est très recommandé pour les signaux d'utilisation directe par l'utilisateur (parole, image, vidéo,…).

Dans le cas de la parole, la qualité est mesurée par l'intelligibilité spécifiquement définie par le pourcentage de mots ou phonèmes correctement écoutés et avec une sonorité naturelle

Il existe trois types de mesures subjectives de la qualité généralement utilisées [34].

- Le test DRT (Diagnostic Rhyme Test)
- Le test DAM (Diagnostic Acceptability Measure)
- Le test MOS (Mean Opinion Score)

MOS	Qualité
1	Mauvais
2	Médiocre
3	Passable
4	Bon
5	Excellant

Tableau .1.1: Qualité avec la mesure MOS [14].

Les mesures subjectives ne peuvent pas être calculées automatiquement, donc elles ne peuvent pas être utilisées comme des critères d'optimisation dans un algorithme de codage.

1.8.2. Tests objectifs

Bien que les méthodes subjectives soient le seul moyen d'atteindre le jugement des utilisateurs, les opérateurs de télécommunications cherchent à éviter le recours à de telles méthodes, du fait du coût et du temps qu'elles demandent.

Les mesures objectives utilisent des fonctions ou des critères mathématiques pour comparer les formes d'onde, les spectres ou les cepstres codées et originales.

Les mesures de distorsion objectives simples les plus couramment utilisée dans le domaine temporel mentionnant :

- Le Rapport Signal sur Bruit (SNR) : Si s(n) est le signal de parole original, ŝ (n) est le signal parole reconstitué comportant Nτ échantillons, alors le SNR représente le rapport de la puissance du signal à la puissance du bruit est définit comme suit [16]:

$$SNR(dB) = 10\log_{10}\frac{\sum_{n=0}^{N\tau-1} s^2(n)}{\sum_{n=0}^{N\tau-1}(s(n)-\hat{s}(n))^2} \qquad (1.42)$$

D'après (1.42) le SNR ne peut prendre décision qu'après avoir écouter le fichier parole entier.

- Le rapport signal sur bruit segmental (SEGSNR) : pour manipuler la nature dynamique des signaux non stationnaires tels que la parole, Le signal est découpé en NF segments de NS échantillons chacun, et on calcule une moyenne [16] (s(n) est le signal original et ŝ(n) le signal synthétisé). Le SEGSNR est une meilleure mesure que le SNR, mais ce n'est pas toujours le cas dans la parole quand la trame entière est presque silencieuse. Pour cela en fait appel à d'autres mesures objectives.

$$SEGSNR = \frac{1}{N_F}\sum_{i=0}^{N_F-1} 10\log_{10}\frac{\sum_{j=0}^{N_S-1} s^2(N_S*i+j)}{\sum_{j=0}^{N_S-1}(s(N_S*i+j)-\hat{s}(N_S*i+j))^2} \qquad (1.43)$$

Il existe, des méthodes objectives plus poussées que les mesures objectives simples telles que le rapport SNR et SEGSNR. Ce ne pas le lieu de les cités tous, nous sommes suffisant avec ces deux mesure suivant notre travail.

1.9. Conclusion

Ce chapitre est une introduction qui rehausse l'importance de compression des données au domaine de communication. Et une représentation d'une manière générale de principales définitions et outils de compression. Parce que ce sujet fait partie du domaine de théorie de l'information, on a listé juste ce qu'on avait besoin, ou ce qui a une relation avec notre travail.

Chapitre 2 : Généralités sur la parole

2.1. Communication par la parole

La parole est la faculté de communiquer la pensée par un système de sons articulés ; c'est le moyen de communication privilégié entre les humains qui sont les seuls êtres vivants à utiliser un tel système structuré.

L'information d'un message parlé réside dans les fluctuations de la pression de l'air, engendrées, puis émises, par l'appareil phonatoire. Ces fluctuations constituent le signal vocal ; elles sont détectées par l'oreille, laquelle procède à une certaine analyse. Les résultats sont transmis au cerveau, qui interprète. Un message vocal est une suite d'images auditives, éléments minimaux dénués de sons, mais dont l'association permet de distinguer les éléments constitutifs des niveaux supérieurs (syllabe, mots, phrases,…).

Le contenu d'un message vocal au sens strict est simplement son intelligibilité ; il ne se distingue pas en cela du message écrit. Au sens large ; il faut bien sûr prendre en compte toutes les intonations, ce qui en accroît fortement la richesse par rapport au message écrit.

Le signal vocal est caractérisé par une très grande redondance, condition nécessaire pour résister aux perturbations du milieu ambiant. La redondance est également présentée au niveau sémantique, ce qui facilite la compréhension du message par le cerveau.

On peut tenter de déterminer la cadence maximale à laquelle un auditeur peut assimiler un message. On va dans ce but définir l'information associée à un message constitué par des éléments discrets x_i appartenant à un ensemble donné X. Si $p(x_i)$ est la probabilité à priori d'occurrence du symbole x_i, sa sélection apporte une information [8] :

$$I = -\log_2 p(x_i)$$

L'information moyenne associée à l'occurrence du message $[X] = [x_0 \ x_1 \ \dots \ x_n]$vaut [8] :

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \qquad (2.1)$$

C'est l'entropie de la source, exprimée en bits.

2.2. Mécanisme de la phonation

la parole est le résultat de l'action volontaire et coordonnée des appareils respiratoire et masticatoire. Cette action se déroule sous le contrôle du système nerveux central qui reçoit en permanence des informations par rétroaction auditive et par les sensations cénesthiques.

L'appareil respiratoire fournit l'énergie nécessaire lorsque l'air est expiré par la trachée-artère . au sommet de celle-ci se trouve le larynx où la pression de l'air est modulée avant d'être appliquée au conduit vocal, qui s'étend du pharynx jusqu'aux lèvres (figure .2.1)

Le larynx est un ensemble de muscles et de cartilages mobiles qui entourent une cavité située à la partie supérieure de la trachée (figure.2.2a). Les cordes vocales sont en fait deux lèvres symétriques placeées en travers du larynx ; ces lèvres peuvent fermer complètement le larynx et, en s'écartant, déterminer une ouverture triangulaire appelée glotte (figure .2.2b). l'air y passe librement pendant la respiration et la voix chuchotée, et aussi pendant la phonation des sons sourds ou non voisés. Les sons voisés résultent au contraire d'une vibration périodique des cordes vocale ; des impulsions périodiques de pression sont ainsi appliquées au conduit vocal.

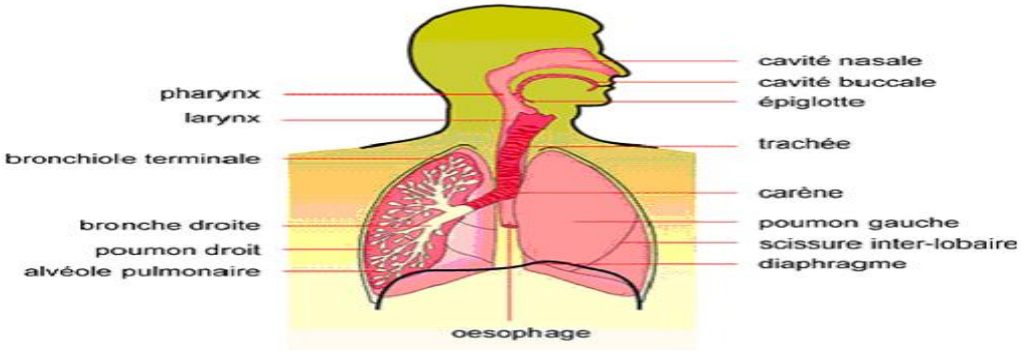

Fig.2.1 : Coupe sagittale de l'appareil phonatoire :

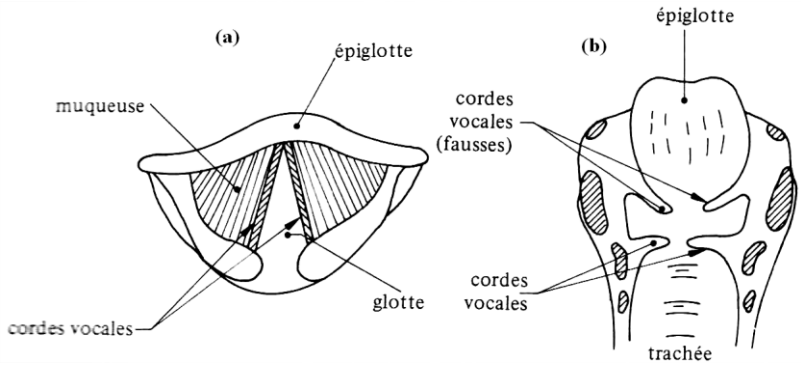

Fig.2.2 : vues de larynx :(a) vue de haut ;(b) coupe verticale.

Ce dernier est un ensemble de cavités situées entre la glotte et les lèvres ; on peut sur la figure.2.1 distinguer la cavité pharyngienne, la cavité buccale et , en dérivation, la cavité nasale.

Le conduit vocale peut être considéré comme une succession de tubes ou cavités acoustiques de sections diverses.

Les sons voisés résultent donc de l'excitation du conduit vocal par des impulsions périodiques de pression liées aux oscillations des cordes vocales ; l'ouverture brusque de la glotte libère la pression accumulée en amont ; elle se referme ensuite plus graduellement.

Les voyelles orales (i,e,u,…) sont émises sans intervention de la cavité nasale qui est alors isolée par la fermeture du voile du palais ; les voyelles nasales (ã,e□,…) et les consonnes nasales (m, n,…) font intervenir la cavité nasale [9].

Le couplage entre la glotte et le conduit vocal est faible : la déformation de ce dernier influence peu l'onde de pression engendrée par l'ouverture de la goltte.

L'intensité du son émis est liée à la pression de l'air en amont du larynx ; sa hauteur est fixée par la fréquence de vibration des cordes vocales, appelée fréquence du fondamental ou pitch.

La fréquence du fondamental peut varier [8] :

- de 80 à 200 Hz pour une voix masculine
- de 150 à 450 Hz pour une voix féminine,
- de 200 à 600 Hz pour une voix d'enfants.

Deux sons de même intensité et de même hauteur se distinguent par le timbre, qui est déterminé par les amplitudes relatives des harmoniques du fondamental.

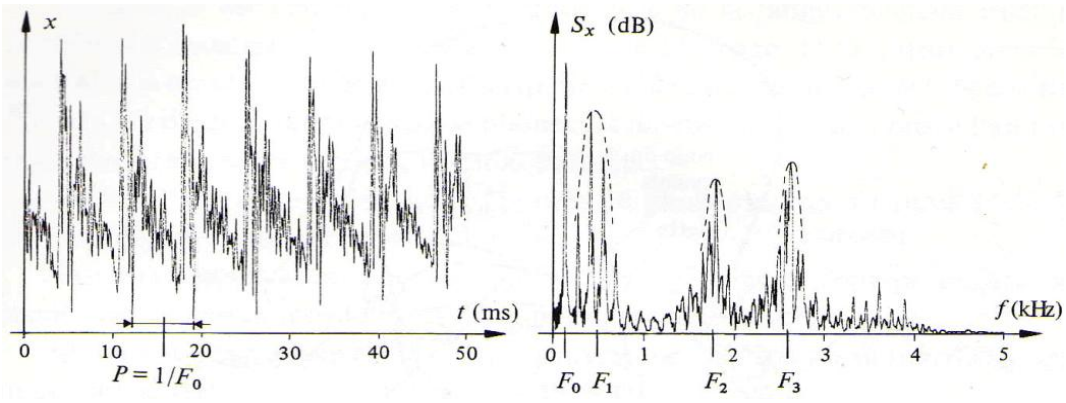

Fig.2.3 : un signal vocal et son spectre.

Un son voisé est un signal quasi périodique dont le spectre est esquissé à la figure.2.3, on y observe les raies qui correspondent aux harmoniques du fondamental F0 (structure de pitch) ; l'enveloppe de ces raies présente des maximums appelés formants et qui correspondent aux fréquences propres Fi (i=1, 2, 3…) du conduit vocal (structure formantique) [8].

Les trois premiers formants sont essentiels pour caractériser le spectre vocal ; les formants d'ordres supérieurs ont une influence plus limitée [9].

Les sons fricatifs résultent de l'écoulement de l'air dans une constriction étroite située en un point du conduit vocal, en particulier au niveau des lèvres et des dents. Les sons fricatifs sont non voisée (f, s, …) ou voisés (v, z, …).

Un son plosif (ou occlusif) est produit par une occlusion momentanée du conduit vocal en un point donné, suivie par une ouverture brusque ; il peut être voisé (b, d,…) ou non voisé (p, t, …).

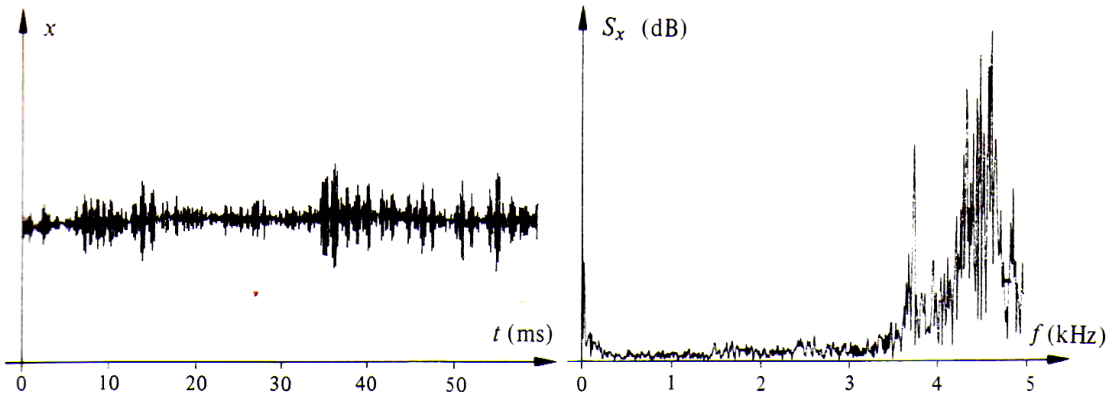

Fig.2.4 : Son non voisé et son spectre.

Les sons non voisés sont le résultat du passage du flux d'air par une étroite constriction au niveau du conduit vocal causant des turbulences, c'est-à-dire du bruit. Contrairement aux sons voisés, les sons non voisés ne présentent pas de structure périodique. Ils peuvent être modélisés par un bruit blanc filtré par la transmittance de la partie du conduit vocal située entre la constriction et les lèvres (figure.2.4) ; son spectre ne présente donc pas de structure de pitch. La classification qui vient d'être esquissée est forcément un peu sommaire et surtout elle concerne la production normale de la parole. Ainsi une voyelle peut être chuchotée, c'est-à-d produite avec la glotte largement ouverte ; dans ce cas le spectre du signal résulte de l'excitation du conduit vocal par une source aléatoire : c'est un spectre continu qui présente une structure formantique semblable à celle d'une voyelle voisée. Par contre, il ne possède pas de structure de pitch (raies dues aux harmoniques du fondamental).

PHONEMES								
VOYELLES		SEMI - VOYELLES	CONSONNES					
					FRICATIVES		OCCLUSIVES	
ORALES	NASALES		LIQUIDES	NASALES	voisées	non voisées	voisées	non voisées
[i]	[ɛ̃]	[w]	[l]	[m]	[v]	[f]	[b]	[p]
[ɔ]	[ɑ̃]	[j]	[r]	[n]	[z]	[s]	[d]	[t]
[a]				[ɲ]	[ʒ]	[ʃ]	[g]	[k]
[u]								
[o]								
[y]								
[ɛ]								

Tab.2.1 : Les phonèmes français

2.3. Analyse de la parole

Le traitement du signal vocal s'inscrit dans une succession de procédures, que ce soit pour la reconnaissance automatique ou pour la synthèse de la parole. Analyse et synthèse sont deux activités duales, l'analyse fournit une description du signal acoustique, que la synthèse utilise pour le reproduire.

Le traitement est aussi utilisé pour réduire la redondance du signal vocal et permet ainsi de comprimer l'onde avant l'enregistrement, la transmission, ou l'extraction des paramètres pertinents pour la reconnaissance. Dans cette dernière perspective, les paramètres les plus couramment extraits sont issus d'une modélisation paramétrique autorégressive sous codage à prédiction linéaire.

2.3.1. Modélisation de la parole

L'analyse de la parole est une étape indispensable à toute application de synthèse, de codage ou de reconnaissance. Elle repose en général sur un modèle. Il existe de nombreux modèles de parole. On distingue les modèles articulatoires, les modèles de production, et les modèles phénoménologiques. Dans le processus de codage, on s'intéresse aux modèles de production. On y décrit la parole comme le signal produit par un assemblage de générateurs et de filtres numériques (modèle source filtre). Les paramètres de ces modèles sont ceux des générateurs et filtres qui les constituent. Le modèle Autorégressif (AR) en est l'exemple le plus utilisé.

Avant d'entamer les détails de modèle on donne une note sur les principales caractéristiques de signal parole. La parole est un signal réel, continu, d'énergie finie, non stationnaire.

Sa structure est complexe et variable dans le temps : tantôt périodique (plus exactement pseudopériodique) pour les sons voisés, tantôt aléatoire pour les sons fricatifs, tantôt impulsionnelle dans les phases explosives des sons occlusifs. Cette structure reflète l'organisation temporelle des gestes de production et sur l'onde sonore apparaissent quelques caractéristiques de la source et du conduit tel que la fréquence fondamentale F0 et la fréquence des formants Fi.
La structure formantique est attribuée au conduit vocal qui agit comme un filtre ayant comme résonances les pôles de la fonction de transfert ou formants, et comme antirésonances les zéros de la fonction de transfert.

2.3.2. Le modèle AR de la parole

Un modèle électrique a été proposé par Fant en 1960 [10]. Qui spécifie qu'un signal voisé peut être modélisé par le passage d'un train d'impulsions $u\ (n)$ à travers un filtre numérique récursif de type tous pôles. On montre que cette modélisation reste valable dans le cas des sons non voisés, à condition que $u\ (n)$ soit cette fois un bruit blanc. Le modèle final est illustré à la figure.2.5. Il est souvent appelé modèle auto régressif (AR), parce qu'il correspond dans le domaine temporel à une régression linéaire de la forme [8] :

$$X(n) = G.U(n) - \sum_{i=1}^{N} a(i)\, X(n-i) \tag{2.2}$$

L'absence de couplage entre la glotte et le conduit vocal permet de modéliser séparément la source et le système de production.

La modélisation proposée utilise le formalisme des systèmes échantillonnés : Les fonctions de transfert sont donc des fonctions de z.

Ce choix est amplement justifié par l'évolution de la technologie des processeurs numériques, utilisés universellement en traitement de la parole.

Pour les sons voisés, la source est un train périodique d'ondes de forme particulière (montée rapide en pression suivie d'une chute plus graduelle). Ce train d'ondes est modélisé par la réponse d'un passe-bas d'ordre 2 à pôle réels et dont la fréquence de coupure est de l'ordre de 100 Hz ; sa transmittance est de la forme :

$$G(z) = \frac{\mathrm{A}}{(1+\alpha \mathrm{z}^{-1})(1+\beta \mathrm{z}^{-1})} \tag{2.3}$$

Pour les sons non voisés, la source est un bruit blanc.

On a vu que le conduit vocal peut être assimilé à une succession de tubes acoustiques élémentaires. L'étude de la propagation d'une onde acoustique plane conduit à une modélisation par une cascade de résonateurs dont la transmittance est de la forme :

$$V(z) = \frac{\mathrm{A}}{\prod_{\mathrm{k}=1}^{\mathrm{k}}(1 + \mathrm{b}_{1\mathrm{k}}\mathrm{z}^{-1} + \mathrm{b}_{2\mathrm{k}}\mathrm{z}^{-2})} \tag{2.4}$$

Chaque résonateur correspond à un formant dont la fréquence centrale est donnée par :

$$F_k = \frac{1}{2\pi} f_s \cos^{-1}\left[\frac{-b_{1k}/2}{\sqrt{b_{2k}}}\right] \tag{2.5}$$

Où f_s est la fréquence d'échantillonnage.

Le son est finalement émis à travers l'ouverture des lèvres ; celle-ci représente une charge acoustique ; le rayonnement des lèvres peut être modélisé par la transmittance :

$$R(z) = C(1 - \mathrm{z}^{-1}) \tag{2.6}$$

Qui exprime que la pression de l'onde observée à une certaine distance des lèvres est proportionnelle à la dérivée du débit volumique aux lèvres.

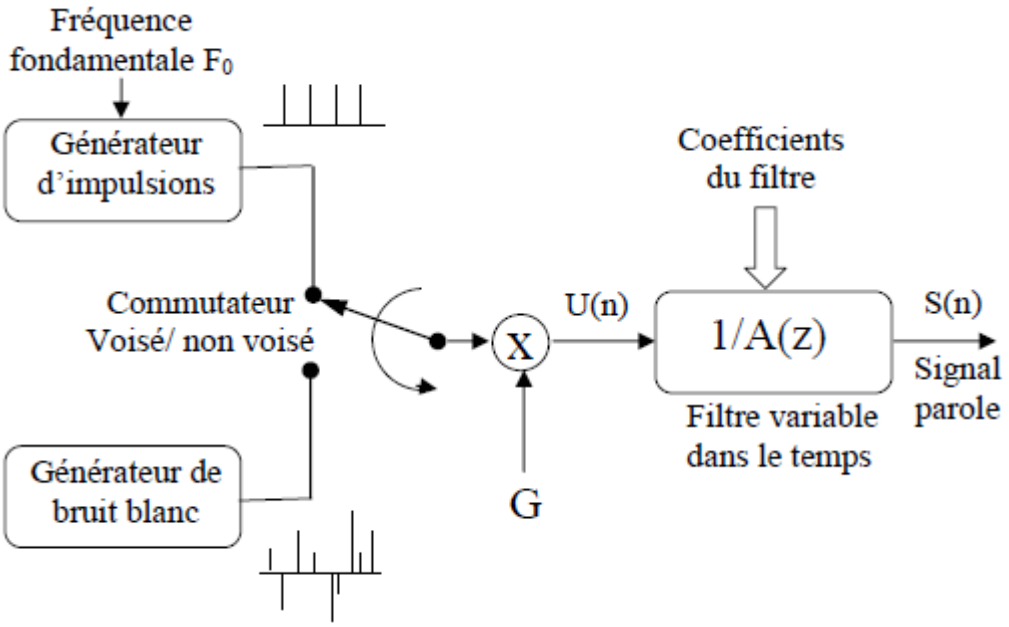

Fig.2.5 : Modèle simplifié de production de la parole. G est le gain, *U(n)* l'excitation, *s(n)* le signal de parole.

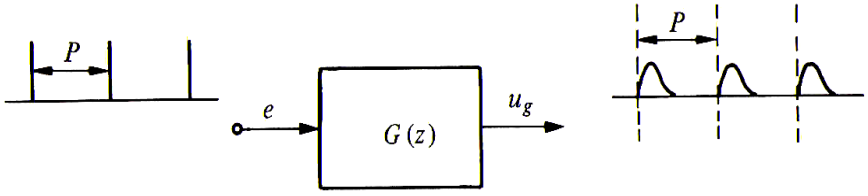

Fig.2.6 : modélisation de la source pour les sons voisés.

En résumé, la transmittance globale entre le train d'impulsions de la figure.2.6 et le signal émis serait :

$$T(z) = G(z)V(z)R(z) \tag{2.7}$$

$$T(z) = \frac{\sigma(1 - \mathrm{z}^{-1})}{(1 + \alpha \mathrm{z}^{-1})(1 + \beta \mathrm{z}^{-1}) \prod_{\mathrm{k}=1}^{\mathrm{k}} (1 + \mathrm{b}_{1\mathrm{k}} \mathrm{z}^{-1} + \mathrm{b}_{2\mathrm{k}} \mathrm{z}^{-2})}$$

Si on considère que l'un des pôles de $G(z)$ est très voisin de l'unité, on obtient la forma simplifiée :

$$T(z) = \frac{\sigma}{(1 + \alpha \mathrm{z}^{-1}) \prod_{\mathrm{k}=1}^{\mathrm{k}} (1 + \mathrm{b}_{1\mathrm{k}} \mathrm{z}^{-1} + \mathrm{b}_{2\mathrm{k}} \mathrm{z}^{-2})}$$

$$T(z) = \frac{\sigma}{A(z)} \tag{2.8}$$

On a posé:

$$A(z) = (1 + \alpha \mathrm{z}^{-1}) \prod_{\mathrm{k}=1}^{\mathrm{k}} (1 + \mathrm{b}_{1\mathrm{k}} \mathrm{z}^{-1} + \mathrm{b}_{2\mathrm{k}} \mathrm{z}^{-2})$$

$$A(z) = 1 + \sum_{i=1}^{2k+1} a_i z^{-i} \tag{2.9}$$

Alors

$$T(z) = \frac{X(z)}{U(z)} = \frac{\sigma}{A(z)} = \frac{\sigma}{1 + \sum_{i=1}^{2k+1} a_i z^{-i}} \tag{2.10}$$

La transmittance de ce modèle est dite tous pôles ; son inverse , le polynôme $A(z)$ est la transmittance du filtre inverse. On trouvera une justificatin théorique et une vérification expérimantale de ce modèle. Ses limitations sont cependant évidentes.

En premier lieu, la soure est soit un train périodique d'impulsions, soit un bruit blanc ; les sons fricatifs voisés ne peuvent pas être produits par ce modèle.

En second lieu, la productin de sons nasalisés fait intervenir deux cavités associées en parallèle ; la transmittance correspondante est donc de la forme

$$\frac{\sigma_1}{A_1(z)} + \frac{\sigma_2}{A_2(z)} = \frac{\sigma_1 A_2(z) + \sigma_2 A_1(z)}{A_1(z) A_2(z)}$$

Et elle présente des zéros en z distincts de l'origine.

Toutefois, on verra que la transmittance tous pôles est la base de la modélisatin par prédiction linéaire ; la présence d'un numérateur qui ne serait pas une simple constante complique énormément l'estimation des paramètres du modèle. On spécule donc sur l'identité ;

$$(1 - az^{-1}) \cong \frac{1}{1 + \mathrm{az}^{-1} + \mathrm{a}^2\mathrm{z}^{-2} + \cdots}$$

Pour substituer à un zéro en un ou deux pôles ; en d'autres termes, on accepte de surestimer le degré du dénominateur pour pouvoir assimiler le numérateur à une constante.

Enfin il est essentiel de rappeler que le signal vocal n'est pas un signal stationnaire : le conduit vocal se déforme d'une façon continue ; les paramètres du modèle sont donc variables dans le temps. Toutefois, les déformations sont suffisamment lentes pour que les coefficients puissent être maintenus constants pendant des intervalles de temps de l'ordre de 20 ms.

2.3.3. Analyse par prédiction linéaire

La méthode du codage à court terme par prédiction linéaire LPC, est devenue une technique prédominante dans l'estimation des paramètres de base de la parole, tels que pitch et formants. En téléphonie cellulaire, elle constitue un standard de transmission à faible débit de la parole. L'importance de cette méthode est liée à la fois, aux estimations précises des paramètres de la parole qu'elle procure et sa relative rapidité en calcul.

L'idée de base de l'analyse à prédiction linéaire est qu'un échantillon de la parole peut être approximé par une combinaison linéaire d'échantillons passés de la parole. Par minimisation de la somme des carrées de différence (principe des moindres carrées) entre les échantillons présents de la parole et ceux prédits linéairement, un ensemble unique de coefficients de prédiction peut être déterminé, identifiant ainsi le signal de la parole.

2.3.4. Formalisme de LPC

De (2.10), les échantillons de la parole $X(n)$ sont liés à l'excitation $u(n)$ par une équation de différence :

$$X(n) = -\sum_{i=1}^{P} a(i)\, X(n-i) + \; G.U(n) \tag{2.11}$$

Ainsi, chaque échantillon du signal original $X(n)$ est approché par une combinaison linéaire des p échantillons qui le précèdent. Cette prédiction n'est possible que si $X(n)$ est autocorrélé. Les coefficients $a(k)$ sont appelés coefficients de prédiction d'ordre p, et le signal $G\, u(n) \;=\; e(n)$ qui est l'excitation, est aussi interprété comme erreur de prédiction (ou résidu) d'ordre p qu'il faut minimiser (excitation comme bruit blanc de moyenne nulle).

$$e(n) = \sum_{i=0}^{P} a(i)\, X(n-i) \qquad avec\ a(0) = 1 \tag{2.12}$$

Nous supposons d'abord que le signal $X(n)$ est aléatoire et stationnaire. Les coefficients $a(k)$ sont donc indépendants du temps et leur estimation est basée sur la minimisation de la variance de l'erreur de prédiction ou énergie résiduelle de prédiction soit :

$$\sigma_e^2 = \sum_{n=-\infty}^{+\infty} e^2(n) = \sum_{n} \left[X(n) + \sum_{i=1}^{P} a(i)\, X(n-i) \right]^2 \tag{2.13}$$

Trouver les a(i) qui minimisent σ_e^2, revient à annuler les dérivées partielles de σ_e^2 par rapport à ces coefficients eux même.

$$\begin{aligned}
\frac{d\sigma_e^2}{da(j)} &= 2 \sum_{n=-\infty}^{+\infty} \left[X(n) + \sum_{i=1}^{P} a(i)\, X(n-i) \right] X(n-j) \qquad pour\ j = 1 \dots P \\
&= 2 \left\{ \sum_{i=1}^{P} \left[a(i) + \sum_{n} X(n-i)X(n-j) \right] + \sum_{n} X(n)X(n-j) \right\} \\
&= 2 \left\{ \sum_{i=1}^{P} a(i) \sum_{n} X(m)X(m+n-j) + \sum_{n} X(n)X(n-j) \right\} \qquad avec\ m = \; n-i \\
\frac{d\sigma_e^2}{da(j)} = 0 \Longrightarrow \sum_{i=1}^{P} \left[a(i) \sum_{n} X(m)X(m+n-j) \right] &= -\sum_{n} X(n)X(n-j) \quad pour\ j = 1 \dots P
\end{aligned} \tag{2.14}$$

Ce sont les équations de Yule-Walker.

On sait que pour un signal stationnaire, la fonction d'auto-corrélation vérifie :

$$\Phi_X(i) = \Phi_X(-i) = \sum_1 X(1)X(1+i) \qquad et \quad \Phi_X(0) = \sigma_X^2$$

Les équations de Yule-Walker deviennent alors :

$$\sum_{i=1}^{P} a(i)\,\Phi_X(i-j) = -\Phi_X(j) \qquad pour\ j = 1 \dots P \tag{2.15}$$

En posant :

$$\Phi^{(p)} = [\Phi_X(1), \Phi_X(2), \dots \Phi_X(P)]^T$$

Et :

$$\Phi^{(p)} = \begin{bmatrix} \phi_x(0) & \phi_x(1) & \cdots & \phi_x(p) \\ \phi_x(1) & \phi_x(0) & \cdots & \phi_x(p-1) \\ \vdots & \vdots & & \vdots \\ \phi_x(p) & \phi_x(p-1) & \cdots & \phi_x(0) \end{bmatrix} = \begin{bmatrix} \sigma_x^2 & \phi^{(p)T} \\ \phi^{(p)} & \Phi^{(P-1)} \end{bmatrix}$$

Les équations de Yule-Walker (2.14) en notation matricielle, deviennent :

$$\begin{bmatrix} \phi_x(0) & \phi_x(1) & \bullet & \bullet & \bullet & \phi_x(p-1) \\ \phi_x(1) & \phi_x(0) & & & & \phi_x(p-2) \\ \bullet & & & & & \bullet \\ \bullet & & \bullet & & & \bullet \\ \bullet & & & \bullet & & \bullet \\ \phi_x(p-2) & & & & \bullet & \phi_x(1) \\ \phi_x(p-1) & \bullet & \bullet & \bullet & \bullet & \phi_x(0) \end{bmatrix} \bullet \begin{bmatrix} a(1) \\ a(2) \\ \bullet \\ \bullet \\ \bullet \\ a(p-1) \\ a(p) \end{bmatrix} = - \begin{bmatrix} \phi_x(1) \\ \phi_x(2) \\ \bullet \\ \bullet \\ \bullet \\ \phi_x(p-1) \\ \phi_x(p) \end{bmatrix}$$

Soit :

$$\Phi^{(p-1)}A = -\Phi^{(p)} \tag{2.16}$$

La matrice $\Phi^{(p-1)}$ est dite de Toplitz, puisqu'elle vérifie :

- la symétrie.
- Elle est définie positive.
- Les éléments situés sur une diagonale parallèle à la diagonale principale sont identiques.

Plusieurs algorithmes ont été proposés pour la résolution de ce système, afin de tirer le vecteur inconnu A relatif aux coefficients de prédiction $a(k)$. Le plus utilisé est celui de Levinson-Durbin [11].

En revenant à l'expression (2.13), celle-ci s'écrit encore :

$$\sigma_e^2 = \sum_n \left[\sum_{j=0}^{P} a(j)X(n-j) . \sum_{i=0}^{P} a(i)X(n-i) \right]$$

$$= \sum_{j=0}^{P} a(j) \sum_{i=1}^{P} a(i) . \sum_n X(n-j)X(n-i)$$

Pour $n - j = m$, et $n - k = m + j - k$, on obtient :

$$\sigma_e^2 = \sum_{j=0}^{P} a(j) \sum_{i=1}^{P} a(i) . \Phi_X(j-i) \tag{2.17}$$

En remplaçant (2.15) dans (2.17), on obtient la valeur minimisée de la variance de l'erreur de prédiction :

$$\sigma_{e_{min}}^2 = G_P = \sum_{j=0}^{P} a(j)\Phi_X(j) \tag{2.18}$$

Le signal de la parole étant fortement non-stationnaire, ce type de modélisation ne reste guère valable plus de 30ms. On notera donc que l'analyse LPC d'un signal de la parole, implique la résolution d'un système de 10 équations à 10 inconnues toutes les 10 ms (à 30 ms au maximum).
D'autre part, si l'on réunit (2.16) et (2.18), on obtient :

$$\Phi_X(j) \begin{bmatrix} 1 \\ A \end{bmatrix} = \begin{bmatrix} G_p \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{2.19}$$

Comme $e(n) = \boldsymbol{G}\, u(n)$, et pour une excitation $u(n)$ du type bruit blanc de moyenne nulle et de variance unitaire, le gain de modèle devient :

$$G = \sigma_{e_{min}} = \sqrt{G_P} \tag{2.20}$$

2.3.5. Considérations pratiques

La meilleure performance de l'analyse LP se base essentiellement sur [4] :

Le choix de la fréquence d'échantillonnage est fonction de l'application visée, et de la qualité du signal à analyser. On choisira plutôt 8 kHz pour les signaux téléphoniques, 10 kHz pour les applications de reconnaissance, et 16 kHz pour les applications de synthèse…etc.

- L'ordre d'analyse conditionne le nombre de formants que l'analyse est capable de prendre en compte. On estime en général, que la parole présente un formant par 1kHz de bande passante, ce qui correspond à une paire de pôles pour A(z). Si on y ajoute une paire de pôles pour la modélisation de l'excitation glottique, on obtient les valeurs classiques de N=10, 12, et 18 pour fe=8, 10 et 16 kHz
- La durée des tranches d'analyse et leur décalage sont souvent fixés entre 30 et 10 ms. Ces valeurs ont été choisies empiriquement ; elles sont liées au caractère quasi-stationnaire du signal de parole.
- Enfin, pour compenser les effets de bord, on multiplie en général préalablement chaque tranche d'analyse par une fenêtre de pondération w(n) de type fenêtre de ***Hamming*** [12] :

$$w[k+1] = 0.54 - 0.46\cos\left(2\pi\frac{k}{n-1}\right) \quad \text{Avec } k = 0,\ldots, n\text{-}1 \tag{2.21}$$

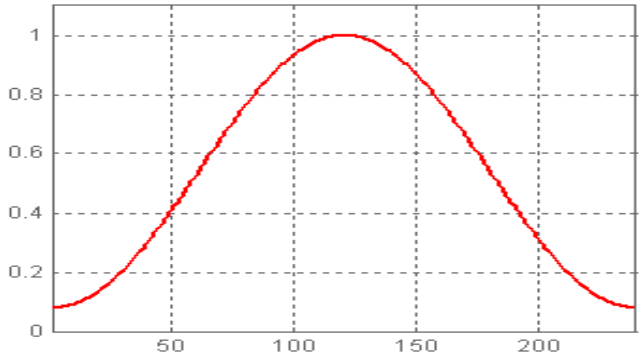

Fig.2.7 : Exemple d'une fenêtre de Hamming de 240 points

2.3.6. Transformation dans le domaine des LSP – LSF

Dans la pratique, on ne quantifie pas directement les coefficients LP, car ils ne sont pas appropriés au codage. Plusieurs transformations équivalentes ont été développées, afin de les convertir en paramètres beaucoup plus appropriés à la quantification.

2.3.6.1. Extraction des LSP

Parmi les représentations qui se sont avérées efficaces [15], les lignes de fréquences spectrales LSF (Line Spectral Frequencies) ou les lignes de raies spectrales LSP (Line Spectrum Pairs). Ces paramètres sont dérivés de la décomposition du filtre d'analyse A(z) d'ordre N, en polynômes prédicteurs symétrique P(z) et anti-symétrique Q(z) qui vérifient l'égalité :

$$A(z)=\frac{P(z)+Q(z)}{2} \tag{2.22}$$

Les paramètres LSF, qui sont liés aux zéros de polynômes dérivés de A(z), présentent un certain nombre de propriétés intéressantes. Exploitant ces propriétés, divers schémas de codage basés sur la quantification scalaire et vectorielle ont été suggérés pour la quantification efficace des paramètres LSF.

La méthode utilisée pour l'extraction des LSP est celle de Kabal et Ramachadran [17], qui utilise les polynômes de Tchebychev, c'est une méthode peu coûteuse en calcul.

On forme deux polynômes d'ordre N + 1 symétrique et antisymétrique, $P_{N+1}(z)$ et $Q_{N+1}(z)$. Ils sont donnés respectivement par la somme et la différence des filtres directs et rétrogrades.

$$P_{N+1}(z)=A(z)+z^{-(N+1)A(z^{-1})} \tag{2.23}$$

$$Q_{N+1}(z) = A(z) - z^{-(N+1)A(z^{-1})} \tag{2.24}$$

On montre que si $A(z)$ est à phase minimale [18], alors $P_{N+1}(z)$ et $Q_{N+1}(z)$ auront toutes leurs racines sur le cercle unité. D'autre part, ces racines auront comme propriétés d'être conjuguées distinctes, et alternées sur ce cercle. Faisons comme hypothèse, que l'ordre du polynôme du filtre $A(z)$ est pair ($N=2n$). $P_{N+1}(z)$ a pour racine évidente -1, et $Q_{N+1}(z)$ a pour racine évidente +1. Factorisons donc (2.23) et (2.24)

$$P_{N+1}(z) = \left(1+z^{-1}\right)\prod_{i=1}^{N/2}\left(1-2\cos(w_{2i-1})z^{-1}+z^{-2}\right) \tag{2.25}$$

$$Q_{N+1}(z) = \left(1-z^{-1}\right)\prod_{i=1}^{N/2}\left(1-2\cos(w_{2i})z^{-1}+z^{-2}\right) \tag{2.26}$$

Avec w_i étant la fréquence de la raie spectrale. Elle a comme propriété principale d'être croissante. En effet, $0 < w_1 < \cdots ... < w_{2i} < \cdots ... < w_N < \pi$ [12].

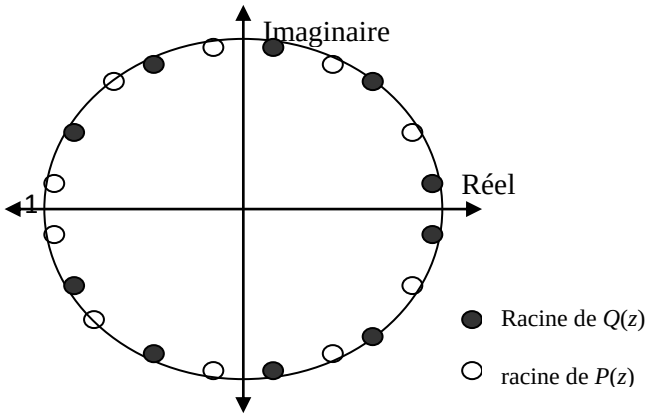

Fig.2.8 : Localisation possible des racines pour $P(z)$ et $Q(z)$ d'ordre pair

2.3.6.2. Lissage des coefficients LSP

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

La technique consiste à interpoler linéairement les coefficients LSF calculés sur une trame de durée T de façon à les appliqués, à la synthèse, sur des sous-trames de durée plus faible.

Ainsi pour deux trames d'analyse consécutives de 20ms, on interpolera avantageusement les coefficients LSF afin d'appliquer les filtres de synthèse correspondants sur des trames de durée plus courte, (Typiquement de T/8 = 2.5 ms).

De nombreuses études étaient faites, sur l'efficacité des différentes représentations des coefficients du filtre $A(z)$, pour l'interpolation ; ce sont les LSF qui fournissent en général la meilleure performance [16,19].

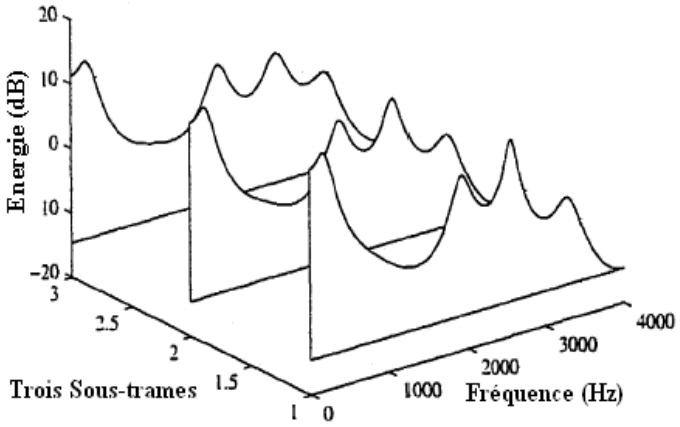

Fig.2.9 : Représentation spectrale de l'interpolation des coefficients LP.

La deuxième sous trame est le résultat de l'interpolation entre la première et la troisième sous trame.

2.3.7. Principe de la prédiction à long terme

La mise en œuvre d'une Prédiction à Long Terme LTP (Long Term Predictor), lors d'un codage par prédiction linéaire, est un moyen efficace de représenter la périodicité du signal de parole. Cette analyse n'a pas d'effet sur des trames de parole non voisées qui n'ont pas de structure harmonique. Ainsi la redondance à long terme peut être modélisée en utilisant un filtre linéaire $P_A(z)$ du premier ordre. La fonction de transfert de ce filtre est définie par :

$$\boldsymbol{P_A(z)=1-\beta.z^{-p_i}} \tag{2.27}$$

Tel que β représente le gain de prédiction, correspondant au degré de périodicité, avec

$0 \leq \beta < 1$, p_i est l'estimation en nombre d'échantillons de la période fondamentale.

Dans le domaine temporel, le filtre d'analyse de pitch soustrait un échantillon de parole retardé d'un délai estimé à partir des échantillons de la trame courante. Dans le domaine fréquentiel, le filtre d'analyse LTP enlève la structure harmonique du signal d'entrée.

2.3.8. Expansion de la largeur de bande

Concernant le problème de la sous-estimation de la largeur de bande des formants par l'algorithme de la LP classique. Il existe deux méthodes permettant d'améliorer cette estimation. La première consiste à appliquer une fenêtre de forme gaussienne, sur le signal d'auto-corrélation.

Ceci correspond à la convolution du spectre de puissance avec une fonction gaussienne et par conséquent a l'élargissement des formants.

L'autre méthode consiste à multiplier les coefficients a_k du filtre $A(z)$ par un facteur γ, avec γ typiquement compris entre 0.99413 et 0.97671. Cette multiplication a pour effet de décaler les pôles vers le centre du cercle de rayon 1 dans le plan z et donc une expansion de largeur de bande des pôles.

L'expansion de la largeur de bande peut être calculée comme suit :

$$\Delta B = -\frac{1}{\pi T} Ln(\gamma) \tag{2.28}$$

Tel que T représente la période d'échantillonnage.

2.3.9. La préaccentuation

La procédure de la conversion A/D sur un signal parole analogique a pour effet de réduire l'énergie des composantes de haute fréquence, ceci est indésirable dans l'analyse LP car une énergie relativement faible dans les hautes fréquences peut engendrer une matrice d'auto-corrélation mal conditionnée, donc il serait plus commode d'effectuer une préaccentuation avant l'analyse LP.

La préaccentuation consiste à faire passer les tranches du signal dans un filtre passe-haut du premier ordre, de transmittance

$$T(z) = 1 - \alpha . z^{-1} \tag{2.29}$$

Où α détermine la fréquence de coupure de filtre tous zéros d'ordre 1(α toujours positive inférieure à1).

Le but de cette préaccentuation est de diminuer l'influence des basses fréquences du signal et, par là, d'augmenter la précision de l'analyse LP. En effet, ce filtrage va diminuer la hauteur relative entre les formants et le reste du spectre, il va "aplatir" le spectre. Cela va permettre de calculer un filtre AR (Auto Régressive) qui collera mieux à l'enveloppe du spectre. En effet, les variations du spectre étant moindre, il sera plus facile de les suivre.

Pour éliminer l'effet de la préaccentuation, un filtre de désaccentuation est utilisé au décodeur.

$$G(z) = 1/T(z) \tag{2.30}$$

2.4. Conclusion

La parole est le moyen de changement d'information le plus important chez l'humain. Son analyse nécessite de lui attribuer le modèle qui convient.

Dans ce but nous avons décrit son mécanisme de production et remplacé ce mécanisme par des filtres électriques, formant dans leur ensemble le modèle AR ; ce dernier nous conduit au calcul LPC qui est la base de la majorité des procèdés de traitement de la parole.

L'un de ces procèdés est le codage WI (Waveform Interpolation) qui sera étudié en détail au chapitre suivant.

Chapitre 3 : Codeur WI

3.1. Introduction

Les techniques du codage de la parole sont très vastes. Leur rôle est d'effectuer la compression ou la réduction du débit binaire en enlevant la redondance, bien sur en respectant la qualité du signal.

Un système de codage de la parole comprend deux parties : le codeur et le décodeur (codec). Le codeur analyse le signal pour en extraire un nombre réduit de paramètres pertinents qui sont représentés par un nombre restreint de bits pour archivage ou transmission. Le décodeur utilise ces paramètres pour reconstruire un signal de parole synthétique.

On distingue trois catégories des codeurs :

- Codage de forme d'onde (waveform coding).

Ces codeurs ont l'avantage principal de la facilité de la mise en oeuvre, mais la qualité du signal reconstruit se dégrade rapidement avec des débits inferieurs à 16 Kbits/s.
Les codeurs qui sont réalisés dans ce sens :

  - PCM (*Pulse Code Modulation*).
  - DPCM (*Differential PCM*).
  - ADPCM (*Adaptive Differential PCM*)
  - ADM (*Adaptive Delta Modulation*).

- Codage paramétrique (parametric coding).

Ces codeurs sont destinés à fonctionner pour des bas débits et sont destinés à maintenir l'intelligibilité de la parole. Dont La plupart de ces codeurs sont basés sur le codage linéaire prédictif LP. La performance de ce type de codage dépend du modèle de production de la parole.

- Codage hydride (hybrid coding).

La qualité des codeurs de formes d'ondes chute rapidement pour des débits inférieurs à 16 Kbits/s, et comme les vocodeurs apportent une amélioration négligeable dans la qualité à des débits supérieurs à 4 Kbits/s, Les codeurs hybrides sont alors utilisés pour combler ce vide, donnant ainsi une qualité de la parole à des débits moyens. Cependant, ces codeurs ont tendance à nécessiter un nombre d'opérations plus élevé. Virtuellement, tous les codeurs hybrides reposent sur l'analyse LPC pour l'obtention des paramètres du modèle de synthèse. Les techniques de formes d'ondes utilisées pour coder le signal d'excitation et les modèles de production du pitch peuvent être incorporées pour améliorer les performances.
La figure 3.1 montre la différence de qualité de parole qui existe entre les trois types de codecs.

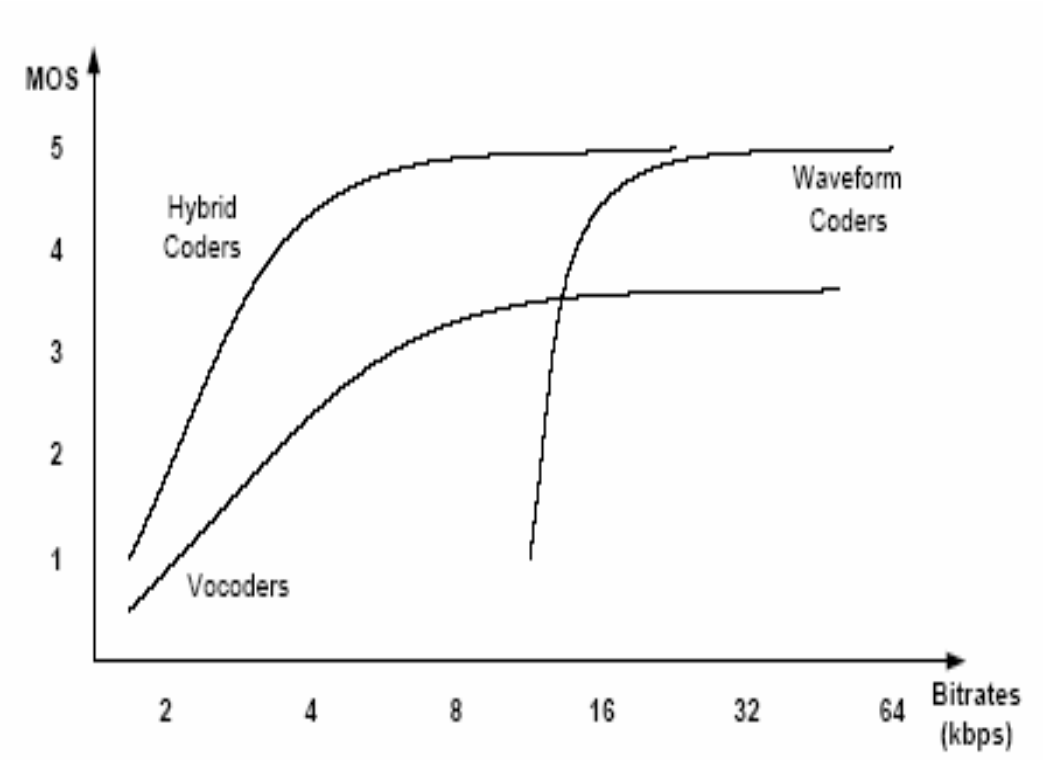

Fig.3.1 : Comparaison de la qualité du codage de la parole.

3.2. Origine et principe du codage WI

L'importance de la perception de la périodicité dans la parole voisée est à l'origine du développement de la technique de codage par interpolation de la forme d'onde. C'est l'une des méthodes qui permettent de réaliser des codeurs à un taux plus bas que 4.8 kbps, en même temps en améliorant la qualité perceptuelle de la parole. Cette technique a été introduite et développée par W. B. Kleijn[2], qui est considérée comme étant la première version et a été baptisée PWI (Prototype Waveform Interpolation). La PWI codait les segments voisés seulement et, par conséquent, elle était utilisée en combinaison avec un autre codeur tel que le CELP pour les segments non voisés.

La PWI exploite le fait que les formes d'ondes de longueur égale à la période du pitch (période fondamentale) évoluent lentement dans le temps. Cette évolution lente des formes d'ondes suggère qu'on n'a pas besoin de transmettre toutes les périodes de la trame au décodeur ; au lieu de cela, on peut les transmettre à des intervalles réguliers. Au décodeur, les formes d'ondes non transmises sont retrouvées au moyen d'une interpolation. De cette manière, le degré de Périodicité de la parole voisée sera mieux contrôlé et, par conséquent. Dans la PWI, les périodes du signal sélectionnées pour être transmises sont dites formes d'ondes prototypes (Prototype Waveforms).

Bien que la PWI travaille remarquablement bien avec les segments voisés, elle a le défaut de ne pas pouvoir être appliquée aux segments non voisés. En d'autres termes, elle doit toujours être utilisée avec une autre méthode de codage de la parole pour manipuler les segments non voisés. Ainsi, la commutation entre les codeurs devient inévitable et réduit considérablement la robustesse du codeur. En 1994, la PWI a été raffinée pour devenir la WI qui est capable de prendre en charge les sons . Similaire à la PWI, la WI représente un signal parole avec une séquence de forme d'onde. Pour la parole voisée, ces formes d'ondes sont simplement de longueurs égales à la période du pitch (pitch cycles).

Pour la parole non voisée et le bruit de fond, les formes d'ondes sont de différentes longueurs et contiennent des signaux assimilables du bruit. Puisque les formes d'ondes ne sont plus limitées à la période du pitch, il n'est plus approprié d'utiliser le terme forme d'onde prototype ou pitch-cycle. À la place, on adopte le terme forme d'onde caractéristique (Characteristic Waveform) qui sera abrégé par CW par la suite.
La différence entre la WI et la PWI est que les formes d'ondes dans la WI sont prélevées à une fréquence plus grande. Cependant, une augmentation de la fréquence de prélèvement de ces formes d'ondes entraînera une augmentation du débit. Pour contrer ce problème, la WI décompose la CW en une forme d'onde à évolution lente (SEW) *(Slowly Evolving Waveform)* et une forme d'onde a évolution rapide (REW) *(Rapidly Evolving Waveform)*. La SEW représente la composante quasi-périodique du signal parole tandis que la REW représente la composante non périodique et le bruit restants dans le signal. Puisque les deux formes d'ondes ont des propriétés différentes du point de vue perception, elles sont quantifiées séparément pour améliorer l'efficacité du codage.

La figure 3.2 présente un schéma bloc du codeur WI. On peut le diviser en deux couches : la couche d'analyse-synthèse et la couche de quantification. Dans la première couche, le bloc d'analyse exécute, d'abord, l'analyse LPC sur le signal parole entrant et fournit les paramètres pertinents et le signal résiduel. Puis, le pitch est estimé et le signal résiduel est, alors, décomposé en une suite de CW. Ces CW sont, alors, alignées et normalisées en puissance pour donner une surface (signal à deux dimensions) qui illustre l'évolution des formes d'ondes à travers la trame. L'étage de synthèse effectue l'opération inverse de celle de l'analyse. En utilisant les paramètres extraits à la synthèse Le signal résiduel est reconstruit à partir des CW et envoyé au filtre de synthèse LP où le signal parole est, finalement, reconstitué. Le commutateur permet au codeur d'éviter la couche de quantification et nous permet de mesurer la performance de la couche d'analyse-synthèse.

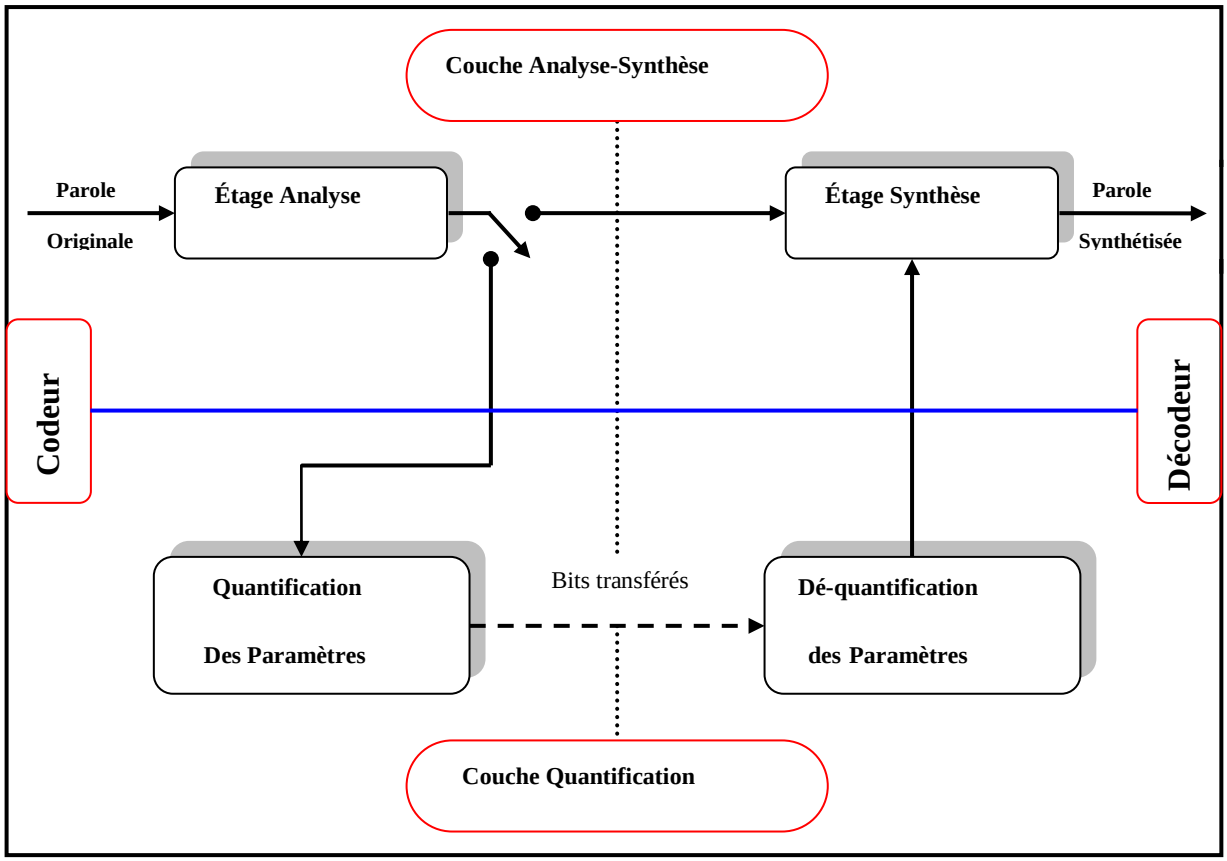

Fig.3.2 : Schéma bloc d'un système de codage WI.

3.3. L'étude des couches de codec WI

3.3.1. Le codeur

3.3.1.1. La couche d'analyse

Comme il est déjà mentionné, le but fondamental de la couche d'analyse est de décomposer le signal parole en une série d'ondes (CW) qui sera alors convertite en surface bidimensionnelle, ainsi que d'extraire d'autres paramètres orthogonaux tels que les coefficients LSF, l'énergie et le pitch. La figure 3.3 montre tous les processeurs que comprend la couche d'analyse.

Notons qu'avant tout traitement, le signal de parole d'entrée est échantillonné et quantifié sur 16 bits, avec une fréquence d'échantillonnage de 8 kHz. La taille de la trame L_f est de 160 échantillons (20 ms) et la longueur des sous-trames L_{sf} est de 20 échantillons.

Comme la WI est basée sur le modèle de codage Prédictif qui utilise une analyse par synthèse, le signal de parole est converti au signal résiduel en utilisant un filtre d'analyse d'ordre 10 [7]. Qui comporte les coefficients de prédiction linéaire LP pour chaque trame d'analyse.

Avant cela, l'opération d'analyse LP est précédée par une préaccentuation avec un facteur $\alpha = 0.6$, Cette opération a pour but de compenser la perte de l'énergie des composantes hautes fréquences due au filtrage passe–bas pendant la conversion A/D. Et une pondération par une fenêtre de hamming de longueur Lw = 240. En d'autres termes, la fenêtre couvre 120 échantillons de la trame courante et 120 de la trame future. Ces 120 échantillons futurs provoquent un retard algorithmique de 15ms. La méthode d'auto-corrélation est appliquée à cette fenêtre de parole pour générer les coefficients du filtre $\{a_k\}$. Ces derniers subiront une expansion de la largeur de bande de 60 Hz, cela consiste à prendre $\gamma = 0.976$. L'extension de la largeur de bande est très bénéfique à l'opération LP car elle assure la stabilité de ces filtres.

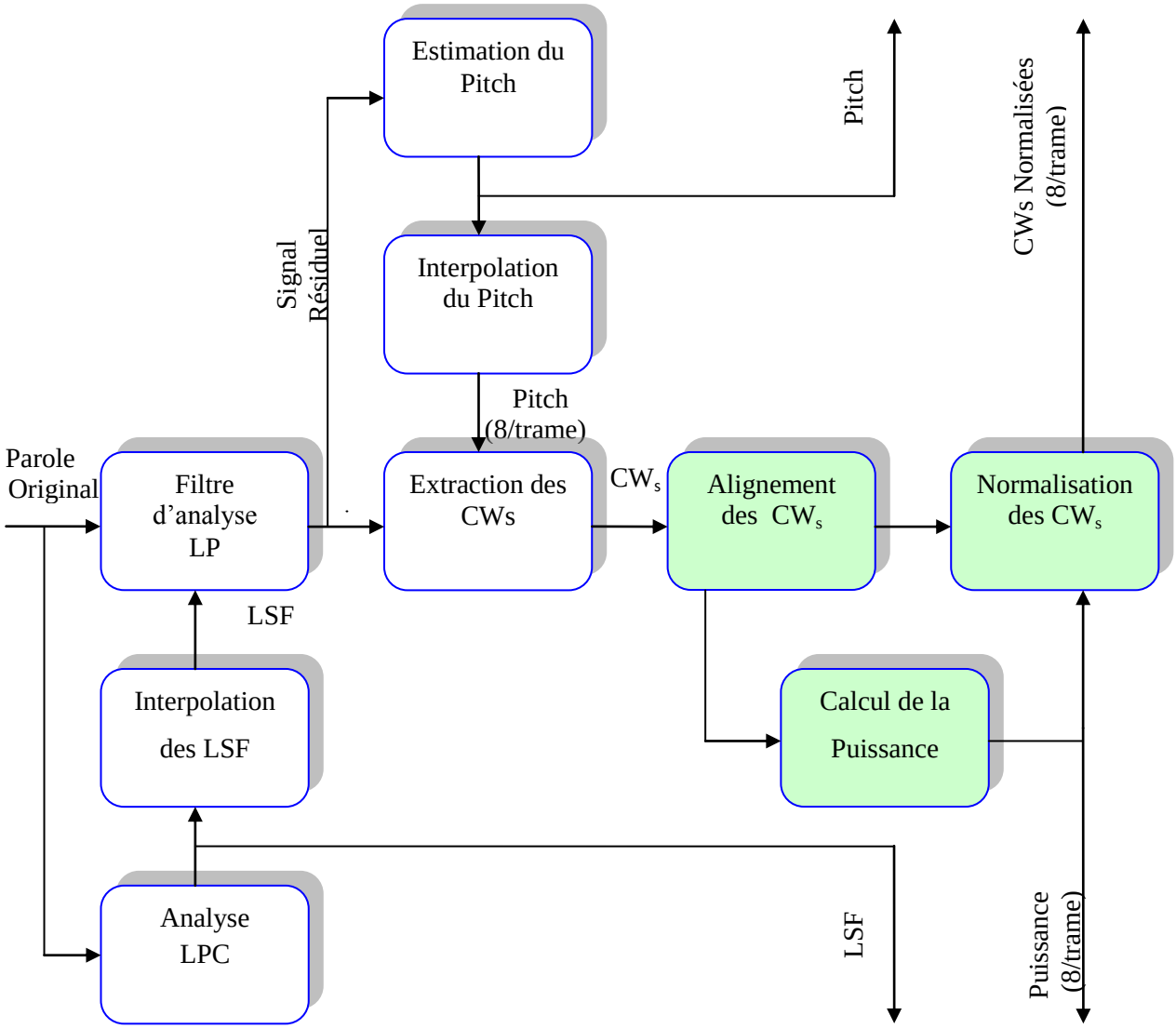

Fig.3.3 : Schéma bloc de l'étage d'analyse de la WI. Les processeurs colorés travaillent à la fréquence des sous-trames tandis que les autres travaillent à celle des trames

Les coefficients résultants sont alors convertis en coefficients LSF. Les valeurs de ces derniers, utilisés dans la détermination de l'excitation des huit sous-trames, sont obtenues par interpolation linéaire de deux ensembles de coefficients LSF. Ces coefficients sont calculés pour deux trames d'analyse successives n et n+1et forment un ensemble intermédiaire pour chacune des 8 sous-trames de la trame à coder (Figure .3.4).

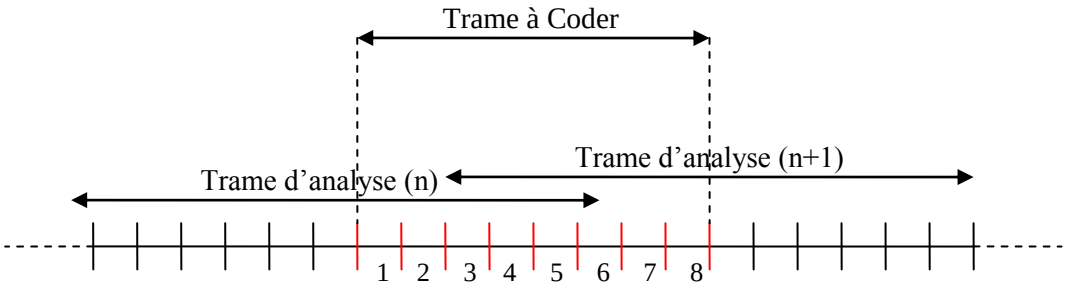

Fig.3.4 : fenêtrage des trames.

La trame d'analyse désigne la trame de parole (de 30ms) qui est pondérée par la fenêtre de Hamming et à partir de laquelle les coefficients LSF sont calculés. La trame à coder désigne la trame de parole à partir de laquelle le codage LPC est effectué. Notons que le centre de la trame d'analyse est aligné avec le centre de la première sous-trame de la trame à coder, ce qui implique que les coefficients LSF calculés à partir de cette trame d'analyse sont ceux de la première sous-trame, et les coefficients des autres sous-trames (numéro 2, 3 et… 8) sont obtenus par interpolation entre les coefficients LSF calculés à partir de la trame d'analyse courante et de la trame d'analyse suivante [4].

3.3.1.2. Détection de pitch

Les échantillons du signal résiduel (y compris les 120 échantillons de la trame future) sont envoyés au processeur qui effectue l'estimation du pitch. Dans la technique WI, la précision de l'estimateur du pitch est très cruciale pour la performance du codeur. En particulier, l'opération d'extraction au codeur et l'interpolation au décodeur reposent lourdement sur la valeur estimée du pitch.

Il existe plusieurs procédures d'estimation du pitch. Quelques unes sont basées sur la localisation des « marqueurs de pitch » (le pic dominant dans chaque période pitch du signal résiduel) tandis que d'autres sont basées sur la recherche de la position du maximum d'auto-corrélation ou du gain de prédiction pour une trame d'échantillons. Dans cette implémentation de la WI, on adopte l'algorithme tiré en [20], qui appartient à la seconde catégorie. Ce dernier est suivi de certaines modifications [31, 34] qui peuvent influencer sur la précision et le temps de calcule de pitch.

L'estimation du pitch est effectuée une fois par trame. Pour chaque trame de données, l'estimateur fait deux calculs indépendants sur deux fenêtres qui se recouvrent. La première comprend la trame courante entière (160 échantillons) et la deuxième fenêtre comprend la seconde moitié de la trame courante et la première moitié de la trame future.

Avant tout calcul, les deux trames sont sous échantillonnées de 8000 Hz à 2000 Hz, cette opération permet de diminuer le temps de calcul dans l'estimation de pitch à premier niveau qui consiste à réduire le nombre d'opérations en même temps à limiter l'intervalle de calcul du pitch. Pour sous-échantillonner un signal $s(n)$ dans un rapport M, une méthode consiste à effectuer un filtrage de gain M dans la bande $(-1/2M, +1/2M)$ suivie d'une opération de décimation de 1 sur M valeur.

La valeur du pitch estimée à premier niveau (d_{max}) et les bandes inférieures et supérieures du pitch (P_{min} ,P_{max}) sont définies par :

$$\boldsymbol{d}_{\max} = \arg\max\Big[\sum_{\boldsymbol{k}=0}^{40-\boldsymbol{d}} \boldsymbol{r_{dec}}(\boldsymbol{d})\boldsymbol{r_{dec}}(\boldsymbol{k}+\boldsymbol{d});.....5 \le \boldsymbol{d} \le 30 \Big] \tag{3.1}$$

$$\boldsymbol{P}\min = \boldsymbol{Max}(20, \boldsymbol{d}_{\max} - 3) \quad \text{Et} \quad \boldsymbol{P}\max = \boldsymbol{Min}(120, \boldsymbol{d}_{\max} + 3) \tag{3.2}$$

Sachant que $r_{dec}(.)$ représente le signal sous échantillonné du signal résiduel $r(.)$ et λ peut prendre les valeurs 1 où 2.

Ensuite ; les calculs des gains de prédiction pour toutes les valeurs possibles du retard sont faits séparément pour chaque fenêtre. Ce gain de prédiction, noté β, est défini par :

$$\beta = \boldsymbol{Max}\left\{0, \boldsymbol{Min}\left\{1, \frac{\boldsymbol{R}(\boldsymbol{d})}{\boldsymbol{R_s}(\boldsymbol{d})}\right\}\right\} \qquad \boldsymbol{P}\min \le \boldsymbol{d} \le \boldsymbol{P}\max \tag{3.3}$$

$$\boldsymbol{R}(\boldsymbol{d}) = \sum_{\boldsymbol{k}=0}^{160-\boldsymbol{d}} \boldsymbol{r}(\boldsymbol{d})\boldsymbol{r}(\boldsymbol{k}+\boldsymbol{d}) \qquad \boldsymbol{P}\min \le \boldsymbol{d} \le \boldsymbol{P}\max \tag{3.4}$$

$$\boldsymbol{dd}_{\max} = \arg\max\left[\boldsymbol{R}(\boldsymbol{d})\right] \qquad \boldsymbol{P}\min \le \boldsymbol{d} \le \boldsymbol{P}\max \tag{3.5}$$

$$\boldsymbol{R_s}(\boldsymbol{d}) = \boldsymbol{sqrt}\left(\left(\sum_{\boldsymbol{k}=0}^{160-\boldsymbol{dd}_{\max}-1} (\boldsymbol{r}(\boldsymbol{k}))^2\right) \times \left(\sum_{\boldsymbol{j}=0}^{160-\boldsymbol{dd}_{\max}-1} (\boldsymbol{r}(\boldsymbol{k}))^2\right)\right) \tag{3.6}$$

Où $R(d)$ et $Rs(d)$ représentent respectivement la fonction d'autocorrection et l'énergie du signal résiduel, d'après [31,34] on peut mettre $Rs(d)$ sous la forme :

$$R_s(d) = sqrt\left(\sum_{k=dd_{max}}^{160-1} (r(k))^4\right) \tag{3.7}$$

Cette nouvelle formule de $R_s(d)$ peut réduire le nombre d'opérations dans le calcul de β sans influencer sur la valeur finale du pitch.

Maintenant, après avoir trouvé le retard optimal pour chaque fenêtre, on utilise quelques seuils pour combiner les retards optimaux des deux fenêtres afin d'obtenir le retard le plus fiable dans la trame courante. Soit (d_0 ; β_0) le retard optimal et le gain correspondant de la première fenêtre et (d_1, β_1) ceux de la deuxième fenêtre, le retard final estimé d_{opt} est obtenu par :

si (β0 > β1 +0.4) {
  si (| d0 - d1 | >15) {
    dopt = d0 }
  sinon {
    dopt = (d0 + d1) /2
  }
sinon si (β0 > β1 +0.4 et | d0 - dernier pitch | < 7) {
  si (| d0 - d1 | >15) {
    d_{opt} = d0
  }
  sinon {
    $\boldsymbol{d_{opt}} = \arg\max\left[\sum_{\boldsymbol{k}=40}^{160-\boldsymbol{d}} \boldsymbol{r(k)r(k+d)}\right]$ Avec ((d0 + d1) /2)-1 ≤ d ≤ ((d0 + d1) /2) +1
  }
sinon {
  dopt = d1
}

3.3.1.3. Interpolation de pitch

Comme le pitch est estimé une seule fois par trame. Cependant, la WI exige une valeur de la période du pitch à chaque point d'extraction[1] pour exécuter l'extraction. Pour résoudre ce problème tout en gardant le même degré de complexité, on utilise un interpolateur de pitch pour calculer les pitchs intermédiaires. Bien qu'il existe plusieurs algorithmes d'interpolation du pitch, la technique d'interpolation linéaire classique est suffisante pour la WI.

Si on définit $P(n_1)$ et $P(n_2)$ comme étant les valeurs des pitchs aux extrémités de la trame courante telles que $n_1 < n_2$ et $n_2 - n_1 = L_f$, alors le pitch peut être linéairement interpolé par :

$$P(n) = \frac{(n_2 - n)P(n_1) + (n - n_1)P(n_2)}{n_2 - n_1} \qquad n_1 \le n \le n_2 \tag{3.8}$$

Comme le risque d'apparition d'un doublement ou d'un triplement de pitch est toujours persistant, alors la manière d'interpolation sera différente d'un cas à l'autre :

Premier Cas : $P(n_2)$ est multiple de $P(n_1)$:

$$P(n) = \begin{cases} \dfrac{C(n_2 - n)P(n_1) + (n - n_1)P(n_2)}{C(n_2 - n_1)} & si \quad n_1 \le n \le \dfrac{n_1 + n_2}{2} \\ \dfrac{C(n_2 - n)P(n_1) + (n - n_1)P(n_2)}{(n_2 - n_1)} & si \quad \dfrac{n_1 + n_2}{2} \le n \le n_2 \end{cases} \tag{3.9}$$

Où la constante C est définie comme étant le rapport $P(n_2)$ sur $P(n_1)$ arrondi au plus proche entier.

Deuxième Cas : $P(n_1)$ est multiple de $P(n_2)$:

$$P(n) = \begin{cases} \dfrac{(n_2 - n)P(n_1) + C(n - n_1)P(n_2)}{(n_2 - n_1)} & si \quad n_1 \le n \le \dfrac{n_1 + n_2}{2} \\ \dfrac{(n_2 - n)P(n_1) + C(n - n_1)P(n_2)}{C(n_2 - n_1)} & si \quad \dfrac{n_1 + n_2}{2} \le n \le n_2 \end{cases} \tag{3.10}$$

Où la constante C est définie comme étant le rapport $P(n_1)$ sur $P(n_2)$ arrondi au plus proche entier.

La figure 3.5 illustre un exemple d'une telle interpolation dans le cas d'un doublement du pitch et dans celui d'une diminution de moitié.

[1] On aura huit points d'extraction par trame

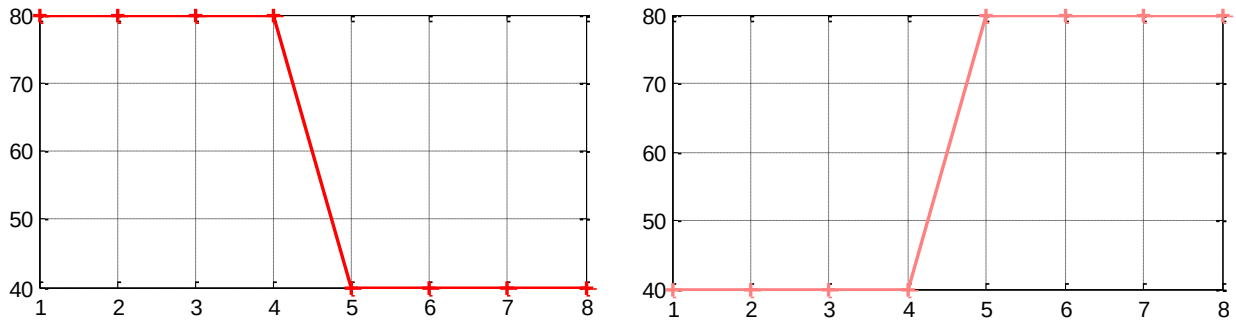

Fig.3.5 : Interpolation du pitch dans le cas d'un doublement de sa valeur (à droite), et dans le cas de diminution de la moitie (à gauche), entre 40 et 80.

3.3.1.4. Extraction des CW

Après avoir estimé et interpolé le pitch, on passe à l'extraction des CW (le processeur Extraction des CWs). L'opération d'extraction est effectuée une fois par sous-trame à une fréquence déterminée par le débit d'extraction R_{extr}. En fait, ce débit est lié aux limites de la fréquence fondamentale (donc de la période du pitch). Comme la limite inférieure de la longueur du pitch est égale à 20 échantillons, le nombre de CW à extraire dans une trame de 160 échantillons ne doit pas être inférieur à 160/20 = 8 CW.

Dans le processus d'extraction, on commence par diviser la trame courante en huit intervalles de même longueur. Le point situé sur l'extrémité droite de chaque intervalle sera un point d'extraction comme illustré dans la Figure .3.6. Par conséquent, deux points d'extraction adjacents seront séparés de 20 échantillons. Cet intervalle définit la longueur de notre sous-trame (L_{sf}).

A chaque point d'extraction, on prend le pitch interpolé dans le processeur et on forme une fenêtre d'extraction de cette longueur. La fenêtre d'extraction est centrée au point d'extraction et le signal résiduel contenu dans cette fenêtre formera notre CW extraite. Cependant, la CW extraite a toujours la longueur de la période du pitch.

Les CW sont étendues périodiquement pendant la conversion au domaine DTFS. Par conséquent, si aucune attention n'est observée vis à vis des extrémités de la CW pendant l'extraction, cela peut mener à des discontinuités importantes dans la CW périodique (à l'endroit où l'extrémité droite rencontre l'extrémité gauche). De telles discontinuités peuvent causer des distorsions audibles dans la parole reconstituée. Pour éviter cela, le point d'extraction de chaque CW est laissé libre de balayer une certaine plage ε de positions à droite et à gauche de sa position initiale.

La position qui donne la plus petite énergie du signal autour des deux extrémités de la fenêtre d'extraction est choisie. La figure .3.6 montre un exemple de l'opération d'extraction.

Dans notre implémentation, ε peut prendre des valeurs entre $-\varepsilon_{min}$ et $+\varepsilon_{max}$ = 15. Des expériences ont montré que, ε_{max} peut aller jusqu'à 16 échantillons sans affecter la qualité de la parole reconstituée.

Pour calculer efficacement l'énergie des extrémités, on crée d'autres fenêtres appelées fenêtres d'énergie des extrémités centrées sur les deux points extrémités de la fenêtre d'extraction, comme montré sur la figure .3.7. L'énergie des extrémités pour une fenêtre d'extraction est la somme des énergies des échantillons qui entourent les deux extrémités de cette fenêtre. La longueur de la fenêtre d'énergie de chaque extrémité est notée δ qu'il est suffisant de mettre égale à 10 échantillons.

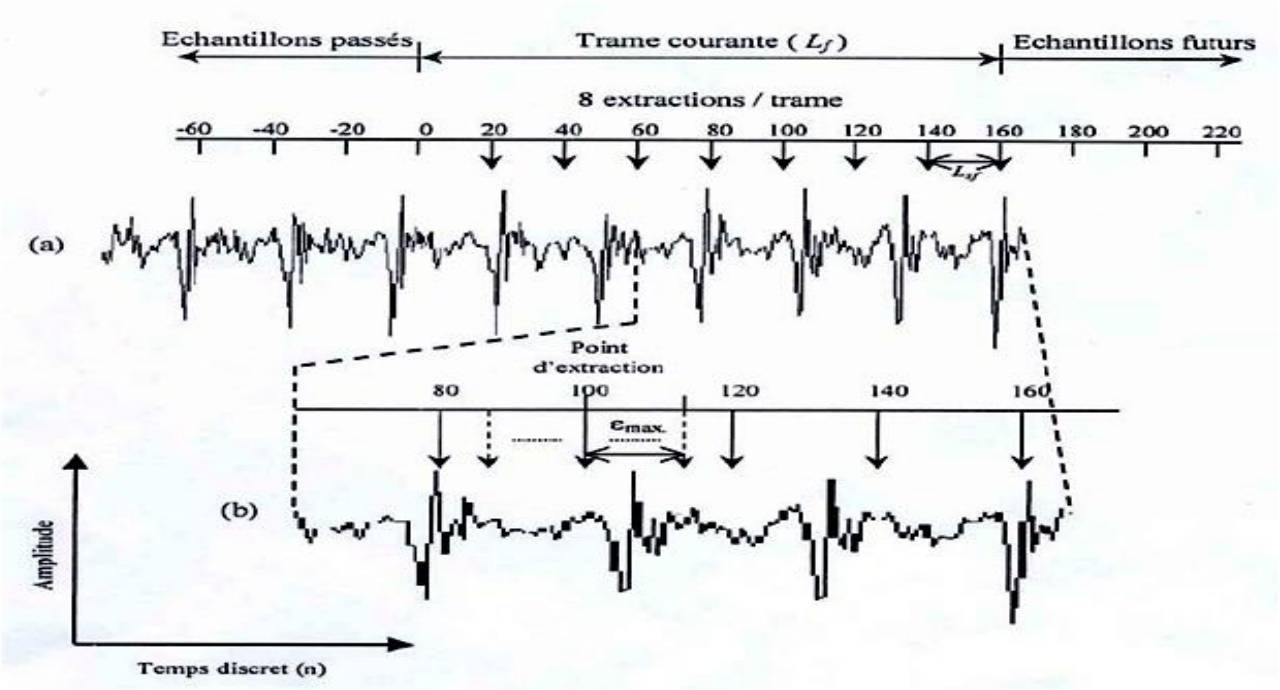

Fig.3.6 : Exemple d'un point d'extraction libre. **(a)** Les positions originales des points d'extractions des 8 CW. Chaque point d'extraction peut être déplacé légèrement jusqu'à ce que les extrémités de la fenêtre d'extraction soient dans des régions de faible énergie. **(b)** Illustration détaillée pour le point d'extraction à n=100.

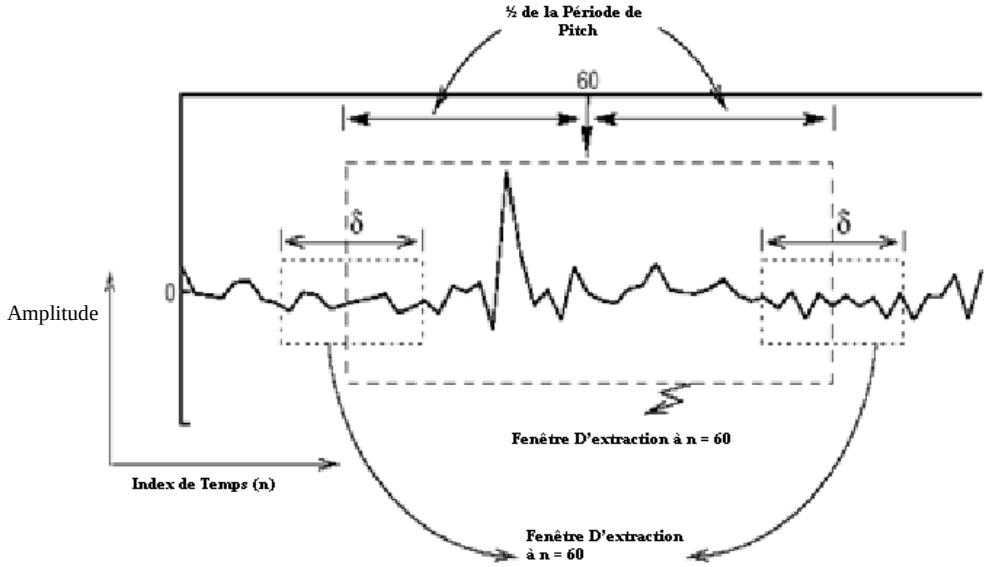

Fig.3.7 : Illustration de l'Opération d'Extraction

3.3.1.5. Représentation des formes d'ondes caractéristiques

Avant de rentrer dans les détails de chaque processeur, on commence, d'abord par choisir une représentation mathématique appropriée pour les CW. Comme on va le voir au fur et à mesure, la majorité des calculs dans la WI sont associés aux CW, il est donc crucial d'avoir la meilleure représentation des CW qui permet de réduire la complexité du codeur.

Les CW sont, finalement, utilisées pour construire une surface bidimensionnelle décrivant l'évolution des formes d'ondes du signal résiduel. Ainsi, la représentation des CW recherchée doit permettre d'avoir un signal bidimensionnel.

Pour commencer, on considère une seule CW unidimensionnelle. La CW est une séquence de valeurs réelles à temps discret de longueur égale à la période du pitch. Donnons la notation $s(m)$ à la CW de longueur P (Pitch period) :

$s(m) \in \Re$ m= 0, 1, …, P-1

Une partie du traitement dans la WI est faite dans le domaine fréquentiel. Ceci implique qu'une représentation temps- fréquence serait très favorable. Nous avons, donc, choisi la représentation en série de Fourier à temps discret (DTFS : Discrete Time Fourier Series) où *s(m)* peut être exprimée par :

$$s(m)=\sum_{k=0}^{P/2}\left[A_k\cos\left(\frac{2\pi km}{P}\right)+B_k\sin\left(\frac{2\pi km}{P}\right)\right] \qquad 0\le m<P \tag{3.11}$$

Où $\{A_k\}$ et $\{B_k\}$ sont les coefficients de Fourier à temps discret (DTFS) calculés à l'aide d'un ensemble d'équations de transformation.

Plus précisément, si P est pair :

$$\left.\begin{aligned} A_k &= \frac{2}{P}\sum_{m=0}^{P-1}\left[s(m)\cos\left(\frac{2\pi km}{P}\right)\right] \\ B_k &= \frac{2}{P}\sum_{m=0}^{P-1}\left[s(m)\sin\left(\frac{2\pi km}{P}\right)\right] \end{aligned}\right\} \quad pour \quad k=1,2,\ldots,P/2-1$$

$$\left.\begin{aligned} A_k &= \frac{1}{P}\sum_{m=0}^{P-1}\left[s(m)\cos\left(\frac{2\pi km}{P}\right)\right] \\ B_k &= \frac{1}{P}\sum_{m=0}^{P-1}\left[s(m)\sin\left(\frac{2\pi km}{P}\right)\right] \end{aligned}\right\} \quad pour \quad k=0 \quad et \quad P/2 \tag{3.12}$$

Quand P est impair :

$$\left.\begin{aligned} A_k &= \frac{2}{P}\sum_{m=0}^{P-1}\left[s(m)\cos\left(\frac{2\pi km}{P}\right)\right] \\ B_k &= \frac{2}{P}\sum_{m=0}^{P-1}\left[s(m)\sin\left(\frac{2\pi km}{P}\right)\right] \end{aligned}\right\} \quad pour \quad k=1,2,\ldots,(P-1)/2$$

$$\left.\begin{aligned} A_k &= \frac{1}{P}\sum_{m=0}^{P-1}\left[s(m)\cos\left(\frac{2\pi km}{P}\right)\right] \\ B_k &= \frac{1}{P}\sum_{m=0}^{P-1}\left[s(m)\sin\left(\frac{2\pi km}{P}\right)\right] \end{aligned}\right\} \quad pour \quad k=0 \tag{3.13}$$

La forme d'une CW peut, maintenant, être décrite par un ensemble de coefficients DTFS$\{A_k, B_k\}$. Notons que l'indice *m* dans (3.11) n'est pas nécessairement entier; il peut prendre n'importe quelle valeur réelle dans l'intervalle $0 \le m < P$. En d'autres termes, les valeurs situées entre deux instants discrets (s(2.3), par exemple) peuvent être calculées aisément par (3.11).

Après avoir obtenu la représentation pour une CW, nous somme, maintenant, prêts à construire une représentation bidimensionnelle pour une séquence de CW. En fait, cette représentation est simplement obtenue en ajoutant une modification à (3.11). Ainsi, on attache un indice de temps discret n à tous les paramètres dans (3.11) qui varient dans le temps. Ces paramètres sont A_k, B_k et P.

L'équation (3.11) peut donc être écrite comme suit :

$$\boldsymbol{s(n,m)=\sum_{k=1}^{M}\left[A_k(n)\cos\left(\frac{2\pi km}{P}\right)+B(n)\sin\left(\frac{2\pi km}{P}\right)\right]} \qquad (3.14)$$

$$\boldsymbol{P==P(n)==P(n-1)}$$
$$\boldsymbol{M==[P(n)/2]==[P(n-1)/2]}$$

P représente la longueur (pitch) des CW et *M* est le nombre d'harmoniques du spectre.

où les coefficients { A_k (*n*)}et { B_k (*n*)}sont, maintenant, variants dans le temps de même que la valeur du pitch P(*n*). Il faut noter que nous avons ignoré les coefficients A_0et B_0 dans l'équation (l'indice k commence à partir de k = 1 au lieu de k = 0). Ceci est du au fait que B_0 dans (3.12) et (3.13) est un coefficient redondant (sin(0)=0). D'un autre côté, A_0 représente la composante DC du signal et n'a aucune importance vis à vis de la perception. Par conséquent, ces deux coefficients peuvent être ignorés.

L'équation (3.14) est, à présent, la représentation d'un signal bidimensionnel où m et n sont les variables courantes. Chaque CW évolue le long de l'axe *m* et la forme des CW évolue à travers le temps le long de l'axe *n*.

Cependant, la longueur de la CW dans (3.14) dépend du pitch P(*n*) variant dans le temps ; les CW à des instants différents peuvent avoir des longueurs différentes. Il est, généralement, plus convenable de normaliser toutes les CW à une longueur commune.

Cette normalisation peut être accomplie en substituant :

$$\phi=\phi(m)=\frac{2\pi m}{P(n)} \qquad (3.15)$$

Dans (3.14) et on peut obtenir :

$$S(n,\phi)=\sum_{k=1}^{P/2}\left[A_k(n)\cos(k\phi)+B_k(n)\sin(k\phi)\right] \qquad 0\le\phi(.)<2\pi \qquad (3.16)$$

De cette manière, toutes les CW ont la même longueur 2л. La figure .3.8 donne une illustration de cette normalisation et un exemple d'une surface bidimensionnelle.

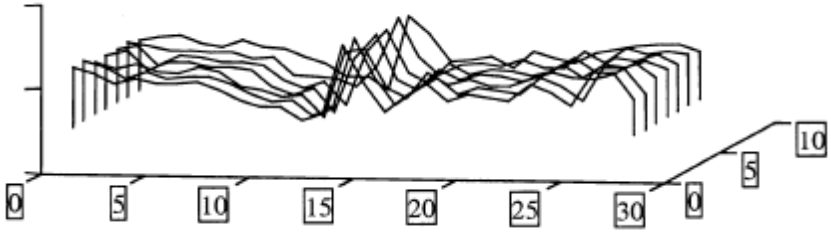

Fig.3.8 : Normalisation.

3.3.1.6. Alignement des CW

La procédure d'extraction donne une description en DTFS pour chaque CW. En général, ces CW ne sont pas en phase, ceci dit, les caractéristiques principales dans les formes d'ondes ne sont pas alignées. Afin d'avoir une description précise des CW et de leur évolution dans la trame, on doit établir un alignement de ces CW. Cet alignement est réalisé pour deux CW successives (la CW courante et la CW précédente). La procédure consiste à aligner la CW courante avec celle précédente en introduisant un décalage temporel circulaire à la trame courante [13], ce décalage temporel circulaire est, en réalité, équivalent à l'addition d'une phase linéaire aux coefficients DTFS.

La figure .3.8 montre un schéma bloc de la procédure d'alignement.

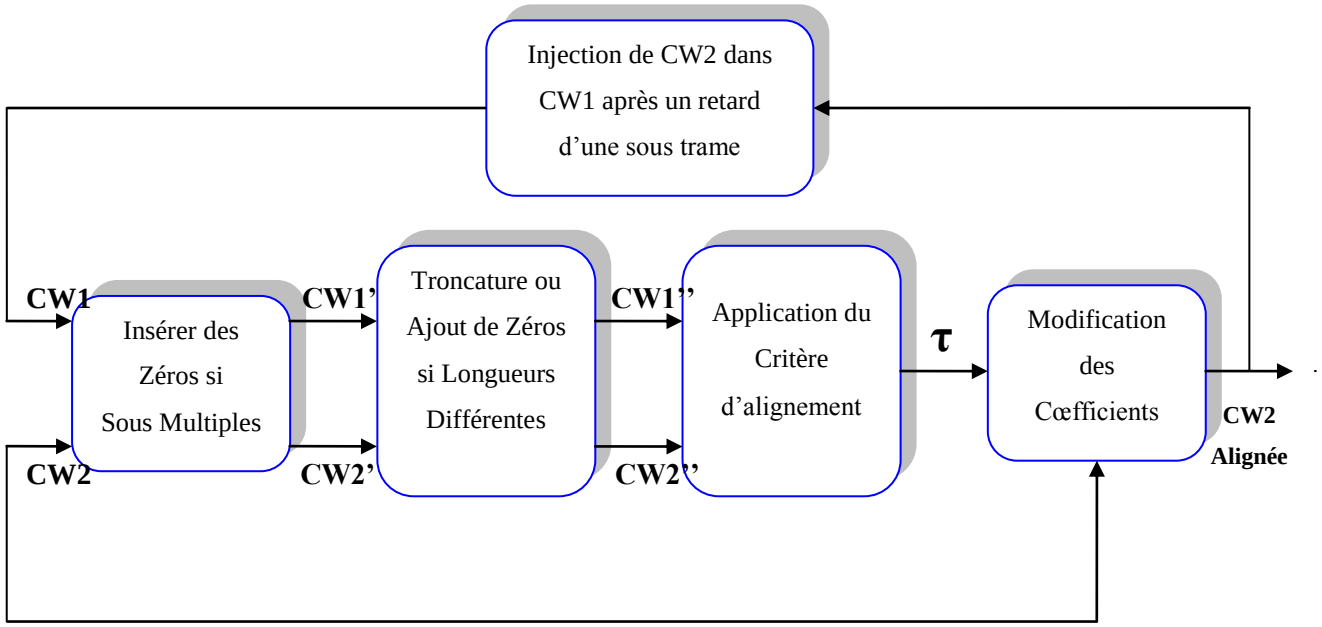

Fig.3.9 : Schéma bloc de la procédure d'alignement.

Puisque les CW n'ont pas toujours la même longueur, ce qui nous mène à étudier chacun des cas [6] :

Premier Cas : Les deux CW ont la même longueur

Les représentations en DTFS des deux CW successives sont :

$$s(n_0,m)=\sum_{k=0}^{M}\left[A_k(n_0)\cos\left(\frac{2\pi km}{P}\right)+B(n_0)\sin\left(\frac{2\pi km}{P}\right)\right] \tag{3.17}$$

$$s(n_1,m)=\sum_{k=0}^{M}\left[A_k(n_1)\cos\left(\frac{2\pi km}{P}\right)+B(n_1)\sin\left(\frac{2\pi km}{P}\right)\right]$$

Où n_0 et n_1 sont les positions dans le temps, respectivement des CW précédente et présente et n_1- $n_0 = L_{sf}$.

Supposons, maintenant, qu'un décalage circulaire de T échantillons est appliqué à la CW courante, $s(n_1,m)$ devient :

$$\boldsymbol{s(n_1,m-T)=\sum_{k=0}^{M}\left[A_k(n_1)\cos\left(\frac{2\pi k(m-T)}{P}\right)+B_k(n_1)\sin\left(\frac{2\pi k(m-T)}{P}\right)\right]} \tag{3.18}$$

II est clair que le décalage circulaire T dans le temps est équivalent à l'adition d'une phase Linéaire $2\pi T/P$ dans le domaine DTFS. Pour trouver la valeur du décalage temporel T nécessaire à l'alignement de CW_1 avec CW_0, on doit maximiser l'inter-corrélation entre les deux CW, ainsi le décalage T est défini par :

$$\boldsymbol{T=\arg\max_{0\le T'\le P}\sum_{k=0}^{M}\left\{\begin{array}{l}\left[A_k(n_0)A_k(n_1)+B_k(n_0)B_k(n_1)\right]\cos\left(\frac{2\pi kT'}{P}\right)+\\ \left[B_k(n_0)A_k(n_1)-B_k(n_1)A_k(n_0)\right]\sin\left(\frac{2\pi kT'}{P}\right)\end{array}\right\}} \tag{3.19}$$

Si en suppose que $\tau=2\pi T/P$, τ représente le décalage normalisé alors on obtient :

$$\tau = \arg\max_{0\le\tau'\le P} \sum_{k=0}^{M} \begin{Bmatrix} [A_k(n_0)A_k(n_1)+B_k(n_0)B_k(n_1)]\cos(k\tau')+ \\ [B_k(n_0)A_k(n_1)-B_k(n_1)A_k(n_0)]\sin(k\tau') \end{Bmatrix} \quad (3.20)$$

Un avantage immédiat de l'exécution de l'alignement dans le domaine DTFS est que cela permet un alignement fractionnel sans calcul additionnel tout en évitant les sur-échantillonnages et sous-échantillonnage conventionnels. Cet alignement fractionnel se fait à n'importe quelle résolution désirée. Pour une fréquence d'échantillonnage de 8000 Hz donne de bons résultats.

La prochaine étape dans l'alignement consiste à incorporer le décalage temporel τ dans les coefficients DTFS de la CW courante. Cela se fait en développant les sinus et cosinus des équations (12) et (13) en utilisant les identités trigonométriques fondamentales. On obtient un nouvel ensemble de coefficients DTFS de la CW décalée.

$$\left.\begin{aligned} A'_k(n_1) &= A_k(n_1)\cos(k\tau)-B_k(n_1)\sin(k\tau) \\ B'_k(n_1) &= A_k(n_1)\sin(k\tau)+B_k(n_1)\cos(k\tau) \end{aligned}\right\} \quad \text{Pour } k = 1, 2,\ldots, M \quad (3.21)$$

Deuxième Cas : Les deux CW ont des longueurs différentes

Dans ce cas le critère d'alignement (3.20), qui est basé sur la supposition d'égalité de dimension, n'est plus applicable directement. Donc on précède l'application de critère d'alignement, d'un pré - traitement qui consiste à :

- dans le domaine fréquentiel, on tronque la CW la plus longue jusqu'à ce qu'elle ait la même longueur que l'autre.
- dans le domaine fréquentiel, on remplit de zéros la plus courte CW jusqu'à ce qu'elle ait la même longueur que l'autre.

La figure suivante montre l'exemple d'une fonction sin(x) contractée et étirée dans le temps.

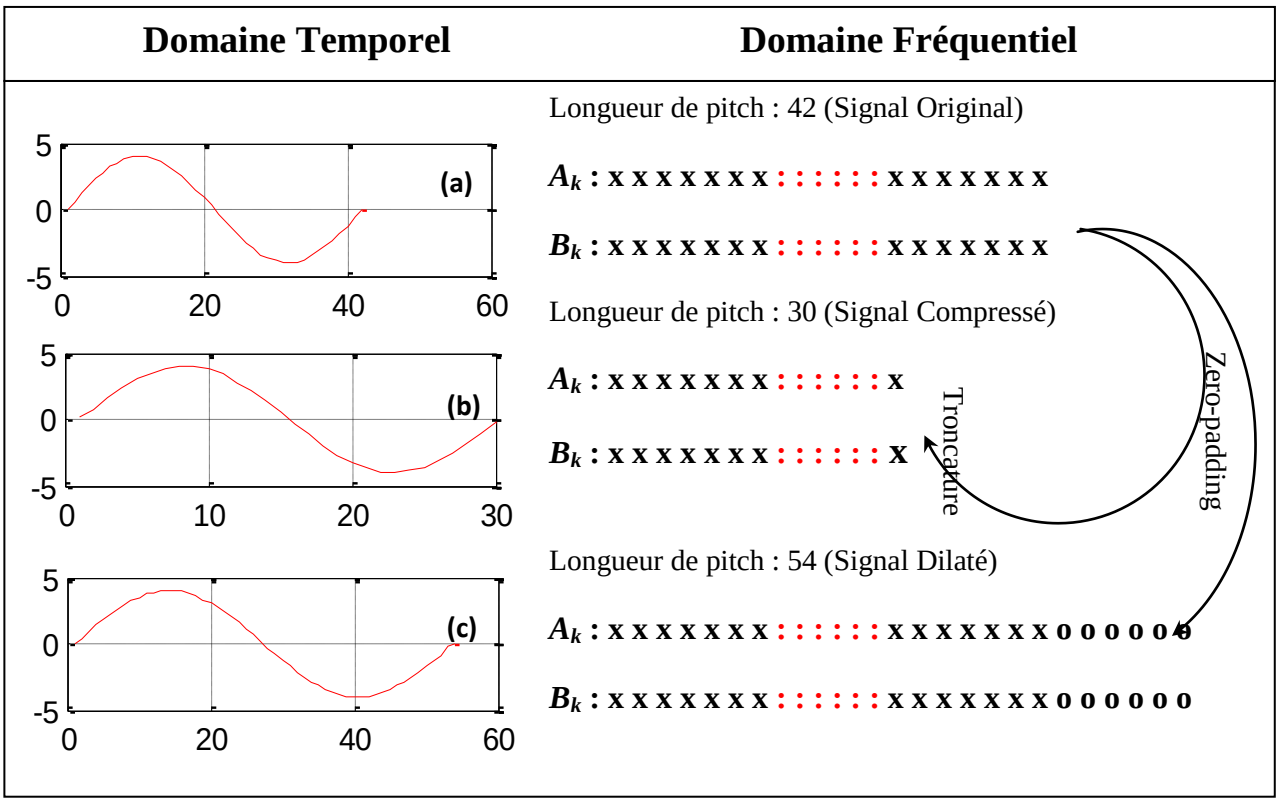

Fig.3.10 : Échelonnage temporel des CW. (a) la fonction sin(x) de longueur 42 échantillons et ses coefficients DTFS correspondants. (b) La version tronquée de longueur 30 échantillons. La forme temporelle globale est préservée mais les détails sont plus ou moins perdus, (c) La version étirée de longueur 56 échantillons. Cet étirement n'introduit aucune information nouvelle à la séquence temporelle, mais il offre une meilleure résolution.

Troisième Cas : Une des CW est de longueur multiple de l'autre

Comme on l'a déjà mentionné, des périodes de pitch multiples, ou sous-multiples peuvent apparaître dans une CW extraite. Afin d'éviter les complications dans le processus d'alignement, le prétraitement consistera donc à l'insertion d'harmoniques d'amplitude nulle entre les harmoniques de la plus courte CW. La figure 3.10 montre comment les zéros sont insérés entre les coefficients DTFS et le résultat correspondant dans le domaine temporel.

Pour détecter l'apparition de sous-multiple du pitch, on opère de la même manière que celle du paragraphe 3.3.1.3 en utilisant l'indicateur *C*. Si cet indicateur est différent de l'unité, alors, on insère *C*-1 zéro entre chaque deux coefficients DTFS.

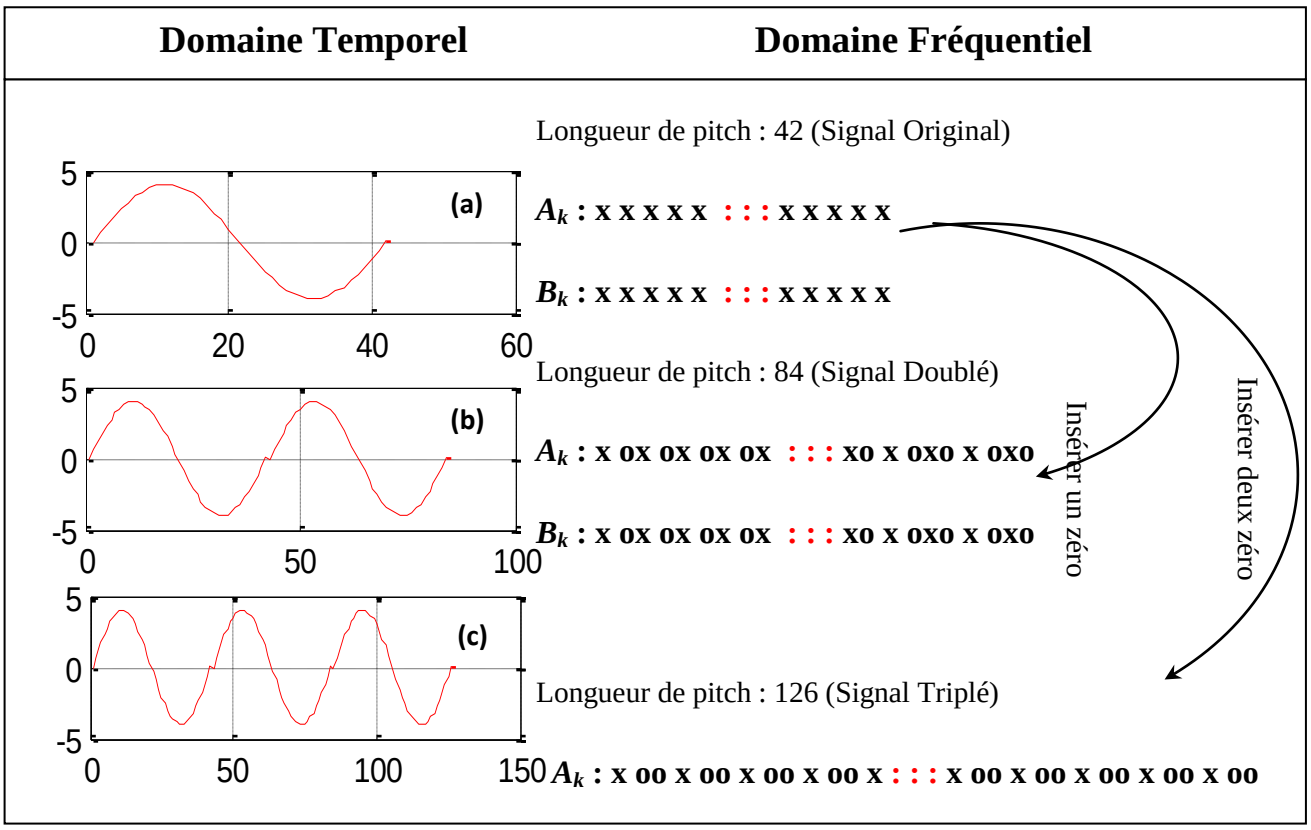

Fig.3.11 : Illustration de l'insertion de zéros entre les composantes spectrales. (a) Une fonction sin(x) de longueur 42 échantillons (b) La forme d'onde de (a) est dupliquée une fois après insertion d'un zéro entre deux harmoniques adjacentes, (c) La forme d'onde de (a) est dupliquée deux fois après insertion de deux zéros entre chaque deux harmoniques adjacentes.

La figure ci-dessous montre un exemple d'une séquence de CW alignées.

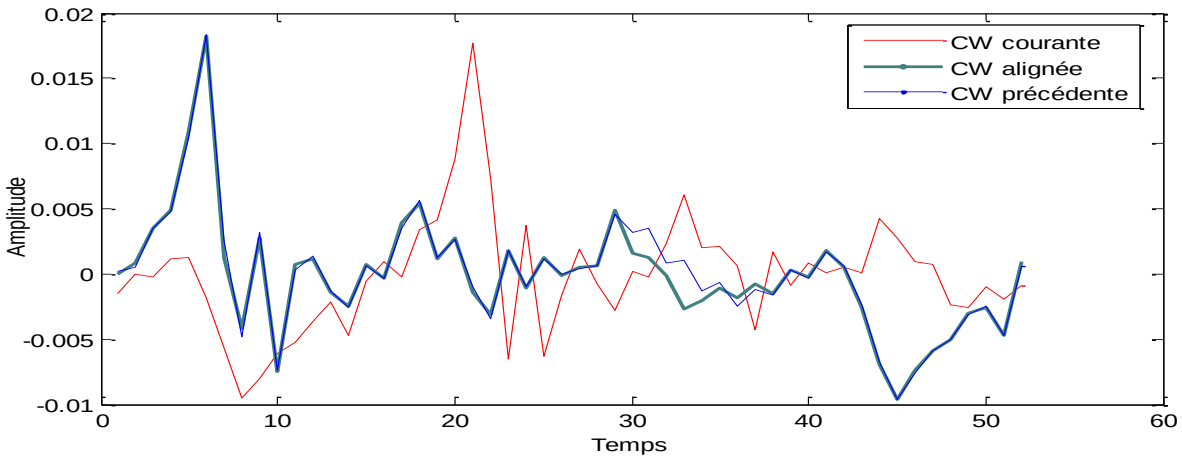

Fig.3.12 : Exemple du processus d'alignement pour deux CW adjacentes.

3.3.1.7. Normalisation des CW

La puissance d'une CW est définie, comme étant l'énergie moyenne par échantillon sur une période du pitch. Ainsi, la relation entre une CW normalisée et sa version non normalisée est exprimée en termes de puissance. Le but principal de cette normalisation est de séparer la puissance et la forme des CW, afin de les quantifier séparément, ainsi d'avoir une meilleure efficacité du codage.

Puisque toutes les CW ont déjà été converties en coefficients DTFS, le calcul de la puissance moyenne d'une CW à l'instant *n*, notée $\Psi(n)$ Peut être exprimée par :

$$\Psi(n)=\frac{1}{P(n)}\sum_{m=0}^{P(n)-1}\left|s(n,m)\right|^2 \tag{3.22}$$

Où $P(n)$ est la longueur de la CW. En remplaçant $s(n,m)$ par ses coefficients DTFS, on obtient :

$$\begin{aligned}\Psi(n)&=\frac{1}{P(n)}\sum_{m=0}^{P(n)-1}s(n,m).s^*(n,m)\\ \Psi(n)&=\frac{1}{P(n)}\sum_{m=0}^{P(n)-1}s(n,m)\sum_{k=0}^{P(n)/2}\left[A_k^*\cos\left(\frac{2\pi km}{P(n)}\right)+B_k^*\sin\left(\frac{2\pi km}{P(n)}\right)\right]\end{aligned} \tag{3.23}$$

Puisqu'on fait le traitement pour une seule position *n*. $\Psi(n)$ devient :

$$\Psi(n)=\left[\frac{1}{P}\sum_{k=0}^{P/2}A_k\sum_{m=0}^{P-1}s(m)\cos\left(\frac{2\pi km}{P}\right)\right]+\left[\frac{1}{P}\sum_{k=0}^{P/2}B_k\sum_{m=0}^{P-1}s(m)\cos\left(\frac{2\pi km}{P}\right)\right] \tag{3.24}$$

En utilisant les règles de conversion en domaine DTFS on aura :

$$\Psi(n)=\begin{cases}\frac{1}{2}\sum_{k=0}^{(P/2)-1}\left(A_k^2+B_k^2\right)+A_{P/2}^2+B_{P/2}^2 & si\ \ P\ pair\\ \frac{1}{2}\sum_{k=0}^{(P/2)}\left(A_k^2+B_k^2\right) & si\ \ P impair\end{cases} \tag{3.25}$$

L'équation (3.25) est la formule utilisée, pour déterminer la puissance de la CW à partir de ses coefficients DTFS. Donc, la normalisation consiste à diviser chaque coefficient DTFS, par la racine carrée de la puissance moyenne.

3.3.2. Le décodeur WI

3.3.2.1. Vue générale de décodeur

L'étage d'analyse décompose un segment de parole en quatre paramètres, le pitch, les coefficients LSF, les puissances et les CW. Les deux premiers ont une fréquence de calcul égale à celle des trames, tandis que les deux derniers sont calculés une fois par sous-trame. A partir des LSF, pitch, puissances et CW normalisées, le signal parole peut être reconstitué dans le processus de synthèse. D'autre part, si le codeur travaille avec la couche de quantification, le bloc de synthèse reçoit les versions quantifiées de ces paramètres.

Le schéma bloc de l'étage de synthèse est donné dans la figure 3.12. Similaire aux processeurs du codeur, la fréquence d'exécution varie d'un bloc à un autre dans la couche de synthèse.

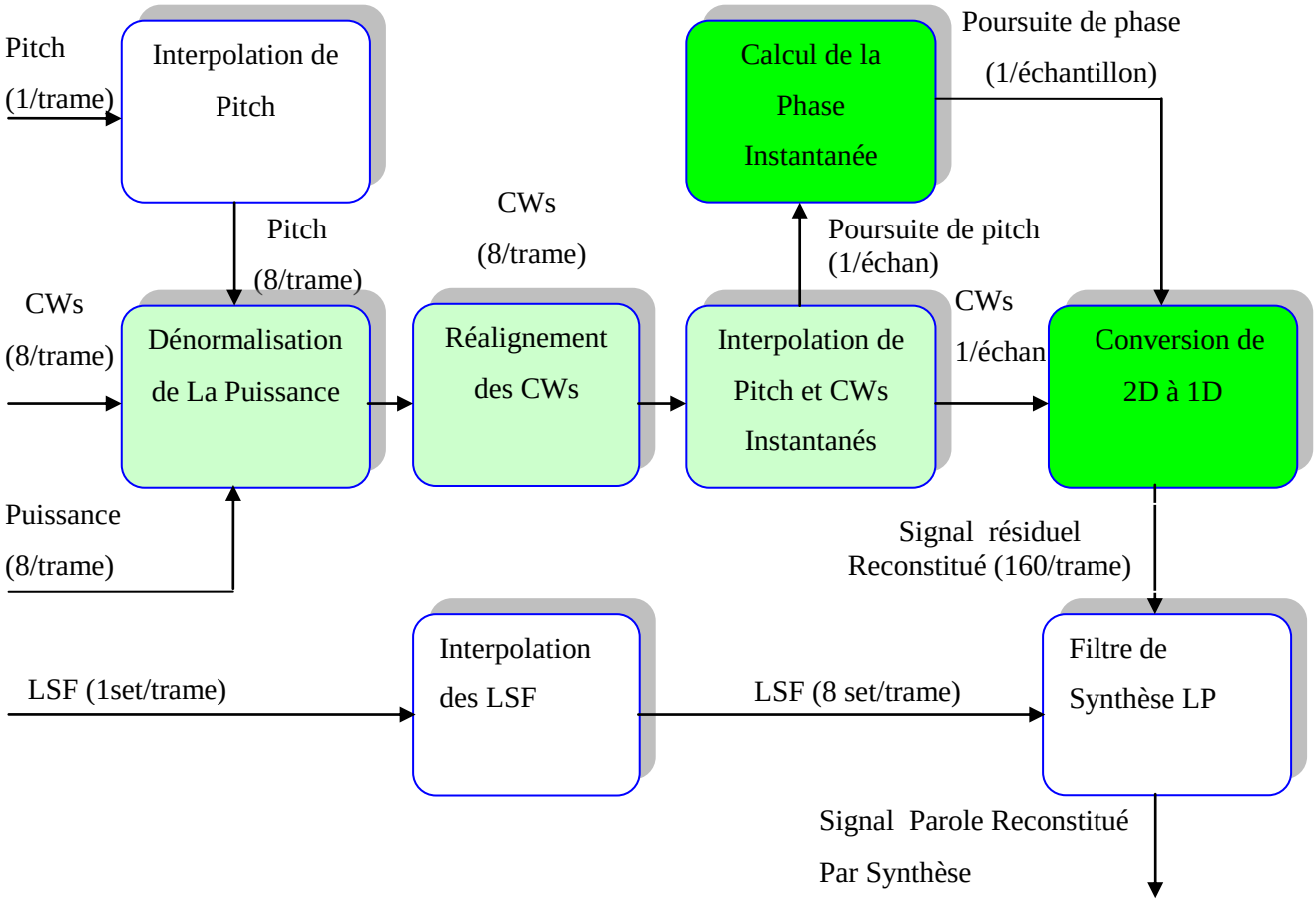

Fig.3.13 : Schéma bloc d'un décodeur WI. Les processeurs colorés en vert clair sont exécutés une fois par sous- trame tandis que ceux colorés en vert foncé sont exécutés à la fréquence des échantillons. Les autres sont exécutés une fois par trame [6].

3.3.2.2. Génération des pitchs et CW instantanés

Après la dénormalisation des CW qui consiste à multiplier chaque coefficient DTFS par la racine carrée de la puissance. Les CW successives peuvent ne plus être bien alignées une fois déquantifiées, ce qui nécessite le réalignement des formes d'ondes.

Maintenant, nous avons une CW reconstruite et alignée dans chaque sous-trame. Dans la technique WI, il est nécessaire d'avoir une CW et une valeur du pitch à chaque point d'échantillonnage pour reconstruire le signal résiduel unidimensionnel.

Une interpolation linéaire peut servir à sur-échantillonner les CW. Quand ce sur-échantillonnage est exécuté entre deux CW de même longueur, une interpolation directe est appliquée. Cependant, si les CW ont des dimensions différentes, des calculs supplémentaires seront nécessaires, pour assurer une bonne interpolation. L'interpolation est linéaire mais n'emploie pas les équations (3.9) et (3.10) de pitch sous-multiple. Il faut bien s'assurer que les valeurs de pitch générées dans cet interpolateur correspondent aux longueurs des CW instantanées.

La figure 3.14 montre le schéma bloc de l'interpolateur qui peut prendre en charge l'interpolation des CW et du pitch dans les trois cas possibles : (I) dimensions égales, (II) dimensions différentes et (III) dimensions sous-multiples du pitch.

Premier Cas : interpolation avec longueurs égales

Si on note par n_0 et n_1 les instants des extrémités de l'intervalle d'interpolation, alors, la CW instantanée $s(n, m)$ à l'instant n peut être calculée par interpolation entre $s(n_0 , m)$ et $s(n_1 , m)$ Dans le domaine temporel, cette opération est exprimée par :

$$s(n,m)=\left(\frac{n_1-n}{n_1-n_0}\right)s(n_0,m)+\left(\frac{n-n_0}{n_1-n_0}\right)s(n_1,m) \qquad n_0 \le n \le n_1,\ \ 0 \le m \le P \tag{3.26}$$

En remplaçant $s(n ,m)$ par ces coefficients DTFS, on obtient :

$$\left.\begin{aligned} A_k(n)&=\left(\frac{n_1-n}{n_1-n_0}\right)A_k(n_0)+\left(\frac{n-n_0}{n_1-n_0}\right)A_k(n_1) \\ B_k(n)&=\left(\frac{n_1-n}{n_1-n_0}\right)B_k(n_0)+\left(\frac{n-n_0}{n_1-n_0}\right)B_k(n_1) \end{aligned}\right\} \quad \text{Pour } k = 1 , 2 , \dots , [P/2] \tag{3.27}$$

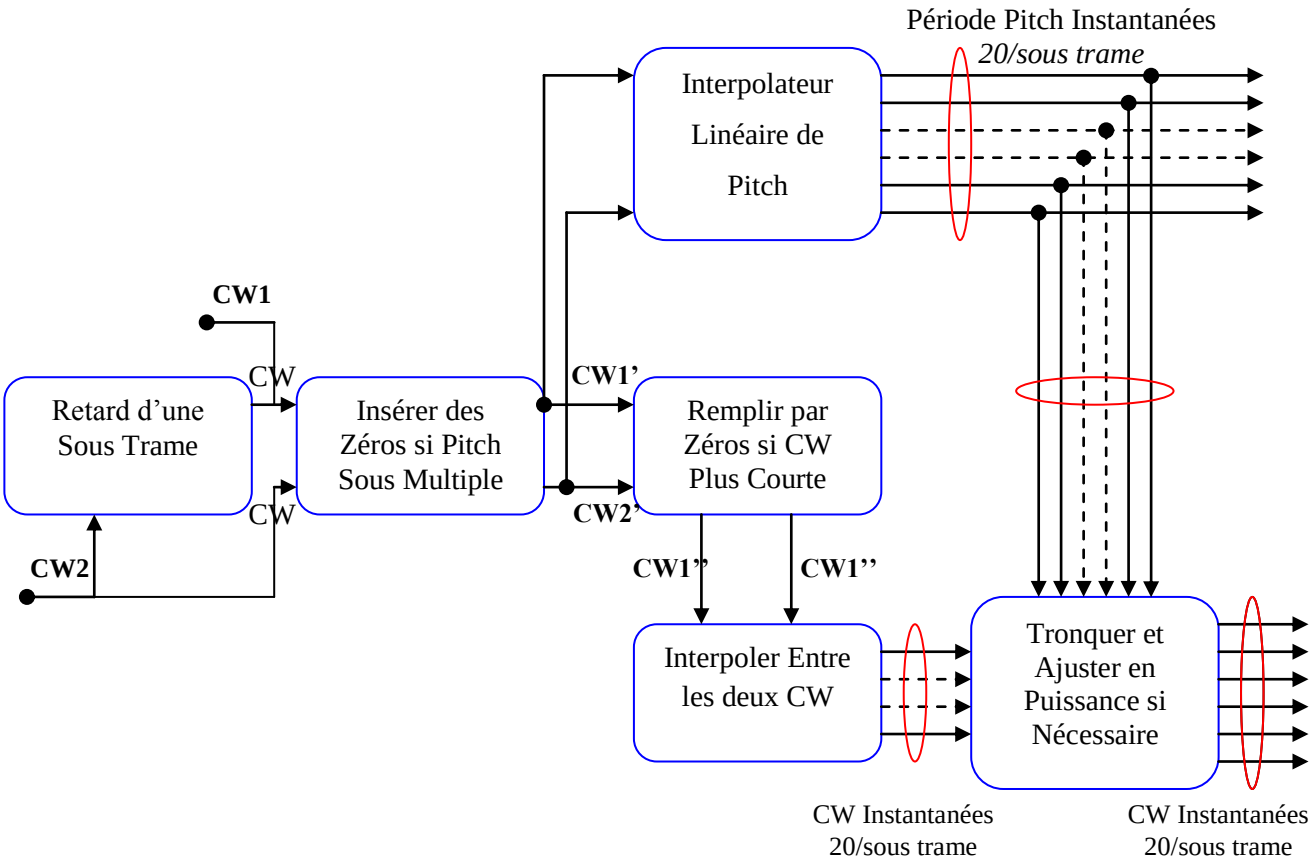

Fig.3.14 : Schéma bloc du processeur d'interpolation.

En d'autres termes, l'interpolation linéaire entre les deux CW dans le temps est équivalente à celle de leurs coefficients DTFS. L'interpolation est exécutée une fois par sous–trame. Puisque les deux CW sont de même longueur, les CW interpolées auront la même longueur également. Par conséquent, on aura un contour constant du pitch interpolé [13].

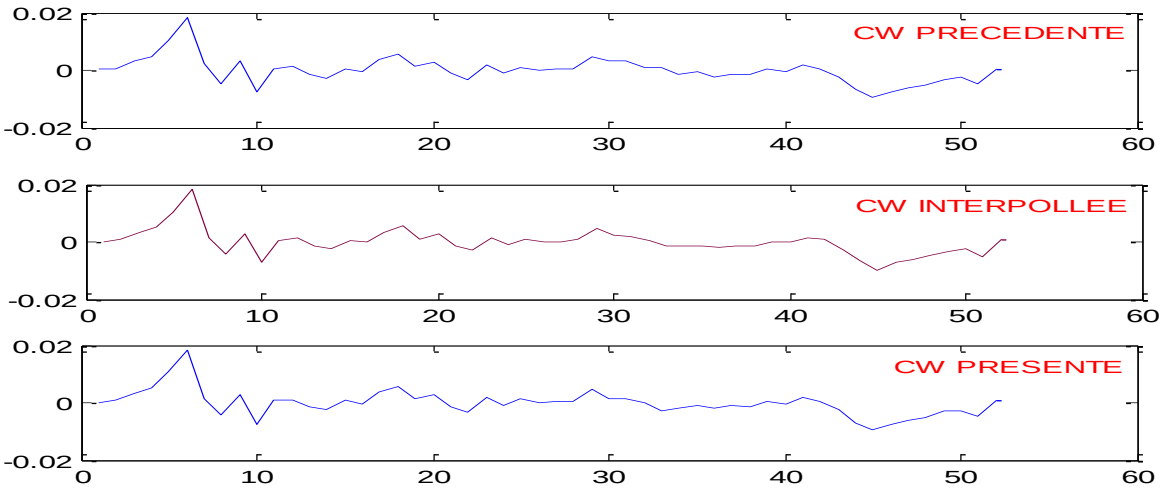

Fig.3.15 : Illustration du processus d'interpolation d'une CW au douzième échantillon.

Deuxième Cas : interpolation avec longueurs différentes

Pour faciliter l'interpolation dans un cas pareil, on peut allonger dans le temps la plus petite CW, pour qu'elle ait la même longueur que la plus longue avant de passer à l'interpolation. Comme déjà fait au paragraphe (alignement)

Ainsi, l'équation d'interpolation linéaire conventionnelle (3.8) peut être utilisée pour sur-échantillonner le pitch. Cependant, les valeurs du pitch sur-échantillonnées résultant peuvent ne pas coïncider avec les longueurs des CW interpolées. Pour éviter un tel problème, on fait coïncider les longueurs des CW avec le contour du pitch avec un ajustement en puissance, pour les CW tronquées.

Troisième Cas : interpolation avec des longueurs sous-multiples du pitch

Si la CW courante est considérablement plus longue ou plus courte que la précédente, cela implique que le pitch actuel est certainement multiple ou sous-multiple, respectivement, du précédent. Comme dans le paragraphe (3.3.1.3) on utilise l'indicateur *C* comme détecteur de sous-multiple de pitch.

Les *C*-1 zéros sont insérés entre les coefficients DTFS, afin d'avoir compensé l'écart de longueur entre les deux CW ; ensuite les CW sont traitées de la même manière que dans le premier cas. La figure 3.15 montre un exemple d'interpolation des CW sur un intervalle d'une sous trame.

3.3.2.3. Estimation de la phase instantanée

Après l'interpolation des CW à la fréquence d'un échantillon, maintenant l'objectif est de convertir les valeurs du pitch en une poursuite de phases instantanées. Ce contour de la phase sera utilisé pour retrouver le signal résiduel unidimensionnel à partir de la surface bidimensionnelle des CW.

Si on désigne par $\phi(.)$ le contour de la phase ; la phase en chaque point d'échantillonnage peut être calculée par l'équation suivante [7]:

$$\boldsymbol{\phi(n)=\phi(n-1)+\int_{n-1}^{n}\frac{2\pi}{P(n')}\,dn'} \qquad (3.28)$$

Où $\phi(n)$ et $\phi(n\text{-}1)$ sont respectivement les phases courante et précédente.

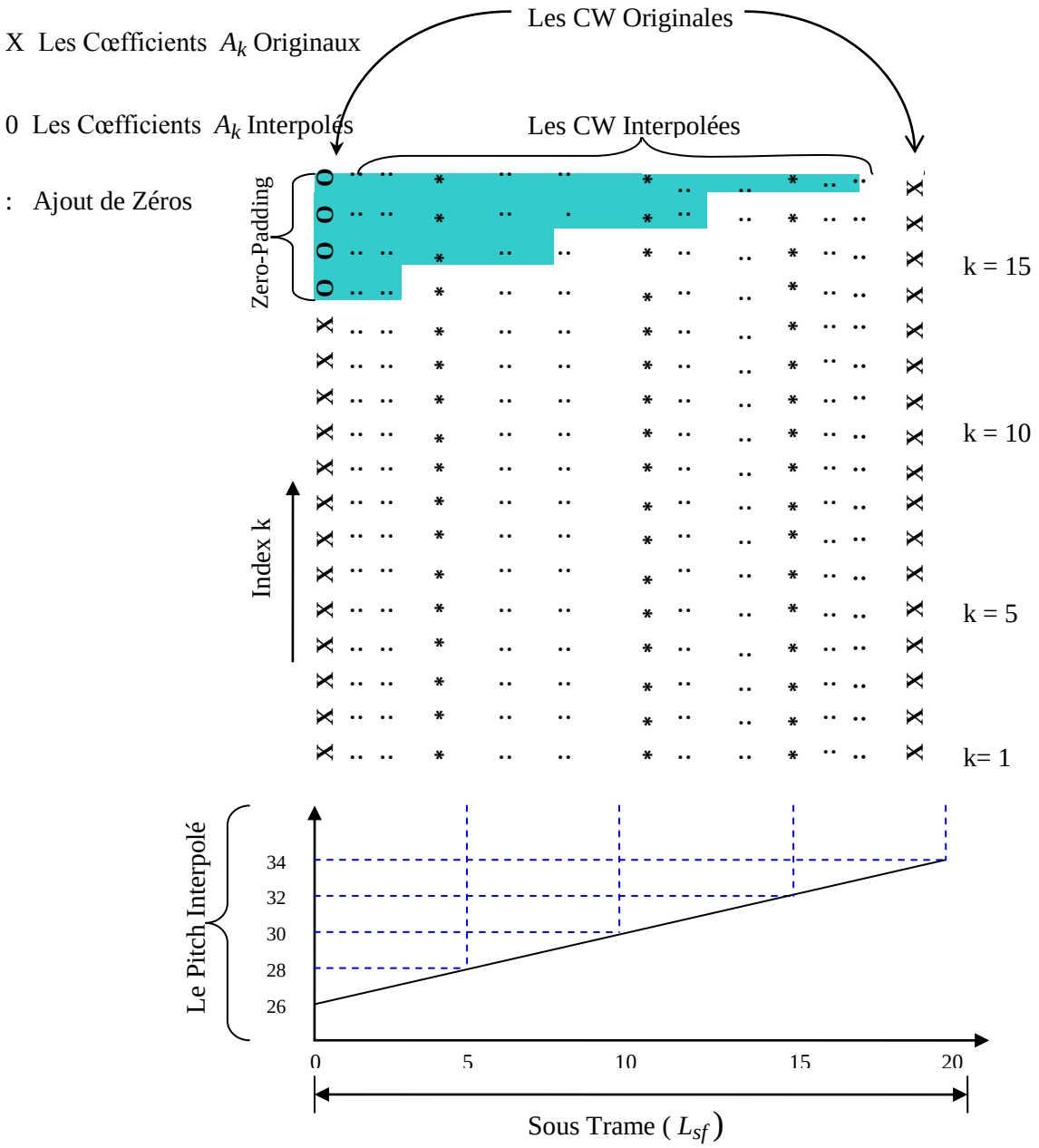

Fig.3.16 : Un exemple d'interpolation des CW sur un intervalle d'une sous-trame.

En supposant que le pitch évolue linéairement sur l'intervalle d'intégration, (3.28) peut être écrite sous la forme :

$$\phi(n)=\phi(n-1)+\int_{n-1}^{n}\frac{2\pi}{(n-n')P(n-1)+(n'-n+1)P(n)}\,dn' \qquad (3.29)$$

Une évaluation rapide de cette intégrale mène à :

$$\phi(n)=\begin{cases}\phi(n-1)+\dfrac{2\pi}{P(n)-P(n-1)}\ln\left[\dfrac{P(n)}{P(n-1)}\right] & si \quad P(n)\neq P(n-1)\\ \phi(n-1)+\dfrac{2\pi}{P(n)} & si \quad P(n)=P(n-1)\end{cases} \tag{3.30}$$

Pour une implémentation en pratique, et afin de réduire la complexité de calcul, la relation (3.31) est une approximation faible de (3.30) [12, 13].

$$\phi(n)=\phi(n-1)+\pi\left(\frac{1}{P(n-1)}+\frac{1}{P(n)}\right) \tag{3.31}$$

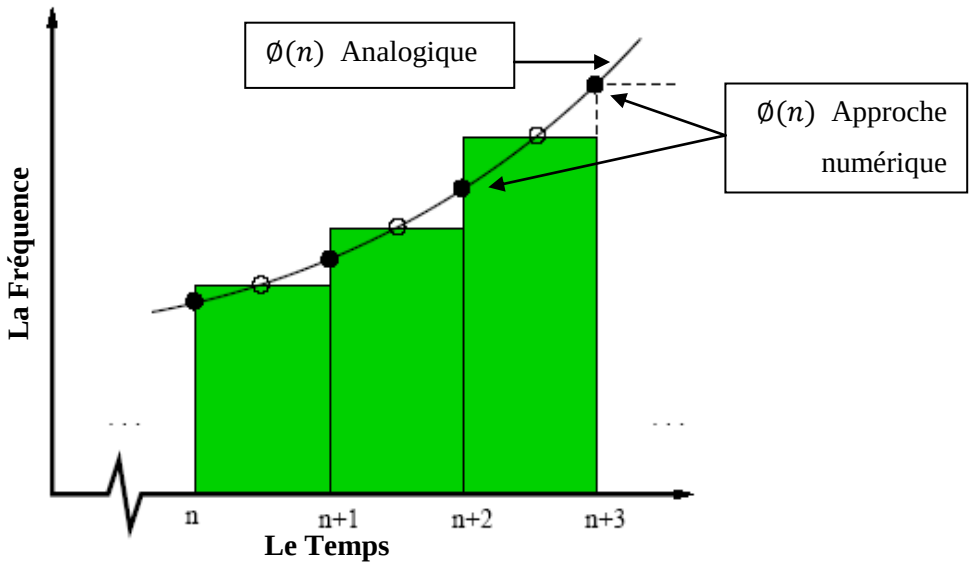

Fig.3.17 : Comparaison entre les deux approches de calcul de phase [6].

La phase initiale $\phi(0)$ au début de la première trame peut être fixée à une valeur arbitraire (aléatoire) car elle n'affecte pas la qualité de perception de la parole reconstituée, mais il reste préférable d'avoir la valeur exacte de la phase initiale.

3.3.2.4. Calcul du signal résiduel

L'opération de conversion en un signal résiduel unidimensionnel $r(.)$ est effectuée échantillon par échantillon comme on peut le voir graphiquement par l'exemple de la figure 3.18 qui montre le processus de reconstitution où chaque CW est normalisée à la longueur 2π. La transformation se fait en superposant les deux graphes. La projection de leur intersection (points de rencontre des droites de poursuite de la phase avec la surface des CW), donne le signal résiduel $r(n)$. Cette transformation est implémentée par l'opération inverse de la décomposition en DTFS [7, 26], ainsi est exprimé par :

$$\boldsymbol{r(n)=s(n,\phi(n))=\sum_{k=0}^{[P(n)/2]}\left[A_k(n)\cos(k\phi(n))+B_k(n)\sin(k\phi(n))\right]} \qquad \boldsymbol{0\le\phi(\cdot)<2\pi} \qquad (3.32)$$

Le signal résiduel reconstitué est utilisé comme signal d'excitation du filtre de synthèse LP pour obtenir le signal parole final. La fonction de transfert du filtre est équivalente à celle de modèle de production de la parole, et les coefficients du filtre sont donnés par la conversion des coefficients LSF en coefficients LP après interpolation.

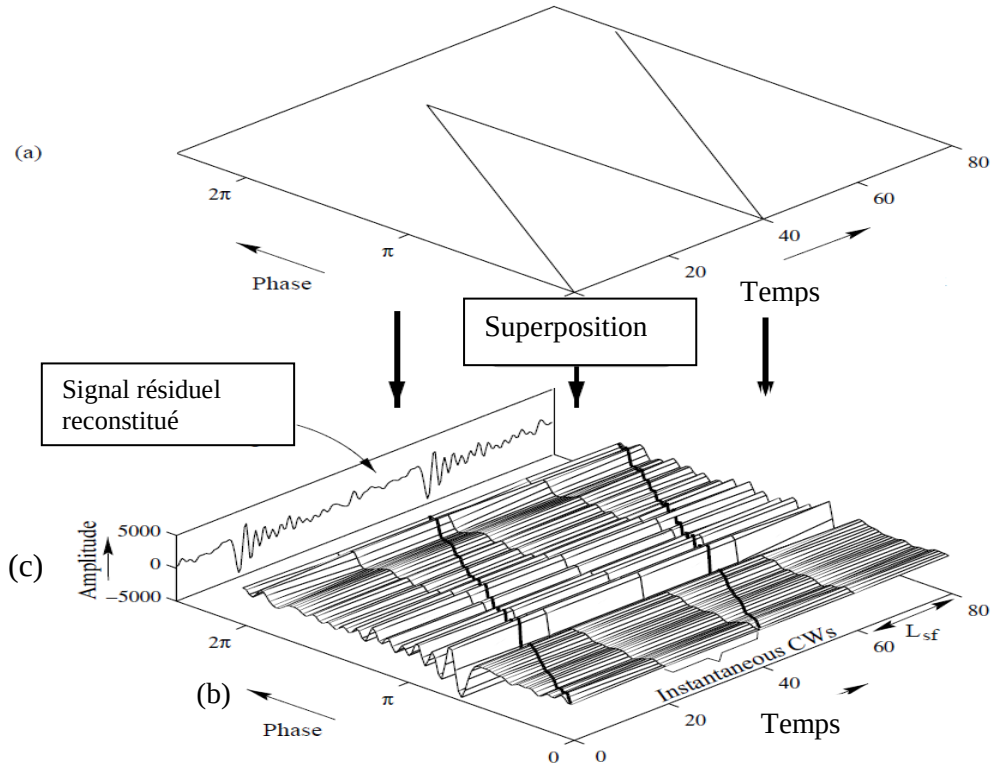

Fig.3.18 : Transformation de la surface 2D à 1D de CW, (a) les droites indiquant la poursuite de la phase interpolée instantanée pour un signal voisé de pitch égal à 40. (b) la surface de CW sur laquelle seront projetées les droites de (a), (c) le signal résiduel résultant.

3.3.2.5. Application de WI sur le signal original

Le problème détecté dans la WI est que, quand le signal d'excitation reconstruit est passé par le filtre inverse afin de synthétiser la parole, des effets indésirables apparaissent. Ils sont dûs au filtrage adaptatif, ainsi la parole reconstruite peut exhiber une enveloppe indésirable, dont le résultat est un gazouillement audible, pour certains sons. Les tests d'écoute persistants et les examens détaillés de la visualisation de l'enveloppe temporelle du signal reconstitué dans la WI, ont montré l'existence de variations d'amplitude indésirables.

Pour enlever ces gazouillement dans la parole reconstruite, ainsi plusieurs tentatives étaient réalisées, telle qu'une compensation d'énergie pour lisser l'enveloppe temporelle de la parole reconstruite [13, 35] ; Mais les résultats n'étaient pas vraiment satisfaisants.

Une autre méthode parait être plus bénéfique, consiste à appliquer les principes de la WI directement sur la parole [13]. En fait, exécuter l'extraction et l'interpolation dans le domaine de la parole peut éliminer les variations de l'enveloppe temporelle dans la parole reconstruite (figure 3.19). Cette méthode peut aussi mener à améliorer l'efficacité du codeur WI.

Quand on applique la WI directement sur le signal parole original il faut prendre en considération, que le pitch est estimé à partir du signal résiduel, pour que sa valeur soit plus précise, et les formes d'ondes sont extraites à partir du signal original, ensuite les mêmes procédures précédentes sont appliquées.

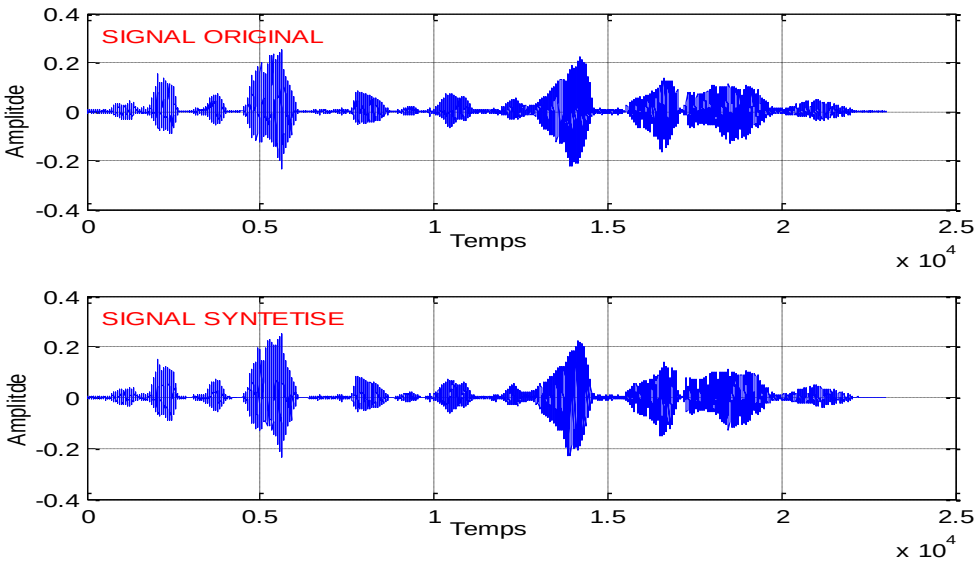

Fig.3.19 : Application de la WI sur le signal original : graphes original (en haut) et reconstitué (en bas).

Les tests d'écoute ont montré que la parole reconstituée contient quelque rugosité (un caractère bruyant ou enrouement), malgré l'amélioration de la forme de l'enveloppe temporelle, et même sans quantification. De telle distorsion a été diagnostiquée, pour être causée par la haute limite d'énergie dans la CW extraite, et qui a mené directement aux discontinuités audibles.

Une solution simple paraît efficace. Elle consistera à sur-échantillonner le signal d'entrée avant la réalisation de la WI ; à rehausser la précision du pitch, à titre d'exemple en réalisant la méthode employée, dans la dernière version de la WI ou la EWI (Enhanced Waveform Interpolation) [28]. cette dernière consiste à faire la détection du pitch deux fois par trame au lieu d'une fois. Finalement, sous- échantillonner la parole décodée.

De telle procédure de sur-échantillonnage augmente la résolution du signal de parole, aussi bien que la précision du pitch estimé, et par conséquent, réduit l'énergie de la limite de la CW extraite. Cependant, cette procédure est associée avec une augmentation de la complexité du codeur, parce que les longueurs des CW se sont étendues par le processus de sur-échantillonnage, donc le nombre d'opérations sera plus élevé.

3.3.2.6. Décomposition des CW

A première vue, il apparaît qu'une représentation précise des CW nécessite un débit de transmission très élevé, plus particulièrement pour les segments non voisés qui possèdent un plus grand débit d'information. Avantageusement, l'oreille humaine n'est pas sensible à toute l'information contenue dans cette surface, la perception humaine des sons voisés est très différente de celle des sons non voisés, ce qui suggère la possibilité d'exploiter une telle différence pour quantifier les CW avec une meilleure précision du point de vue perception.

Au lieu d'adopter une classification (voisée / non voisée), une nouvelle technique de décomposition a été utilisée [7], dans laquelle chaque CW est séparée en deux composantes avant la quantification. Ces deux composantes sont : une forme d'onde à évolution lente SEW (Slowly Evolving Waveform) et une forme d'onde à évolution rapide REW (Rapidly Evolving Waveform), représentant les composantes périodique et non périodique du signal parole. En exploitant la différence dans la perception humaine de ces deux formes d'ondes, une meilleure efficacité de codage est possible en les quantifiant séparément.

La SEW est obtenue en filtrant passe-bas la surface des CW le long de l'axe du temps discret, et la REW peut être obtenue en retranchant la SEW de la CW. Pour un signal parole, la SEW et la REW représentent, respectivement, une forme d'onde ressemblant à une impulsion et une composante de bruit.

Vu la présence de périodicité dans les régions voisées, la SEW possède, généralement, un niveau d'énergie plus élevé que la REW, inversement, pour la parole non voisée où le signal évolue plus rapidement et où il n'y a aucune périodicité apparente.

3.3.2.6.1. Conception du filtre passe-bas

C'est un filtre Anti-repliement non causal à phase linéaire. Cependant, ce filtre nécessite une fréquence de coupure de 20 Hz [7], équivalente à la fréquence normalisée 0.1, ou même la fréquence de coupure de 25 Hz équivalente à la fréquence normalisée 0.125, sa réponse impulsionnelle notée $h_{CW}(i)$, calculée par fenêtrage de la réponse d'un filtre passe-bas idéal (coupure à 20 ou 25 Hz) avec une fenêtre de Hamming de longueur 17 échantillons. Finalement, on peut obtenir $h_{CW}(i)$ en normalisant la réponse fenêtrée :

$$\sum_{i=-8}^{8} h_{CW}(i) = 1 \tag{3.33}$$

La figure 3.20 trace la réponse en amplitude et en phase de $h_{CW}(i)$ et sa réponse impulsionnelle. Notons que la réponse en fréquence possède une bande de transition assez large. Ceci est dû, principalement, au fait que le filtre FIR possède seulement 17 coefficients. On peut augmenter le nombre de coefficients pour avoir une meilleure précision du filtre, mais aux dépends d'un retard algorithmique plus important.

3.3.2.6.2. Calcul des SEW et REW

Sachant que la transformation en DTFS est une opération linéaire, le filtrage passe-bas des CW dans le temps est équivalent au filtrage passe-bas de leurs coefficients DTFS. Pour cette raison, on réalise le filtrage directement sur les coefficients A_k et B_k. De manière plus précise, pour calculer la CW filtrée passe-bas à l'instant n, on peut utiliser la formule :

$$\left.\begin{aligned} \tilde{A}_k(n) &= \sum_{i=-8}^{8} A_k(n - iL_{sf}) h_{CW}(i) \\ \tilde{B}_k(n) &= \sum_{i=-8}^{8} B_k(n - iL_{sf}) h_{CW}(i) \end{aligned}\right\} \tag{3.35}$$

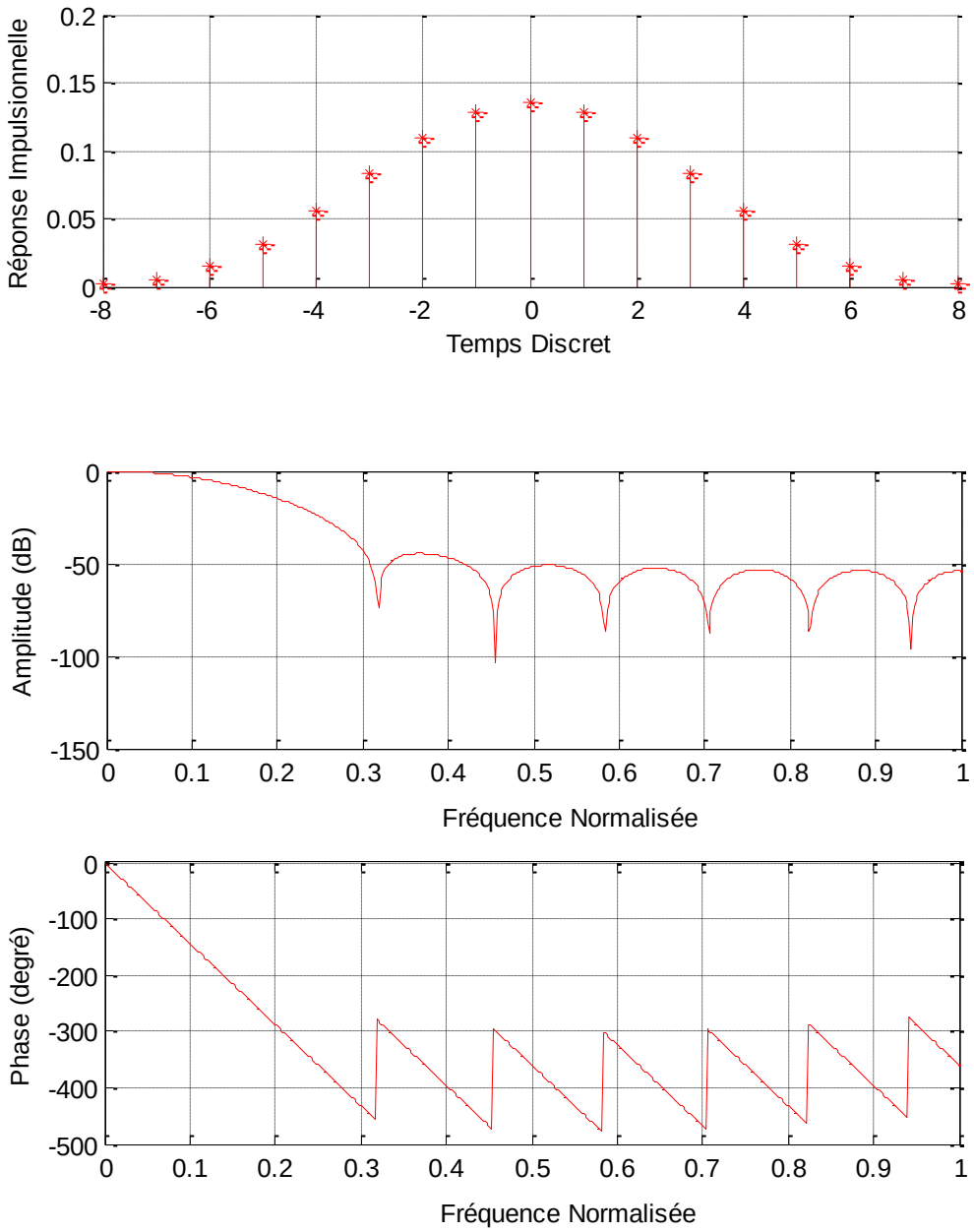

Fig.3.20 : Caractéristiques du filtre passe-bas de décomposition en SEW-REW. En haut : sa réponse impulsionnelle *hcw*. Au milieu : sa réponse en amplitude. En bas : sa réponse en phase. La fréquence de coupure normalisée égale à 0.1 (20Hz).

Or, la dimension des CW varie avec le pitch. Pour faciliter le filtrage, les mêmes techniques des paragraphes précédents (3.3.1.3) sont utilisées, pour allonger ou contracter les CW de manière à ce que toutes les CW à l'intérieur de la fenêtre de filtrage aient la même longueur avant le filtrage. La figure .3.21 décrit l'opération d'ajustements des CW appliquée avant le filtrage.

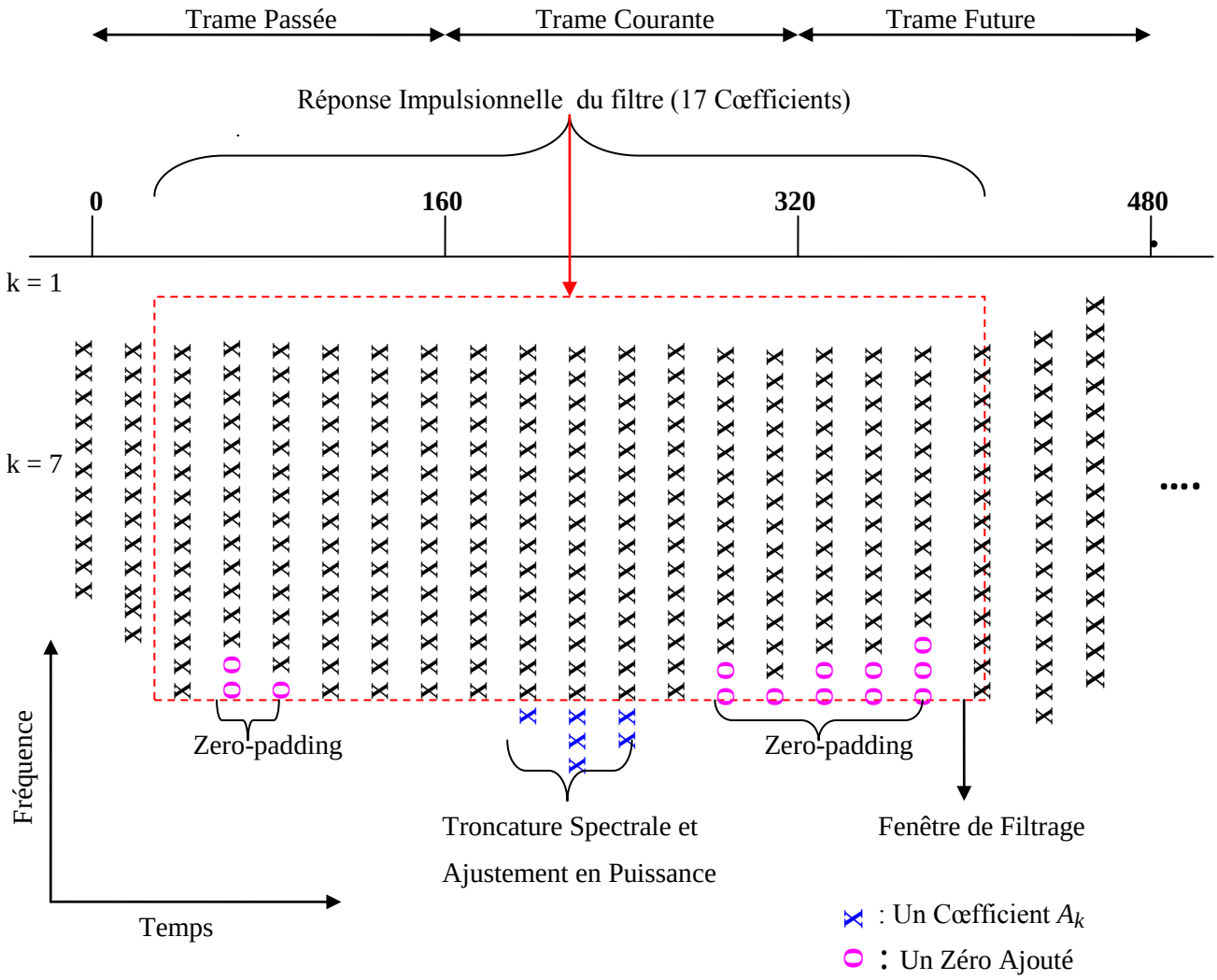

Fig.3.21 : Opération de filtrage passe-bas pour la décomposition en SEW-REW. Le Schéma montre 17 CW successives couvrant trois trames.

Le filtre est centré à la CW du point milieu. Puisque la longueur originale de cette CW est de k = 15, toutes les CW dans la fenêtre de filtrage doivent avoir la même longueur avant le filtrage. Les plus courtes CW (longueurs < 15) seront étendues par ajout de zéros tandis que les plus longues (longueurs > 15) seront tronquées puis ajustées en puissance.

Ensuite, la procédure de filtrage est exécutée *k* par *k* (ligne par ligne). La figure .3.22 illustre un exemple de décomposition en deux surfaces SEW et REW de 17 formes d'ondes successives.

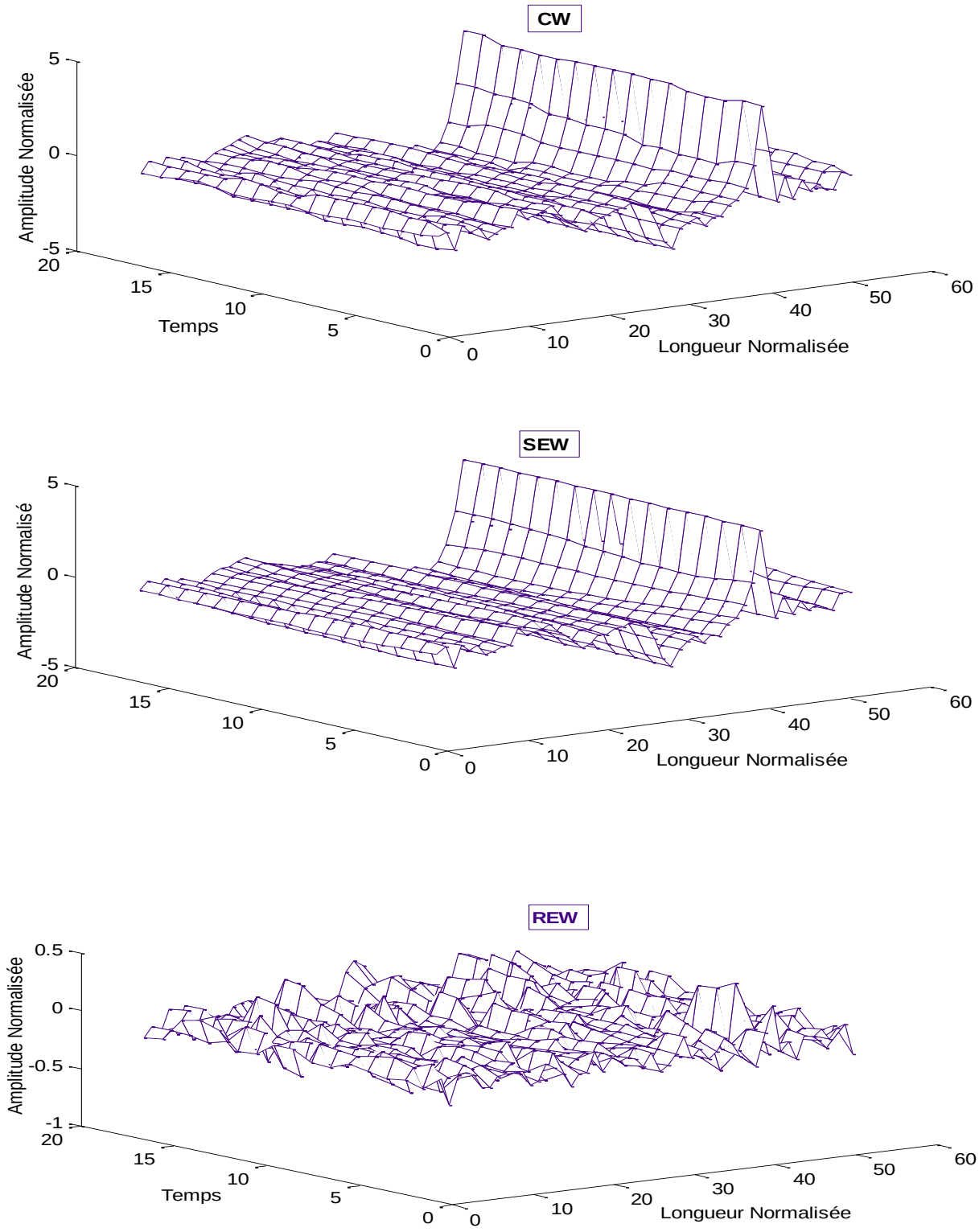

Fig.3.22 : Décomposition d'un segment de longueur 40 ms (17 CW) en surfaces SEW et REW. La fréquence de coupure du filtre passe-bas est de 20 Hz [7].

3.4. Réduction de débit des paramètres de la WI

La couche d'analyse-synthèse de la WI (en l'absence de quantification) fournit une parole de qualité transparente et ferait l'objet d'une excellente base pour le développement d'un codeur de la parole à des débits considérablement réduits.

Il y a quatre paramètres à quantifier dans le schéma de la WI : les paramètres LP (LSF), le pitch, l'énergie et les CW.

Dans l'allocation de bits du codeur WI autours de 4 kbps, on alloue 24 bits [38] pour la quantification de chaque ensemble de LSF dont la fréquence de mise à jour et de transmission est de 50 Hz. On utilise pour cela la quantification vectorielle par segmentation SVQ (Split Vector Quantization). Cependant, le débit de quantification des cœfficients LSF peut aussi être réduit jusqu'à 20 bits par la technique de quantification vectorielle à plusieurs niveaux MSVQ (Multi-Stage Vector Quantization) [28].

La fréquence de transmission de pitch est de 50 Hz (un par trame). Puisque l'estimateur de pitch fournit des valeurs entières, nous avons un total de 101 valeurs possibles (120-20+1), qu'on peut coder à l'aide de 7 bits [7], en utilisant un quantificateur scalaire.

Contrairement aux LSF, la puissance nécessite un traitement supplémentaire avant la quantification. Etant donné que le logarithme du signal puissance est plus significatif que le signal puissance lui-même, les valeurs entrantes de la puissance sont d'abord, transformées au domaine logarithmique. Puis, elles sont filtrées passe-bas et sous-échantillonnées de 400 Hz à 100 Hz (2 valeurs / trame) [7]. Les valeurs sous-échantillonnées sont codées par la technique de quantification vectorielle par analyse et synthèse sur 6 bits. Au récepteur, le signal puissance est décodé et sur-échantillonné à la fréquence de 400 Hz par interpolation ; c'est une interpolation linéaire exécutée directement sur les valeurs du logarithme de la puissance. Une fois le contour de puissance est sur-échantillonné, le signal puissance est obtenu par l'opération inverse, ou exponentielle.

La technique de quantification des CW (REW et SEW), reste toujours un problème inévitable. La technique optimale est celle qui donne un meilleur compromis entre, la qualité perceptuelle des CW reconstituées, le débit de quantification et le temps d'exécution. Ce qui fait appelle à une recherche exhaustive pour la compression de cette composante.

3.4.1. Quantification conventionnelle des CW

Dans cette section, on va parler de la quantification et la déquantification conventionnelle des CW, La figure 3.23 donne les schémas blocs des deux processeurs. Les CW comme la puissance ; nécessitent un traitement supplémentaire avant la quantification. Plus précisément, comme déjà vu, chaque CW est décomposée en deux formes d'ondes (SEW et REW) qui seront quantifiées séparément, avec un traitement spécifique pour chaque composante.

3.4.1.1. Quantification des REW

Commençons tout d'abord par lister trois conclusions importantes [44] :

1- Une faible dégradation dans la qualité de la parole est observée si le spectre de phase des REW est remplacé par un spectre de phase aléatoire.

2- Aucune détérioration n'est observée dans la parole résultante si chaque spectre d'amplitude d'une REW est lissé par une fenêtre carrée de 1000 Hz.

3- Une très petite détérioration audible est produite si le spectre d'amplitude d'une REW est moyenné sur tous les REW dans un intervalle de 5 ms.

La première conclusion montre que le spectre des REW comporte quelques informations perceptibles et ne doit pas être transmis avec un faible débit. Les deuxième et troisième impliquent que la résolution dans le temps du spectre d'amplitude des REW est nettement plus importante que sa résolution en fréquence.

Pour exploiter ces résultats, la REW entrante est sous-échantillonnée à un débit de 200 Hz qui est en accord avec la résolution dans le temps suggérée par la troisième conclusion (intervalle de 5 ms). Chaque REW sous-échantillonnée est, alors, convertie vers sa représentation polaire où le spectre de phase est complètement écarté. Le spectre d'amplitude est quantifié vectoriellement en utilisant la technique VDVQ, avec une dimension égale à L=60 (pitch maximal/2). On utilise un dictionnaire de taille aussi petite car une description grossière du spectre d'amplitude des REW est suffisante pour avoir une bonne qualité de codage selon la deuxième conclusion.

Au récepteur, les spectres des REW sont décodés et sur-échantillonnés par un facteur de 2, du débit 200 Hz à 400 Hz. Cela est effectué en insérant un nouveau spectre après chaque spectre reçu. Ces nouveaux spectres sont obtenus par interpolation linéaire des spectres adjacents ou en choisissant le spectre précédent. Finalement, chaque spectre d'amplitude d'une REW sur-échantillonnée est

combiné avec un spectre de phase aléatoire puis reconverti en coordonnées rectangulaires. Les valeurs de la phase dans les spectres sont indépendantes et uniformément réparties dans [-π, π]. Il est à noter que les spectres de phase aléatoire sont ajoutés aux REW à la fréquence des sous-trames.

3.4.1.2. Quantification des SEW

Puisque la fréquence de coupure du filtre de décomposition, est généralement compris entre (25, 20) Hz, les SEW ont une largeur de bande d'évolution très petite (leur évolution est très lente). Cela suggère qu'on peut les sous-échantillonner de 400 Hz à environ 50 Hz. Cependant il est plus avantageux de les sous-échantillonner à une fréquence un peu plus grande, de 100 Hz afin de compenser l'imprécision du filtre de décomposition. Chaque SEW sous-échantillonnée est convertie en notation polaire dont on écarte le spectre de phase. Le spectre d'amplitude est divisé en trois sous-bandes sans recouvrement, 0 - 1000 Hz, 1000 - 2000 Hz, et 2000 - 4000 Hz. Ces sous-bandes sont quantifiées séparément par la technique VDVQ, où la bande de base est quantifiée avec un débit plus grand que les deux autres sous-bandes. Une telle allocation de bits est due à la grande capacité de résolution de l'oreille humaine pour les basses fréquences.

Au récepteur, après décodage et combinaison des sous-bandes, on applique une interpolation linéaire pour ajuster le spectre combiné aux extrémités des sous-bandes (c-à-d. à 1000 Hz et à 2000 Hz). Un changement brusque ou une discontinuité importante dans le spectre peut causer des distorsions dans la parole reconstituée.

Après avoir reconstitué et lissé le spectre d'amplitude, on lui associe un spectre de phase fixe et on le retransforme en coordonnées rectangulaires. Ce spectre de phase fixe est donné à partir d'un segment voisé d'une voix d'homme à pitch élevé (maximum d'harmoniques) [6]. Après, les SEW sont sur-échantillonnées du débit 100 Hz à 400 Hz.

Quant aux procédures de recherche des dictionnaires, elles sont identiques à celles des REW, on choisi dans le dictionnaire le vecteur qui minimise l'erreur quadratique moyenne, à l'exception pour la bande de base où on emploie le critère de l'erreur modérée par perception (Perceptually Weighted Error). Ce critère est très utilisé dans les codeurs basés sur la technique CELP.

$$W(z)=\frac{1-\sum_{k=1}^{N} a_k \gamma_2{}^k z^{-k}}{1-\sum_{k=1}^{N} a_k \gamma_1{}^k z^{-k}} \qquad 0<\gamma_w \leq 1 \tag{3.36}$$

Où γ_1, γ_2 sont typiquement compris entre 0 et 1 [3]; pour le CELP on a γ_1=0.9, γ_2=0.5.

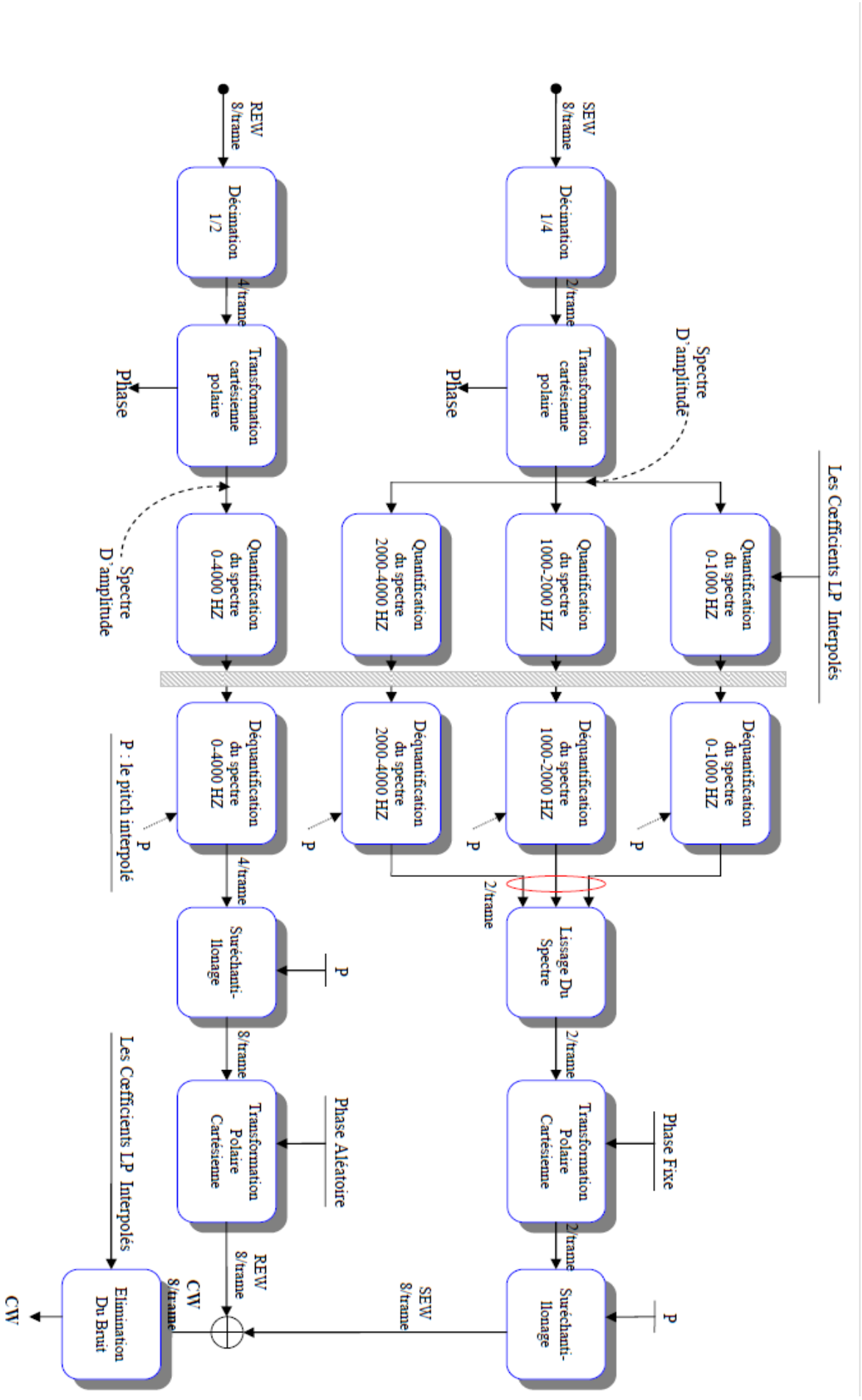

Fig.3.23 : Schéma général de quantification et dé-quantification des SEW et REW.

3.4.2. La technique de quantification des REW étudié en [4]

La quantification directe du spectre d'amplitude des REW est un problème de la quantification à dimension variable VDVQ ; en effet, cette quantification demande un grand effort de calcul, en plus elle conduit à une perte d'information remarquable.

Pour remédier à ce problème, une technique simple et pratique, consiste à transformer les longueurs variables des REW à une seule et unique longueur fixe, le principe de cette transformation est de mettre le spectre d'amplitude des REW ou $R(\omega)$, sous la forme d'une combinaison linéaire de fonctions de base, tel que les fonctions orthogonales $\psi_i(\omega)$

$$\boldsymbol{R(\omega)=\sum_{i=0}^{L-1}\gamma_i\psi_i(\omega)} \qquad \boldsymbol{0\leq\omega\leq 2\pi} \tag{3.37}$$

Une telle représentation rend l'amplitude des REW plus lisse, en fait, le lissage des amplitudes des REW peut améliorer réellement la qualité perceptuelle de la parole reconstruite. Parmi les plusieurs méthodes de conversion, l'approche basée sur la transformée en cosinus discrète DCT (Discrete Cosine Transform) était capable de remplacer la quantification à dimension variable, avec la plus haute exactitude [33].

Il a été constaté que typiquement il y a des ressemblances entre les spectres adjacents des amplitudes des REW [30]. Cela suggère que cette corrélation peut être exploitée dans la quantification, et qu'une représentation simplifiée des REW pourrait être possible avec un taux de mise à jour approximativement de 200 Hz. Cependant, puisque la propriété essentielle des REW, est qu'elles évoluent rapidement, la réalisation d'une quantification vectorielle qui effectue une analyse par synthèse ne résulterait pas nécessairement des améliorations considérables, donc une simple quantification vectorielle sera suffisante.

Donc ici, nous présentons une méthode alternative, dont en exploitant la corrélation entre les spectres des REW consécutifs. La figure 3.24, fournie un schéma détaillé de la couche de quantification et déquantification du spectre d'amplitude des REW, basé sur la DCT avec un taux de décimation des REW égale à 2.

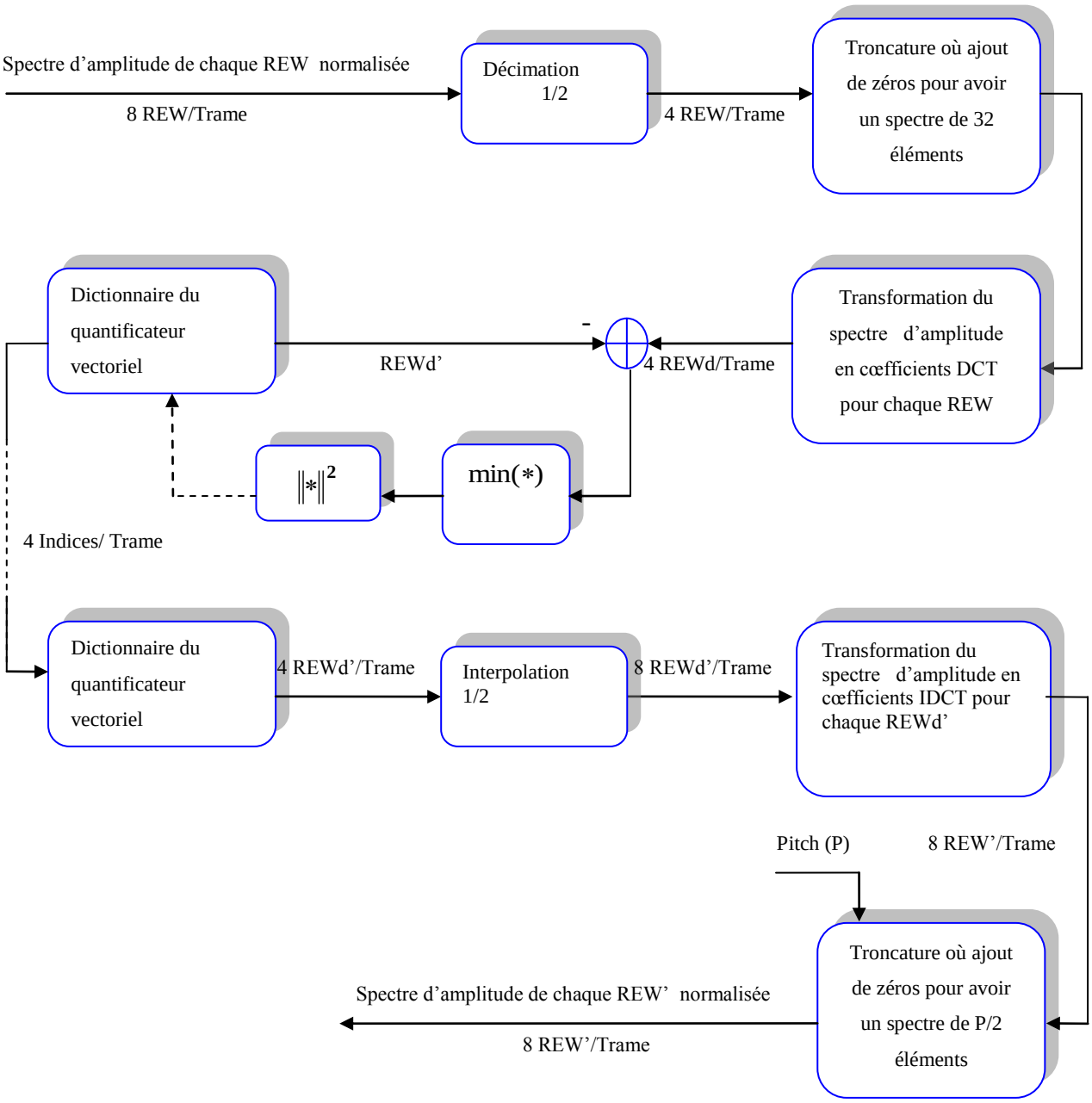

Fig.3.24 : Schéma bloc de la quantification et dé-quantification des REW.

Une étude pratique de cette quantification adoptée comme critère d'évaluation la mesure du rapport signal sur bruit montre les performances de codeur avec les critères suivant :

a) **Débits de quantification :** après la transformation du spectre d'amplitude des REW au domaine DCT avec une longueur choisi égale à 32. chaque spectre DCT est codé par 4 bits, ce qui se traduit par un débit de 16 bits/trame, donc un dictionnaire est constitué de 16 vecteurs DCT.

Le rapport signal sur bruit moyen entre voix masculin et féminin correspond à ce débit est égale à : 6.6289.

b) Ajustement des dictionnaires : parce que l'espace mémoire occupé par le codeur est un caractère de performance de celui-ci. L'objective de l'justement est de réduire l'espace mémoire du dictionnaire de quantification des cœfficients DCT ; la méthode d'ajustement polynomiale permet de représenter une forme d'onde par les cœfficients d'un polynôme qui a la plus proche ressemblance de cette forme, ainsi un signal de L échantillons peut être représenté par N échantillons tel que $N<L$, et N représente l'ordre du polynôme ajusté.

Le rapport signal sur bruit moyen entre voix masculin et féminin correspond au débit 16 bits/trame et un ajustement d'ordre 13 est' égale a : 6.5628.

c) Décimation des REW reconstruites : cette procédure sert à réduire le débit de quantification en modifiant le degré de décimation, les figures suivantes montrent l'effet de l'interpolation des cœfficient DCT après l'exécution d'une décimation d'ordre 4 comparé au résultat d'une décimation d'ordre 2 ; pour 4 REW dans une trame d'analyse.

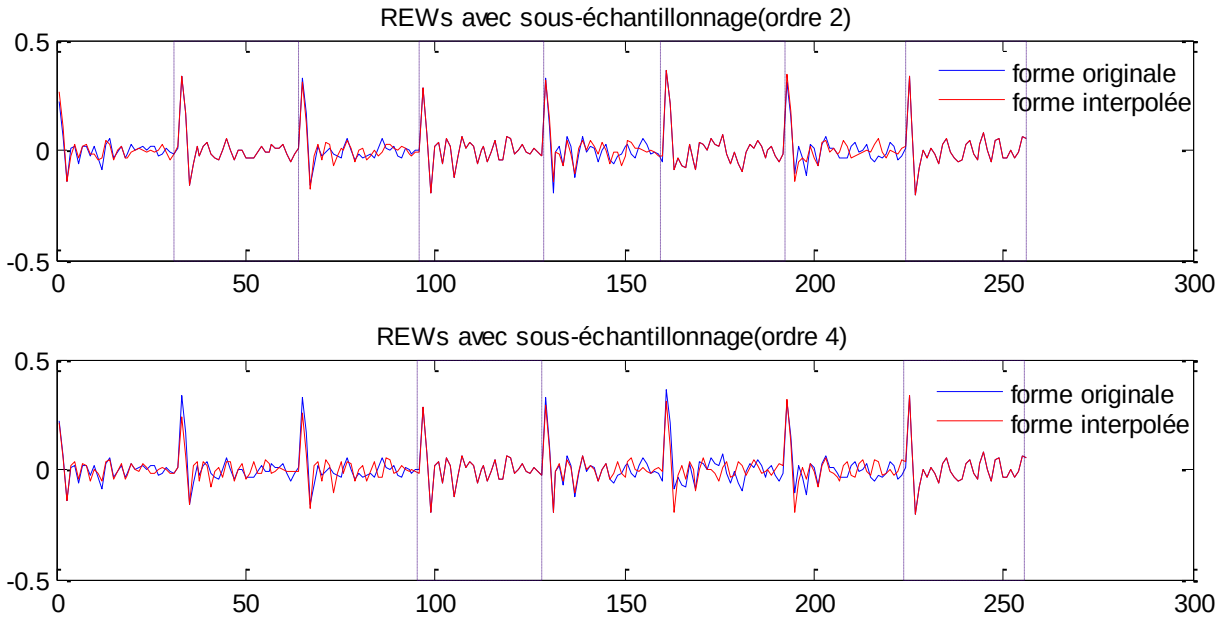

Fig.3.25 : Effet de l'interpolation sur les cœfficients DCT.

D'après des essais effectués sur des REW sous échantillonnées avec un rapport de quatre, la valeur du SNR a subit une faible dégradation. Mais sur le point de vue perception, la dégradation du signal reconstruit est plus importante que celle échantillonnées avec un rapport de deux, c-à-d on n'aura pas un codeur de qualité communication (Toll Quality).

3.5. Conclusion

Le but principal de notre travail est de trouver une amélioration pour le codage WI, cela demande une préconnaissance de tous les détails de ce dernier pour lui adresser la modification qui assure l'augmentation de ses performances qualitatives (intelligibilité de parole) et quantitatives (taux de compression).

Un tel codeur destiné à la compression doit avoir deux partie ; un codeur pour l'analyse et un décodeur pour la synthèse.

Le rôle du codeur est de faire l'analyse du signal et extraire les paramètres nécessaires pour la synthèse d'un signal estimé. Le signal parole est divisé en trame de 160 échantillons. À partir de chaque trame:

- On calcule les LPC, et on les transforme en LSF.

- On estime le pitch.

- Extraction de CW pour chaque sous trame ainsi on calcule la puissance correspondante, on les normalise, aligne et puis à l'aide d'un filtre passe bas on décompose les CW en SEW et REW. Pour réduire le débit, chaque composante est quantifiée avec une quantification convenable. Et puis elles sont près à être stockées ou transmises dans un canal.

Où décodage, les paramètres résultants de codage « (SEW, REW), pitch, puissance, LSF » sont rassemblés d'une manière synchronisée pour former une estimation du signal original.

Les SEW et REW sont associés pour former les CW qui seront dénormalisées à l'aide de puissance.

Les pitch de trame précédente et présente sont interpolés pour créer 8 pitch, un pour chaque sous trame.

Avec les pitch interpolés et les CW alignés, on estime la phase instantanée et puis le signal résiduel qui sera filtré par le filtre constitué par les LSF. Et donne en résultat l'estimation du signal original.

L'efficacité du codeur WI se centralise dans la meilleure qualité de codage des deux composantes SEW et REW .Dans ce chapitre nous avons détaillé la quantification des composantes REW et vérifié les résultats de cette quantification. Il nous reste qu'à chercher une méthode pour diminuer le débit des SEW qui occupent un débit de 18 bit/trame.

Chapitre 4 : Compression des SEW avec sous échantillonnage et interpolation.

4.1. Introduction

Dans le chapitre précédent, nous avons vu que l'étude du décodeur fait au [4] montre des résultats concrets ; des paroles intelligibles, des rapports SNR acceptables et un taux de compression très important. Ainsi, on a déjà vu que le codeur s'articule autour de la quantification des REW afin de réduire le taux de compression.

Maintenant qu'est ce qu'on peut faire pour les SEW pour réduire encore leurs débits au niveau du codeur ?

Le fait que le signal parole est pseudo périodique dans ces parties voisées implique l'existence d'une grande redondance. Cette propriété conduit à penser à l'élimination des redondances.

Si cette tâche est possible, elle sera convenable aux SEW plus que les REW vu la variation lent de ceux-ci.

Pour assurer une telle astuce, il faut d'abord étudier et bien observer le comportement des SEW du codeur suivant les trames et les sous trames.

4.2. Étude de variation des SEW dans une trame

Alors on remarque sur la figure .4.1 que cette trame a une forme pseudo-périodique, nous estimons que sa période est de 40(échantillons) à peu près. Cependant, aucune nouvelle n'est parue. On sait que les signaux voisés ont déjà cette forme au domaine temporelle. Mais est ce que cette périodicité engendre une périodicité dans le domaine des composantes SEW ?

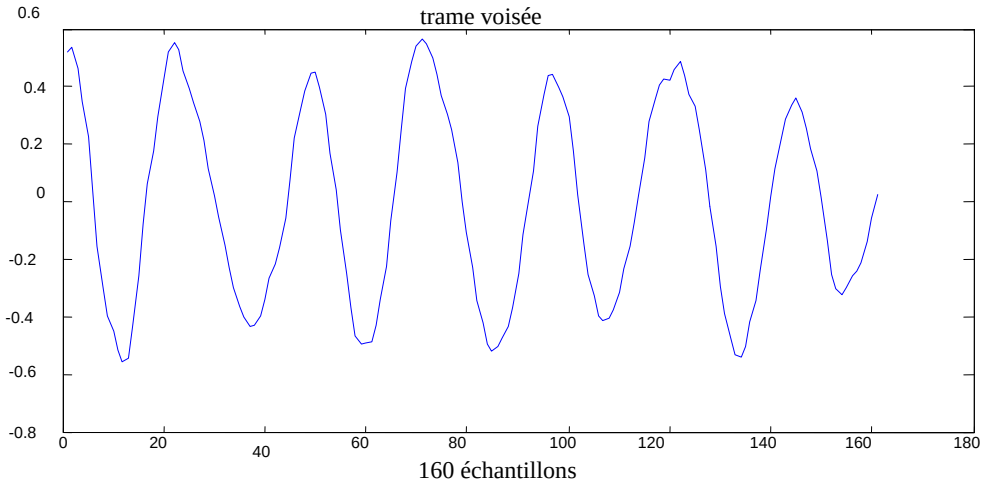

Fig.4.1 : trame voisée de 160 échantillons.

Une trame est divisée en huit sous trames. On va représenter les SEW de chaque sous trame.

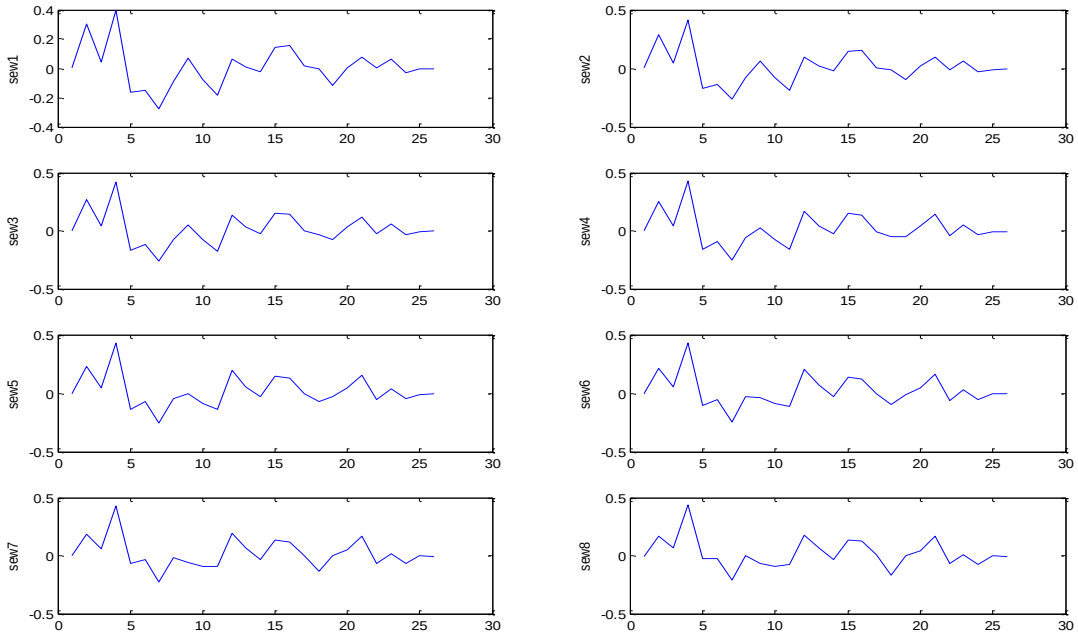

Fig.4.2 : les graphes des 8 SEW d'une trame voisée.

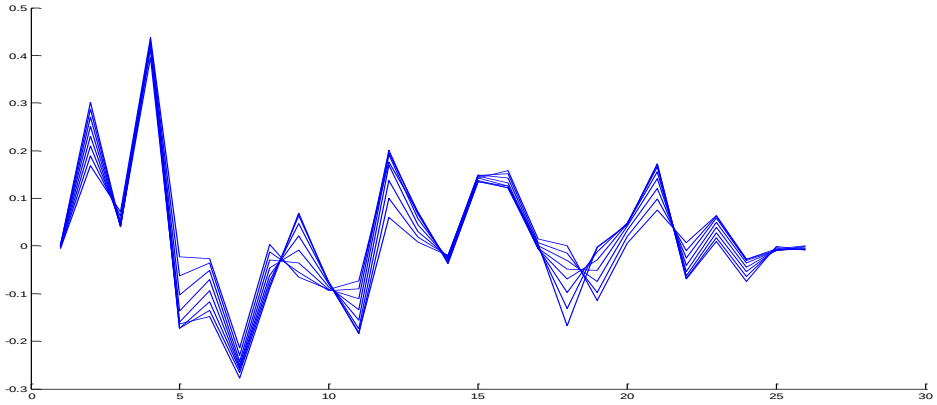

Fig.4.3 : les 8sew superposées.

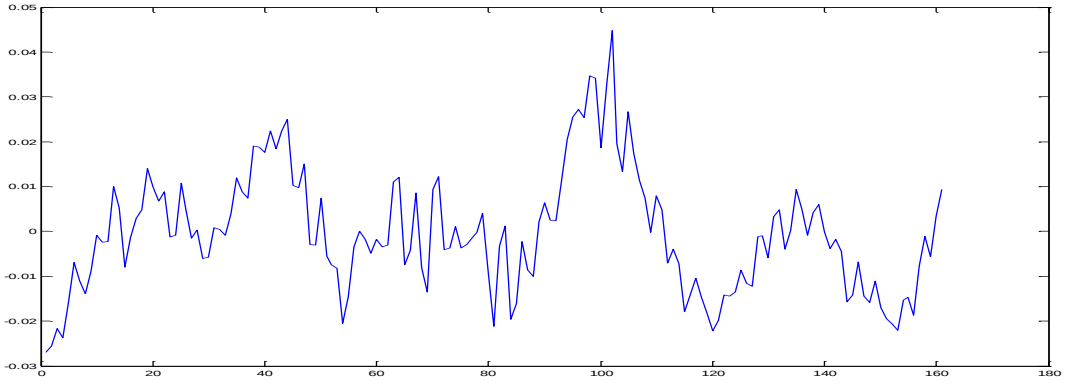

Fig.4.4 : trame non voisée.

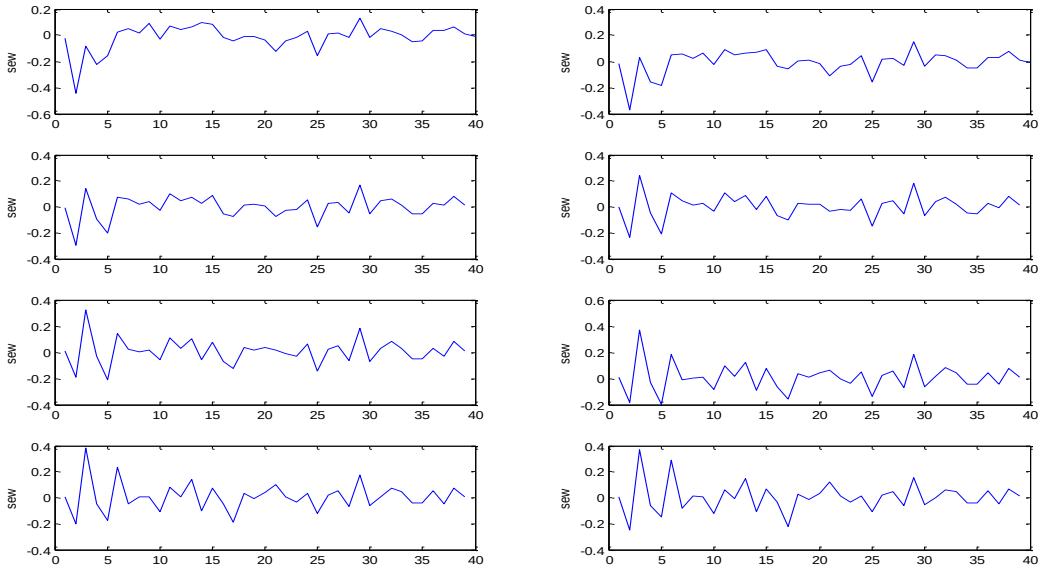

Fig.4.5 : les graphes des 8 SEW d'une parole non voisée.

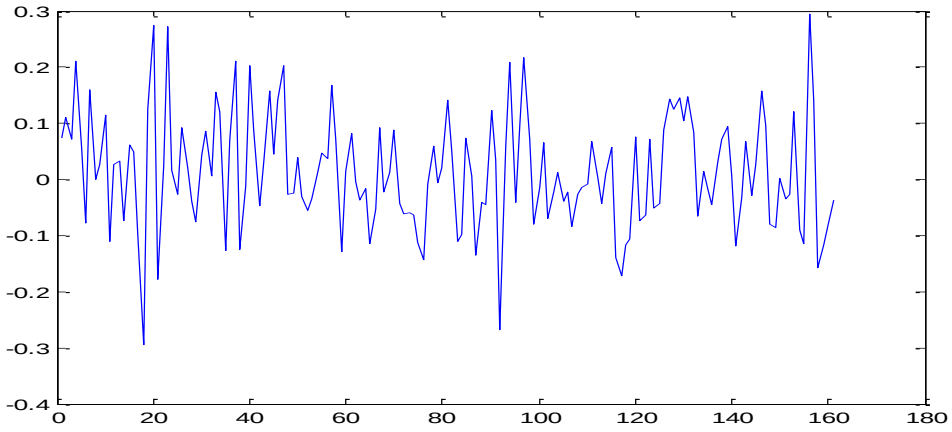

Fig.4.6 : l'allure d'un bruit.

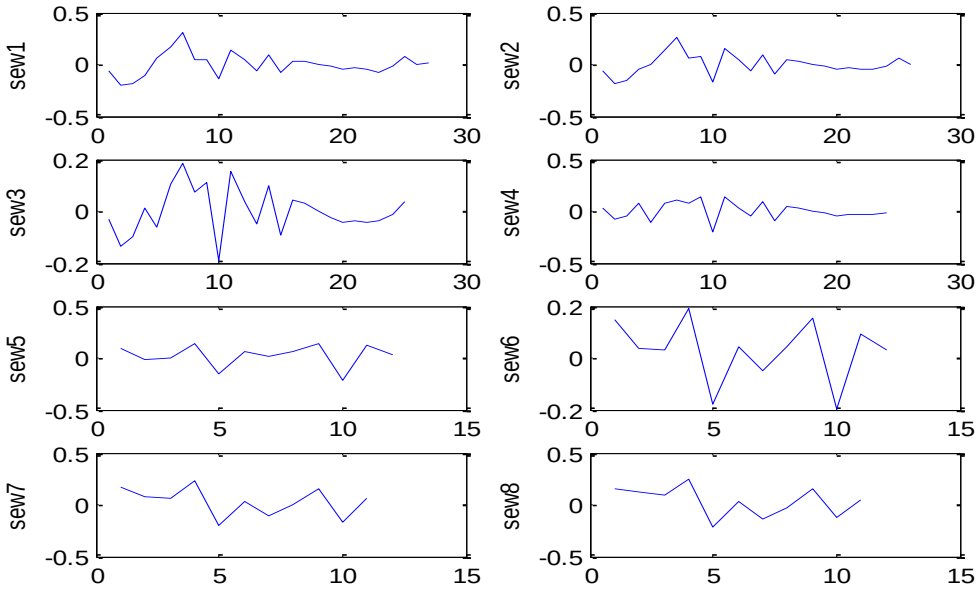

Fig.4.7 : 8 SEW d'un bruit (signal non voisé).

D'après ces figures nous constatons qu'il y a une grande ressemblance de forme entre les SEW, notamment dans le cas d'un signal voisé, de plus nous remarquons sur les SEW ; L'égalité entre leurs pitch (période), malgré la différence de quelques échantillons en longueur dans certains cas.

En effet, nous avons la possibilité d'éliminer certaines SEW dans le codeur, d'une manière par laquelle on peut les recréer ultérieurement au niveau du décodeur.

Pour réaliser une interpolation linéaire, semblable à celle appliquée sur les LSF, nous avons choisi d'éliminer au niveau d'une trame les 6 SEW intermédiaires, et garder les deux SEW d'extrémités, ensuit de les ajuster à la même langueur pour qu'on puisse réaliser l'interpolation.

4.3. Ajustement de longueur des SEW

Cet Ajustement signifie la mise des deux SEW extrêmes qui seront stockées ou transmettes à une même longueur, ainsi, il est conçu pour deux raisons essentielles :

- une interpolation doit être réalisée sur deux vecteurs de même longueur.

- après interpolation ; pour utiliser les SEW dans la régénération de signal original (codé), chaque SEW doit avoir sa longueur originale qui est égale à la longueur du pitch correspondant.

Alors pour ne pas perdre les informations, nous cherchons la SEW de longueur maximale, et nous l'utilisons comme référence de longueur, on peut avoir trois cas possibles :

- La première SEW est de longueur maximale

On garde SEW1, et ses échantillons supplémentaires on les ajoute à la SEW8 pour la rendre à la même longueur que la SEW1.

- la huitième SEW est de longueur maximale

Dans l'inverse, on conserve SEW8 et on ajoute ses échantillons supplémentaires à SEW1.

- Une SEW intermédiaire à la longueur maximale

Dans ce cas, on calcule la différence de longueurs entre cette SEW, et SEW1 et SEW8, et puis on ajoute les échantillons supplémentaires à chaque SEW d'extrémité pour qu'elles soient de même longueur que cette SEW intermédiaire.

Il y a un autre avantage de ce type d'ajustement qui s'assimile dans la conservation des échantillons réels des SEW ajustées au lieu de créer d'autres échantillons aléatoires ou les remplacer par des zéros.

A cette étape nous avons fini le codage de SEW, ainsi nous avons remplacé les huit SEW d'une trame uniquement par deux SEW ajustées à une longueur maximale et contiennent les informations nécessaires pour régénérer les SEW éliminées à l'aide d'une interpolation linéaire.

En somme, le débit des SEW se réduit principalement à 1/4 et avec le sous échantillonnage de 1/4 cité au paragraphe (3.4.1.2) leur débit par conséquent diminue à 1/16.

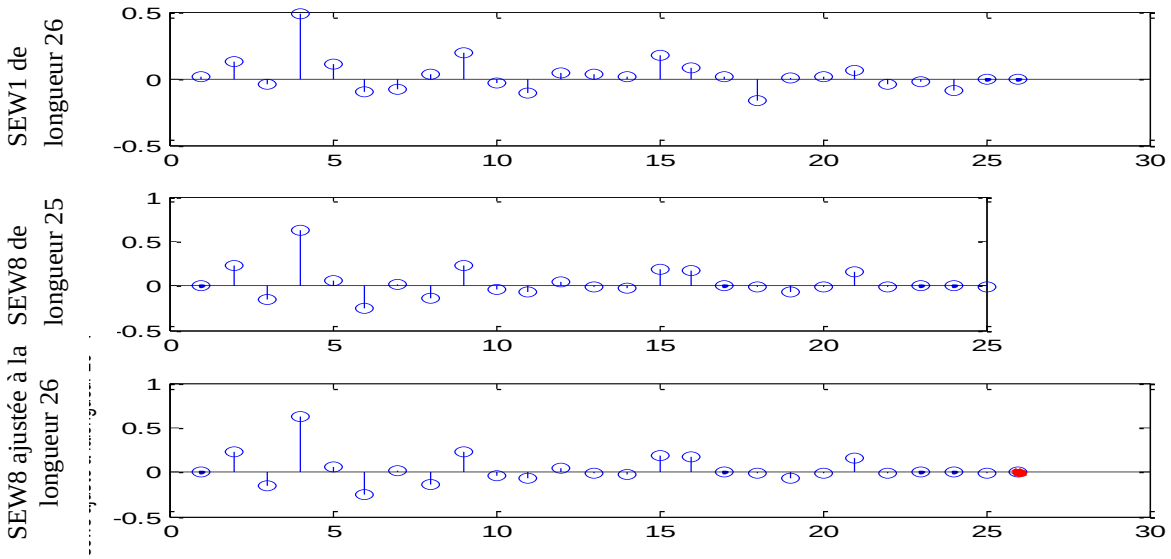

Fig. 4.8 : Ajustement de longueur (exemple où SEW8 est courte, nous avons ajouté un échantillon)

4.4. Interpolation des SEW

(Échantillon)

La partie de décodage consiste à interpoler linéairement les deux SEW avec la formule :

$$sew(i) = (sew1) * w(i) + (sew8) * \big(1 - w(i)\big) \tag{4.1}$$

Où $w(i)$ sont les composantes de vecteur d'interpolation linéaire

$$w(i) = \left[1; \frac{7}{8}; \frac{6}{8}; \frac{5}{8}; \frac{4}{8}; \frac{3}{8}; \frac{2}{8}; \frac{1}{8}\right]$$

Après interpolation, les six SEW résultantes sont à la longueur maximale, on utilise le vecteur qui contenue les longueurs des pitchs pour ajuster chaque SEW à sa longueur, on peut avoir deux cas :

– longueur du pitch inferieure a celle de la SEW : on élimine les derniers échantillons parce qu'ils sont supplémentaires.

– longueur du pitch égale à la longueur de SEW : On garde la SEW tell qu'elle est. Le troisième cas (longueur du pitch supérieure a celle de la SEW) est impossible parce que nous avons déjà utilise la SEW qui a la longueur maximale dans la trame (paragraphe 4.3).

Alors nous avons reconstruit une cellule de 8 SEW qui remplace la cellule original et qui sera ajoute vecteur par vecteur au REW pour reconstruire les CW, qui seront utilisés dans le décodage plus exactement à la reconstitution de signal parole codé. Les trois figures suivantes illustrent un exemple de huit SEW dans les trois étapes de codage (originales – interpolées – réduits à la longueur réelle).

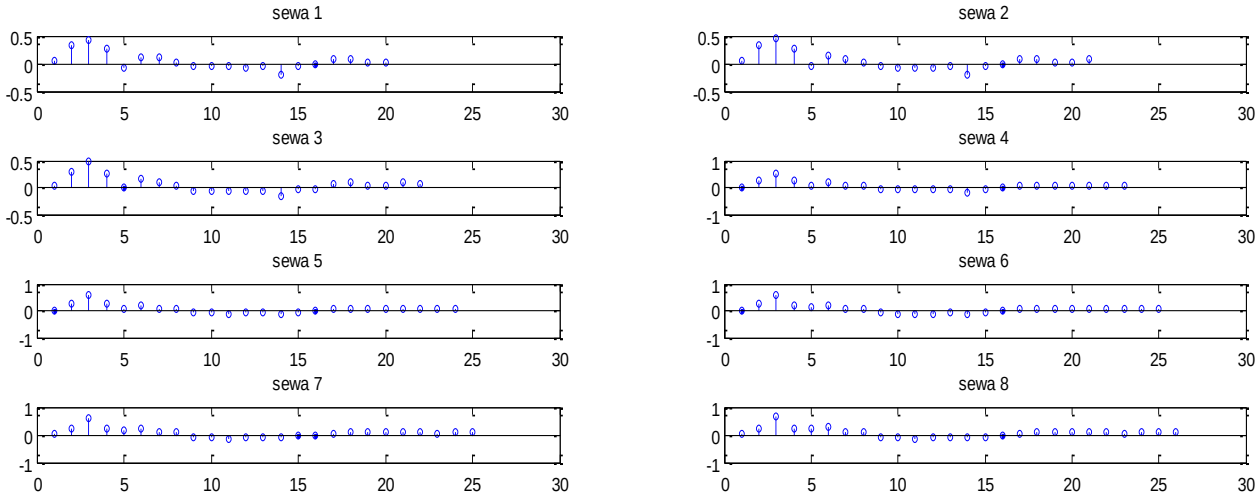

Fig.4.9 : 8 SEW originales en longueurs différents.

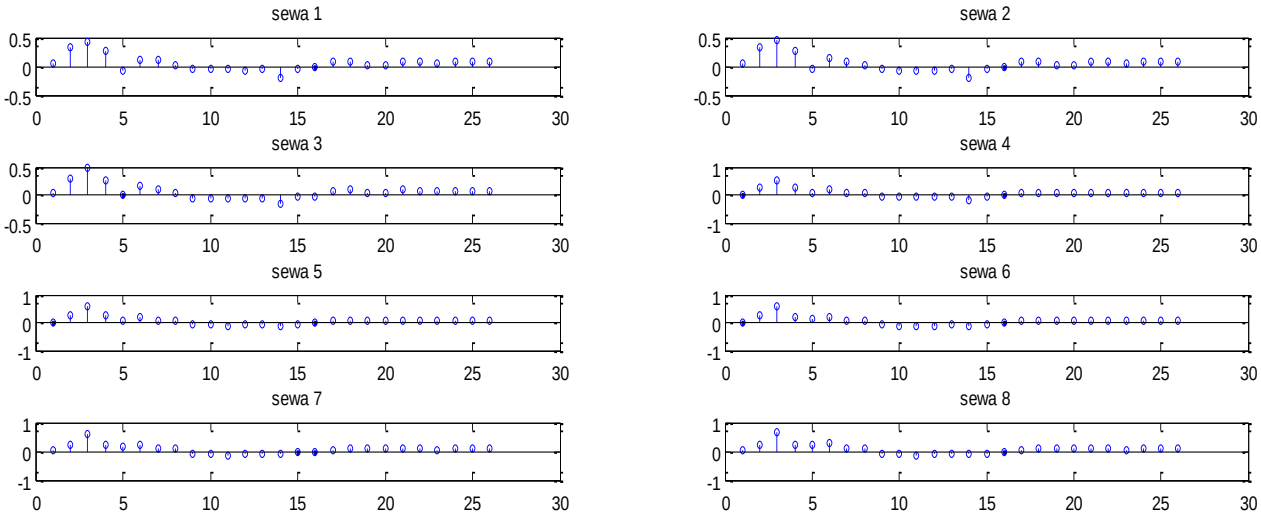

Fig.4.10 : 8 SEW après interpolation (tout les SEW ont la même longueur).

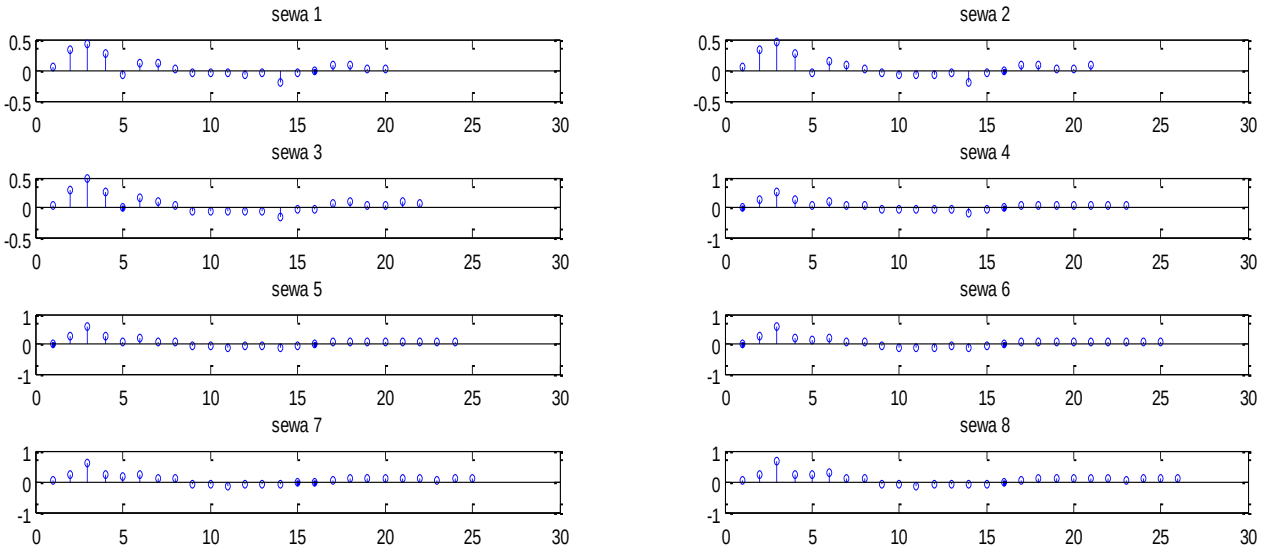

Fig.4.11 : 8 SEW codées (interpolées et ajustées à ses longueurs originales).

Les expériences d'écoute montre que le codage nous donne des SEW qui ont le même comportement comme les originales, avec une légère différence dans certains cas qui n'affect pas l'aspect perceptuel.

4.5. Principe de procédé d'interpolation

On peut résumer notre travaille avec un organigramme qui explique la démarche de procédure (figure 4.12 et figure 4.13).

Remarques

En pratique l'application de l'algorithme sur tout les partie du signal (voisé et non voisé) n'affecte pas son qualité grâce à :

* Si le signal est voisé : il y a ressemblance entre les huit SEW de trame, c'est-à-dire les six SEW interpolés remplacent les originales éliminées.

* Si le signal est non voisé : on a des petites variations entre les SEW originales et interpolées se variation n'affect pas la perceptibilité de parole parce que ces partie sont déjà aléatoires, donc les SEW interpolées seront suffisantes.

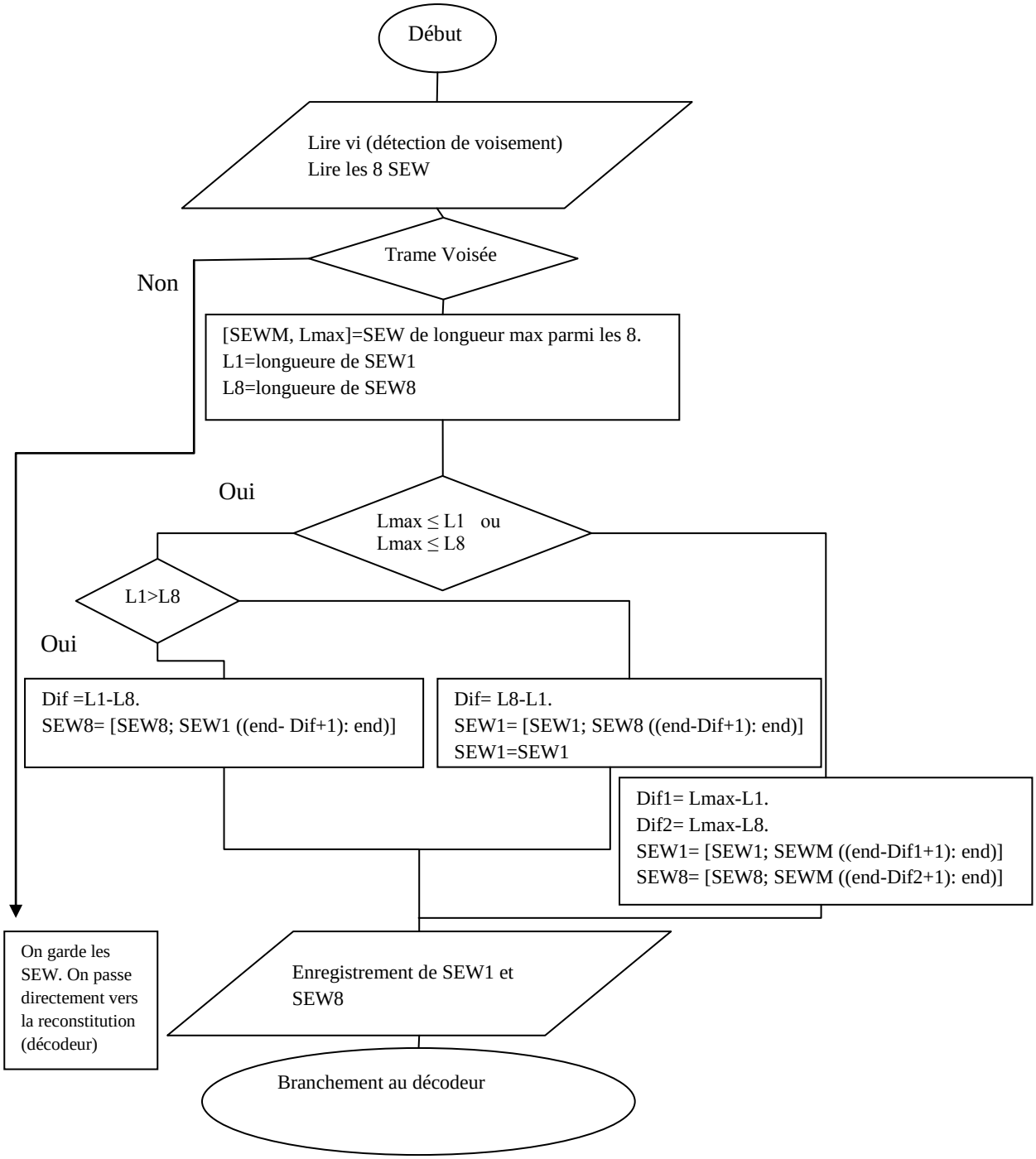

Fig.4.12 : Organigramme de sous échantillonnage des SEW.

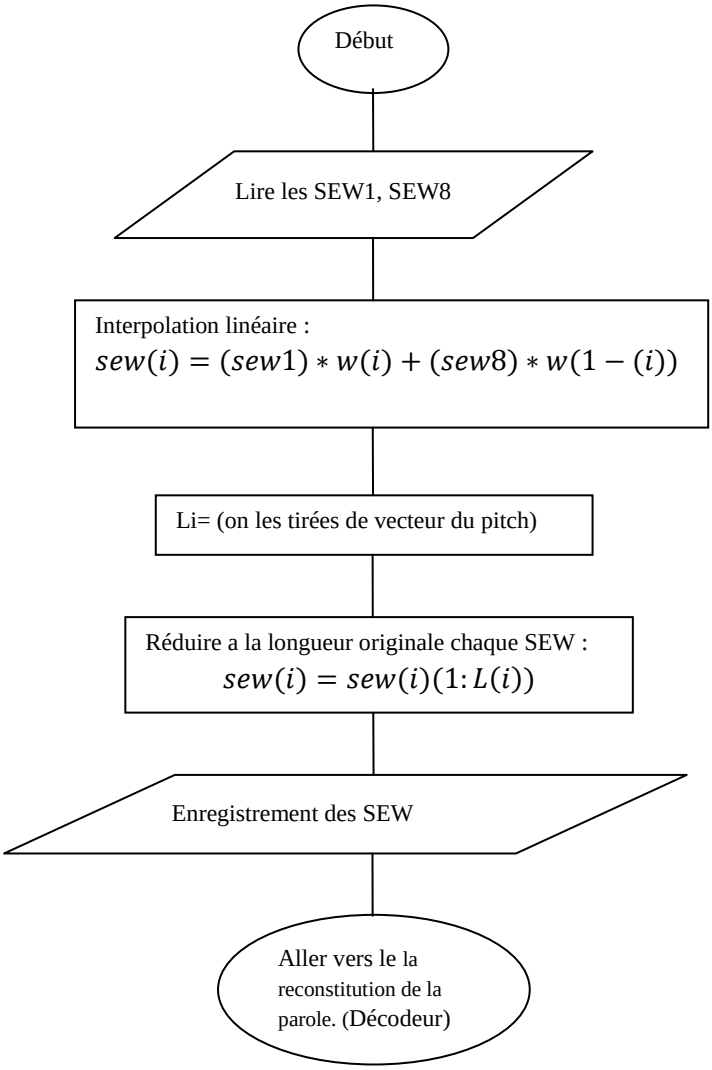

Fig.4.13 : Organigramme de sur échantillonnage et génération des SEW.

4.6. Évaluation des résultats

Pour évaluer notre codage ; nous avons appliqué et teste le codeur sur des signaux normals, des signaux qui son voise comme le (koran) et un signal bruité.

Evaluation subjective

L'évaluation subjective montre des bon résultats, les parole son intelligibles et la qualité perçue acceptable.

Evaluation objective

Généralement les tests objectifs ne conduisent pas à une bonne interprétation du résultat ; parfois la mesure du SNR entre la parole originale et celle codée donne des valeurs négatives.

Ces valeurs négatives des SNR, sont essentiellement dues à la désynchronisation de la WI dans le temps, entre le signal original et celui reconstitué [6]. Cette désynchronisation est interprétée par le fait, de la variation de point d'extraction des formes d'ondes, et l'inexistence d'une méthode, qui calcule la valeur exacte de la phase initiale $\phi(0)$. Par conséquent, il est nécessaire d'évaluer la qualité de la parole reconstruite par une mesure subjective.

Pour remédier à ce problème j'ai essaie de faire une synchronisation basse sur l'erreur minimal. On glisse un partie du signal codé sur le signal original, en enregistre l'erreur commet entre eux. A la fin de glissement on prend l'erreur minimal qui doit correspondre à la partie choisie et sa version codé, et on calcul le SNR entre ces deux morceaux.

L'inconvénient de cette méthode est le grande nombre d'itérations, par conséquence on ne peut pas prendre des parties longues du signal pour les comparées.

Un test de codeur sur cinq types de parole est présenté par le tableau ci-dessous, où les mesures des SNR sont relatives. Il ne compare que certaine trame et pas tout le signal.

Tab.4.1 : SNR relatifs (portion de signal reconstitué/portion de signal original)

Type de parole	**SNR sans interpolation**	**SNR avec interpolation**
homme	41.5480	37.8146
femme	71.3580	68.7286
Signal purement voisé	34.2590	34.2365
Signal bruit	24.5013	22.1098
Section de « koran » voisée	62.2499	61.6859

On tire à partir de tableau, que lorsque le signal est voisé le changement de SNR est très petit. C'est un résultat qui renforce notre proposition et montre la performance de cette méthode pour les signaux pseudo périodiques.

4.7. Évaluation de la performance

Dans cette section, nous donnons une présentation de l'allocation des bits pour un codeur WI standard travaillant à 3.85 Kbps à travers le tableau .4.2, notons que les paramètres sont quantifiés selon la description présentée dans le chapitre.3. Le débit de quantification des SEW est extrait à partir de [33].

Tableau .4.2 : Allocation de bits d'un codeur WI à 3.85 Kbps.

Paramètres	**Bits/trame**	**Bits/seconde**
Coefficients LPC	20	1000
Pitch	7	350
Gain	12	600
SEW (amplitude)	18	900
SEW (phase)	4	200
REW (amplitude)	16	800
REW (phase)	0	0
Total	77	3850

Un codeur standard alloue aux SEW, 18 bits/trame. Avec notre procédure le débit des SEW est réduit jusqu'à 1/4 parce que nous avons pris deux SEW parmi huit, c'est-à-dire 18/4 bits/trame. Ceci est très intéressent : seulement 225 bits/sec pour les SEW. Ce qui donne un codeur de 3.175 bits/trame, avec la possibilité de réduire encore le débit si on augment le nombre des SEW éliminées.

4.8. Conclusion

Dans ce chapitre nous avons profité de la caractéristique des signaux pseudo-périodiques qui est la grande redondance pour les adresser une compression qui convient avec cette propriété.

Comme première étape, nous avons suivi les variations des SEW dans le cas où les trames sont voisées, nos voisées et bruit pur. Nous avons remarqué la grande ressemblance entre les huit formes SEW d'une trame voisée, une différence non importante dans le cas des trames non voisées, mais une différence complète dans le cas de bruit hors que la parole.

Cette observation nous a donné l'idée d'appliquer la duale décimation - interpolation. Comme on a huit SEW ressemblant, nous avons choisi d'éliminer les six SEW intermédiaires comme décimation et garder les deux SEW d'extrémités à condition qu'elles soient ajustées aux mêmes longueurs. Et comme interpolation et reconstitution des SEW éliminées, nous avons appliqué la formule d'interpolation entre les deux SEW conservées.

Les paroles codées par cette méthode sont parues acceptables surtout avec l'évaluation subjective, elles sont intelligibles et ne marquent aucune variation perceptuelle remarquable. L'évaluation objective présente le problème de désynchronisation et donne des valeurs insignifiantes, on a remédié à ce problème par un calcul des SNR relatifs qui ont des valeurs comprises dans les normes. De point de vue débit, la procédure offre au codeur la possibilité de travailler avec un débit environ de 3 kbits/sec.

Conclusion générale :

Durant ce travail nous avons montré l'efficacité des codeurs hybrides, ainsi que la performance des méthodes d'interpolation dans le domaine de la compression à l'égard de relations évidentes entre elles. La compression cherche à éliminer les parties des signaux insignifiantes et garder les parties qui ont une grande quantité d'informations, l'interpolation quant cherche à éliminer les parties qui sont redondantes avec la possibilité d'être générées.

L'interpolation peut suivre les redondances et les éliminer, ce qui implique un grand abaissement de débit qui est proportionnel aux redondances, et par conséquent donne de meilleures solutions pour les signaux redondants avec de fortes corrélations (périodiques, pseudo périodiques).

Nous avons exposé une initiation générale sur la compression, représenté les caractéristiques générales du signal parole ainsi que son prétraitement et son traitement, puis le type de modélisation utilisé avec des détails de l'analyse LPC qui est la base du codage paramétrique.

Le but de notre travail est d'améliorer le codeur WI, par conséquent nous avons présenté le codeur WI à son état final, ainsi que son principe de codage et décodage et sa gestion des paramètres. Parce que la WI nécessite la quantification des ses paramètres chacun suivant sa nature, la quantification des paramètres du codeur a été détaillée et plus particulièrement la quantification des composantes REW.

Nous avons remarqué dans le codeur WI que les formes SEW ont besoin de réduction de leurs débits à cause des redondances qu'elles contiennent. Bien sûr, que l'outil le plus convenable dans ces cas est l'interpolation.

Dans ce but nous avons associé à la partie codage une décimation des formes SEW qui ne contient pas des informations significatives et une interpolation créative des parties tronquées à partir des parties conservées vues comme signifiantes, à l'étape de décodage.

Les résultats obtenus par cette procédure sont très intéressants grâce à la grande redondance existant au niveau des paroles voisées. Nous avons réussi à tronquer des parties de signal sous forme des SEW, puis les interpoler sans toucher à la qualité perceptuelle du signal. Du point de vue débit, si cette méthode est appliquée dans un codeur standard qui travaille avec 3.85 Kbits/s et (18 bits/trame pour les SEW), on arrive à un débit de 3.175 Kbits/s.

Tous les tests sont réalisés à l'aide du logiciel de simulation MATLAB.

Perspectives :

Finalement nous espérons, que ce mémoire sera une aide avantageuse, pour tous ceux qui veulent approfondir leurs idées dans l'amélioration des performances de codeur WI. Le codeur est souple et peut subir d'autres modifications.

Deux points nous motivent dans nos travaux futurs:

- élargir son domaine d'application sur d'autres signaux autre que la parole qui présentent de fortes corrélations.
- application de méthodes d'interpolation plus avancées que la méthode linéaire afin d'augmenter la qualité des signaux résultant de l'interpolation.

Bibliographies

[1] C.E. Shannon, A mathematical theory of communication, Bell System 1948.

[2] W.B. Kleijn and W.Granzow, Methods for Waveform Interpolation In Speech Coding, AT & T Bell Laboratories Digital Signal Processing 1, pp. 215-230, 1991.

[3] G. BAUDOIN, J. CERNOCKY, Ph. GOURNAY, G.CHOLLET, Codage de la parole à bas et très bas débits, Département Signaux et Télécommunications, ESIEE, BP 99 93162 Noisy Le Grand cedex, ANN.TÉLÉCOMMUN., 55, n°9-10, 2000

[4] T. Azzedine, codage d'un signal parole dans le codeur à interpolation de la forme d'onde, mémoire de magister, laboratoire de traitement de signal et communication, ENSP Algérie, 2007.

[5] L'union International des Télécommunications, UIT-T Série P Méthodes d'évaluation Objective et Subjective de La Qualité, Juillet 2006.

[6] L. T. Choy, Waveform Interpolation Speech Coder at 4 kb/s, Department of Electrical & Computer Engineering McGill University Montreal, Canada August 1998.

[7] W. B. Kleijin and J. Haagen, A Speech Coder Based on Decomposition of Characteristic Waveforms, Information Principles Research Laboratory, AT & T Laboratories, Murray Hill. NJ 07974, USA, pp 508-511, IEEE, 1995.

[8] R.Boite et M.Kunt,"*Traitement de la parole*", Presses Polytechniques Romandes, première édition(1987).

[9] Calliope, La parole et son traitement automatique, Masson, 1989.

[10] G. Fant, Acoustic theory of speech production, s-gravenhage, Mouton, 1949

[11] G. H. Golub and C. F. Van Loan, Matrix Computations. Baltimore, Maryland: The John Hopkins University Press, third ed., 1996.

[12] M. A. Khan,Coding of Excitation Signals In a Waveform Interpolation Speech Coder, Department of Electrical & Computer Engineering McGill University Montreal, Canada, Thèse pour obtenir le Master, July 2001

[13] M. Leong, Representing Voiced Speech Using Prototype Waveform Interpolation for Low-rate Speech Coding, Department of Electrical & Computer Engineering McGill University Montreal, Canada November 1992.

[14] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speechco ders," IEEE J. Selected Areas in Comm., vol. 10, pp. 819–829,June 1992.

[15] F. Itakura, "Line spectrum representation of linear predictive coefficients of speech signals," Journal Acoustical Society America, vol. 57, p. S35, Apr. 1975.

[16] T. Islam, Interpolation of Linear Prediction Coefficients for Speech Coding Department of Electrical Engineering McGill University Montreal, Canada April, 2000.

[17] P. Kabal and R. P. Ramachandran, "The computation of line spectral frequencies using chebyshev polynomials," IEEE Trans. Acoustics, Speech, Signal Processing, vol. ASSP-34, pp. 1419–1426, Dec. 1986.

[18] F. K. Soong and B.-H. Juang, "Line Spectrum Pair (LSP) and speechdata compression," Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing, pp. 1.10.1–1.10.4, Mar. 1984.

[19] M. Jelinek, Modélisation Spectrale et Compression de Parole a Bas Débit, Thèse de Doctorat Spécialité: génie électrique, Université De Sherbrooke, Octobre 1998.

[20] Enhanced Variable Rate Codec, Speech Service 2 Option 3 and 68 for Wideband Spread Spectrum 3 Digital Systems, May 2006.

[21] E. López-Soler, N. Favardin. A combined quantization-Interpolation scheme for Very Low bit rate coding of speech LSP parameters. Proc. ICASSP-93. pp.II-21-24, 1993.

[22] S.V. Vaseghi, Advanced Digital Signal Processing and Noise Reduction, Second Edition, John Wiley & Sons Ltd ISBNs: 0-471-62692-9 (Hardback):0-470-84162-1 (Electronic), 2000.

[23] WAI C. CHU, San Jose, California, Speech Coding Algorithms Foundation and Evolution of Standardized Coders, Mobile Media Laboratory DoCoMo USA Labs, John Wiley & Sons ,2003

[24] S.K. Mitra, Digital Signal Processing a Computer-Based Approach, Department Of Electrical And Computer Engineering, University Of California, Santa Barbara ,McGraw-hill.

[25] R.l. Allen, D.w. Mills, (signal analysis Time, Frequency, Scale, And Structure), the institute of electrical and electronics engineers, Canada, IEEE, 2004.

[26] Y. Shoham, Very Low Complexity Interpolative Speech Coding at 1.2 to 2.4 Kbps, Acoustic and Audio Communication Dept Bell Laboratories, Lucent Technologies, 700 Mountain Ave. Murray Hill NJ 07974 USA, pp 1599-1602 IEEE, 1997.

[27] R. Matmti, M. Jelinek, J.P. Adoul, Modulation De L'excitation Pour Codage De La Parole a Très Bas Débit (<4 bits/sec), Group De Recherche Information, Signal Et Ordinateur, Université De Sherbrooke, Canada 1995

[28] O. Gottesman, Member, IEEE, and A. Gersho Fellow, IEEE, Enhanced Waveform Interpolative Coding at Low Bit-Rate Vol. 9, No. 8, November 2001.

[29] Kyung Jin Byun, Ik Soo Eo, He Bum Jeong, and Minsoo Hahn, A Novel Dimension Conversion for the Quantization of SEW in Wideband WI Speech Coding, Springer-Verlag Berlin Heidelberg 2005.

[30] O. Gottesman and A. Gersho, High Quality Enhanced Waveform Interpolative Coding at 2.8 kbps, in Proc. 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing, Istanbul, pp 1363−1366.Turkey, June 2000.

[31] Jin Kyu Choi, Chang-Heon Lee, Hong-Goo Kang, Young-Cheol Park and Dae Hee Youn Improvement Issues on Transcoding Algorithms for the Flexible Usage to the Various Pairs of Speech Codec, MCSP Lab., Yonsei University / LG Electronics Inc., Korea IEEE, 2004.

[32] J. Nurminen, A. Heikkinen, and J. Saarinen, Objective Evaluation of Methods for Quantization of Variable Dimension Spectral Vectors in WI Speech Coding, in Proc. Eurospeech 2001, Aalborg, Denmark, pp 1969−1972, September, 2001.

[33] J. Nurminen, A. Heikkinen, and J. Saarinen, Quantization of Magnitude Spectra in Waveform Interpolation Speech Coding, Institute of Digital and Computer Systems, Tampere University of Technology, speech and audio systems laboratory, Nokia Research Center, Finland.

[34] P. Lupini and V.Cuperman, Subjective Performance of Spectral Excitation Coding of Speech at 2.4kbits/s School of Engineering Science, Simon Fraser University, Burnaby, BC, Canada.

[35] M. Leong and P. Kabal, Smooth speech reconstruction using Prototype Waveform Interpolation, Proc. IEEE Workshop on Speech Coding for Telecom. pp. 39-41, October 1993.

[36] H. FARSI, Speech Pre-Processing for Pitch and Pitch-Cylce Evolutions Smoothing, Department of Electronic and Electrical Eng, Faculty of Eng, University of Birjand, IRAN, 2006.

[37] O. yousef, aide mémoire de mathématiques pour ingénieurs, office des publications universitaires 1994.

[38] K.K.Paliwal and B.S.Atal, Efficient Vector Quantization of LPC Parameters at 24 Bits/Frame, AT & T Bell Laboratories Murray Hill, NJ 07974, IEEE, 1991.

[39] P. Gournay , F. Chartier , a 1200 bps HSX Speech Coder for Very Low Bit Rate Communications, IEEE Workshop on Signal Processing System SiPS'98, Boston, 1998.

[41] L. Buniet, Traitement Automatique de la Parole en Milieu Bruité, Université Henri Poincaré, Février 1997

[42] P. Noll, Speech, and Audio Coding for Multimedia Communication, Proceeding International Cost 254 Workshops on Intelligent Communication Technologies and Application university, Berlin,1999.

[43] N.B. Beng, Robust Spectral Coding in Speech Processing Department of Electrical Engineering McGill University Montreal, Canada May 1998.

[44] W. B Kleijn and J. Haagen, A General Waveform-Interpolation Structure, Proc. European Signal Processing Conf. (Edinburg), pp 1665-1668, September 1994.

[45] D. Wellens, Implémentation et Optimisation d'un Algorithme de Compression Sans Perte Université Libre de Bruxelles, Juin, 2003.

[46] A. Gersho, R.M. Gray, Vector quantization and signal compression, Englewood Cliffs, N.J. Prentice-Hall, 1990.

[47] Y. L. Linde, A. Buzo and R. M. Gray, An algorithm for vector quantizer design, IEEE Trans. Commun., vol. COM-28, pp. 84–95, Jan. 1980.