REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

ECOLE NATIONALE POLYTECHNIQUE

DEPARTEMENT D'ELECTRONIQUE

Laboratoire de Traitement du Signal et Communications

MEMOIRE DE MAGISTER

Option : Traitement du Signal et Communications

Présenté par : TOUAZI AZZEDINE

Ingénieur d'Etat en Electronique (U.S.T.H.B)

Thème :

Codage d'un Signal Parole dans le Codeur à Interpolation de Formes d'Onde

Soutenue le 28 octobre 2007 devant la commission d'examen composée de :

Présidente du jury :

Mme. L. HAMAMI…….…....….Maître de conférences, ENP d'Alger.

Rapporteur :

Mr. D. BERKANI……..…Professeur, ENP d'Alger.

Examinateurs :

Mr. H.BOUSBIA-SALAH…….……... Maître de conférences, ENP d'Alger.

Mlle. M.GUERTI…….…….……Maître de conférences, ENP d'Alger.

Mr. A.NASRI…….……Chargé de cours, Université de Boumerdès.

Année Universitaire 2006-2007

Remerciements

Ce travail n'aurait certainement pas pu être mené à bien sans le soutien, le concours et les conseils de nombreuses personnes. Je tiens à les remercier très sincèrement.

Au terme de ce mémoire, je tiens à remercier et exprimer ma sincère reconnaissance à monsieur D. Berkani, Professeur à l'ENP, pour le temps précieux qu'il m'a accordé en acceptant d'être mon rapporteur.

Je tiens, ensuite, à exprimer toute ma gratitude à madame L. HAMAMI, Maître de conférences à l'ENP, d'avoir accepté de présider le jury de ce mémoire. Un grand merci à monsieur H.BOUSBIA-SALAH, mademoiselle M.GUERTI, tout deux, Maîtres de conférences à l'ENP, et à monsieur A. NASRI, Chargé de cours à l'Université de Boumerdès qui ont été examinateurs. Leurs critiques constructives m'ont été particulièrement précieuses dans l'élaboration finale de ce mémoire.

J'exprime aussi ma reconnaissance à toutes les personnes, évoluant au sein du Laboratoire de traitement de signal et communications, qui m'ont apporté leur soutien, leur amitié et leur expérience tout au long de ce mémoire. Je remercie aussi tous mes professeurs pour leur encouragement durant tout mon parcours éducatif.

Enfin, je remercie aussi vivement mes parents, mes frères et sœurs et tous mes amis, qui de près ou de loin m'ont soutenu tout au long de ce travail.

الملخص

أخضعت أجهزة الاتصالات الحديثة مشفرات الكلام بتدفقات تقارب 4 ك.ب/ثا، إلى توسيع مجالات استعمالاتها، مثل الهاتف المرئي و الاتصالات المتنقلة. في العشرية الأخيرة، تم اقتراح عدد مختلف من الطرق لتشفير الكلام. العمل المعرض في هذه المذكرة يهدف أولا إلى دراسة المشفر الصوتي الذي يعتمد أساسا على مبدأ استكمال شكل الموجة WI، لقد كانت نوعية الأصوات الناتجة عن عملية التصور لهذا الأخير بدون عملية التكميم ذات كفاءة و شفافية عاليتين. ثانيا قمنا بعرض و دراسة طريقة تكميم المركبة المتغيرة بسرعة REW، وتم ذالك اعتمادا على عملية التحويل ألجيبي المتقطع DCT و تقنية تقريب الدوال بعوامل كثير حدود على طريقة المربعات الصغرى.

كلمات مفتاحية : تشفير الكلام، استكمال شكل الموجة , عملية التكميم، المركبة المتغيرة بسرعة، عملية التحويل ألجيبي المتقطع، تقريب الدوال بعوامل كثير حدود، طريقة المربعات الصغرى.

Abstract

The modern communications systems subjected speech coders at bit rates around 4 kbps is expected to be widely used in applications such as visual telephony, mobile communications. During the last decade, a variety of speech coding techniques have been proposed. The work presented in this memory, study the general performance of a speech coder, which it based on the waveform interpolation scheme WI. The WI coder has been simulated without quantization; the simulation of the speech coder gave a good quality with one elevated degree of intelligibility and of natural. Secondly, we presented a method of quantization of the rapidly evolving waveform component REW, this method essentially based on the discrete cosine transform DCT and the polynomial curve fitting technique based on the principle of the least squares Principle.

Key words: Speech coding, Waveform interpolation. Quantification, Rapidly evolving waveform, Discrete cosine transform, Polynomial curve fitting, Least squares Principle.

Résumé

Les systèmes de communications modernes ont soumis les codeurs de la parole à des débits voisins de 4 kbps, d'être largement employé dans des applications telles que la téléphonie visuelle et les communications mobiles. Pendant la dernière décennie, une variété de techniques de codage de la parole ont été proposées. Le travail présenté dans ce mémoire, étudie le fonctionnement général d'un codeur de la parole basé sur le schéma d'interpolation de la forme d'onde WI. Un codeur WI a été simulé sans quantification ; les résultats de simulation ont donné une parole de bonne qualité avec un degré élevé d'intelligibilité et de naturel. De plus nous avons présenté une méthode de quantification de la composante a évolution rapide REW, qui se base essentiellement sur la transformée en cosinus discrète DCT et la technique d'approximation polynomiale basée sur le principe des moindres carrés.

Mots clés : Codage de la parole, Interpolation de la forme d'onde, Quantification, Composante a évolution rapide, La transformée en cosinus discret, approximation polynomiale. Principe des moindres carrés.

Liste des Abréviations

ACELP	Algebraic Code Excited Linear Predictive.
ADPCM	Adaptive Differential Pulse Code Modulation.
AR	Auto Regressive.
CELP	Code-Excited Linear Prediction.
CODEC	Encoder and Decoder.
CS-ACELP	Conjugate Structure Algebraic CELP.
CW	Characteristic Waveform.
DCVQ	Dimension Conversion Vector Quantization.
DCT	Discrete Cosine Transform.
DFT	Discrete Fourier Transform.
DSP	Digital Signal Processing.
DTMF	Duel tone multi frequencies
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.
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.
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

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, permet aussi de réduire la mémoire nécessaire dans les systèmes de stockage de la parole ; mais la volonté d'avoir une représentation numérique de la parole à faible débit, n'est pas souvent compatible avec la demande d'une reconstitution de la parole de haute qualité.

Le défi actuel est de proposer des codeurs de parole, permettant une quantification autours d'un taux de 4 kbps. Il est bien connu que la qualité de la parole à base d'algorithmes CELP (Code-Excited Linear Prediction), se détériore rapidement pour des débits au-dessous de 4 kbps. 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 reconstitution fidèle de la parole à des débits aussi faibles que 4 kbps. Ainsi, l'interpolation d'un signal parole [1] (abrégé WI pour waveform Interpolation), parait une technique efficace.

Dans le paradigme 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 ; dans ce codeur la parole est représentée par l'enchaînement des SEW (Slowly Evolving Waveforms) pour les segments voisés et REW (Rapidly Evolving Waveforms) pour les segments non voisés.

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

Objectif de Notre Travail :

Le premier objectif dans ce travail, est de présenter et d'étudier une méthode rehaussée de la quantification de la composante rapide REW, cette méthode est basée particulièrement sur la transformée en cosinus discrète DCT (Discrete Cosine Transform), et les techniques de représentations polynomiales des fonctions, tel que la méthode d'ajustement polynomiale

(curve fitting polynomial); avec une tentative d'atteindre une qualité de parole perçue acceptable.

Avec l'addition de peu d'affinages, un codeur WI complet est simulé sous environnement Matlab, en tenant compte des différents dérivés de la WI. Par ailleurs, on prend en considération les travaux réalisés par L.T. choy et M. Leong [2,3], nous identifierons certaines problématiques dans le codeur, qui sont concentrées principalement sur la haute qualité de reconstitution de la parole.

Composantes du Mémoire :

Ce mémoire comprend quatre chapitres ; une brève description du système auditif humain, les constituantes principales d'un signal parole, et des notions générales de codage de parole, sont données dans le premier chapitre.

Dans le deuxième chapitre, nous avons octroyé une vue d'ensemble de la théorie de base d'analyse de la prédiction linéaire à court terme avec les méthodes conventionnelles d'obtention des coefficients LP. Les approches communes de la représentation et l'interpolation des paramètres spectraux sont aussi expliquées. Ensuite décrire le principe de la prédiction à long terme, l'effet de la préaccentuation et de l'expansion de la largeur de bande.

Le troisième chapitre introduit le concept et la structure complète de l'algorithme WI. Un bref historique de l'évolution de l'algorithme est donné. Cette partie présente alors la mise en oeuvre de l'algorithme, avec une accentuation sur la couche analyse-synthèse, Chacun des blocs algorithmiques est discuté en détails. Nous avons examiné les résultats de simulations pour les différents dérivés de la WI.

Dans le quatrième Chapitre, le schéma général de la couche quantification est fourni avec une brève description des principes de base de la quantification ; une vue générale de la quantification classique des paramètres d'un codeur WI a été donnée. La deuxième partie de cette section décrit la mise en oeuvre de la quantification de la composante rapide REW.

Enfin une dernière partie conclut ce travail, et expose quelques suggestions pour les futurs travaux.

Chapitre 1 :

Introduction aux Principes de Base du Codage de la Parole

1.1. Introduction

En raison des caractéristiques du conduit vocal humain, le signal de parole est fortement redondant. Ces redondances permettent aux algorithmes de codage de compresser le signal en enlevant l'information non pertinente contenue dans le signal. Donc la connaissance du système vocal et des propriétés du signal de parole est essentielle pour concevoir des codeurs efficaces.

Les propriétés du système auditif humain peuvent également être exploitées, pour améliorer la qualité perceptuelle du signal codé. Avant d'aborder le problème de codage de parole plus précisément, quelques caractéristiques du signal vocal sont présentées, et permettront de mieux apprécier les différentes techniques de codage présentées par la suite.

Dans cette section, une partie simple de la théorie acoustique est présentée et les notions de phonème, de formant, de son voisé, non voisé et de pitch sont définis.

1.2. Caractéristiques d'un signal vocal

Le signal de parole étant un signal réel, continu, d'énergie finie et non stationnaire, les caractéristiques du signal de parole et du conduit vocal évoluent dans le temps. Les positions du système phonatoire agissent comme une opération de filtrage, en augmentant certaines fréquences tout en atténuant d'autres. Dans ce qui suit nous donnerons une description succincte du système vocal.

1.2.1. L'appareil phonatoire

L'appareil phonatoire nous permet de produire des sons très variés dans un espace fréquentiel et énergétique pourtant limité. L'appareil phonatoire humain a été la base de recherches visant à simuler mécaniquement ses capacités, de différentes recherches ayant permis en retour de mieux comprendre son fonctionnement.

La production de la parole est assurée chez l'être humain, par plusieurs organes successifs. Les poumons sont indispensables dans ce processus puisqu'ils assurent la génération d'un composant incontournable de l'air sous pression. Cet air expulsé, traverse alors les cordes vocales qui entrent ou non en action pour produire un voisement, ce voisement correspond à la fréquence fondamentale qui est le timbre de la voix ; cette fréquence fondamentale étant produite, elle est propagée dans l'ensemble du conduit vocal ; ce conduit est de forme et de volume variable.

Plusieurs organes concourent aux possibles modifications du conduit vocal, qui permettent de produire des sons différents. Parmi ces organes en trouve la langue acteur principal de ces modifications qui peut agir par constriction ou occlusion du conduit vocal. Les dents et les lèvres agissent également par occlusion ou constriction, à des degrés cependant moindres [4].

Le conduit vocal est la plupart du temps, constitué du seul conduit buccal. La luette et son prolongement vers le palais, le vélum, assurent notamment la fermeture du conduit nasal pendant la production de la parole. Le conduit nasal peut dans certains cas, être connecté au conduit vocal. Cette connexion permet de générer des sons supplémentaires (voisés ou non voisés) en modifiant le volume de la caisse de résonance qui est constituée par le conduit buccal. Une coupe de l'appareil phonatoire humain est illustrée dans la figure 1.1.

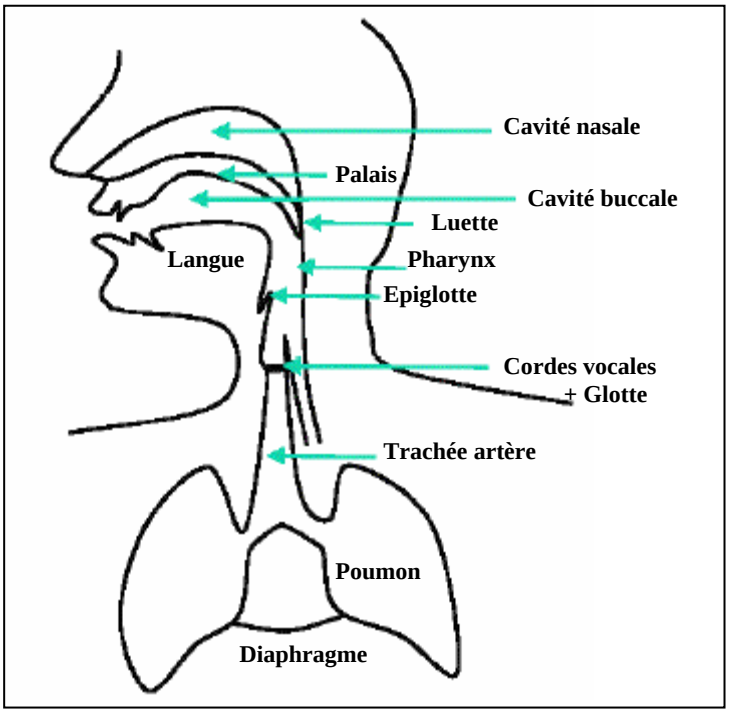


Fig. 1.1: Coupe de l'appareil phonatoire humain [5].

1.2.2. L'appareil auditif

L'appareil phonatoire, émetteur d'informations, ne serait d'aucune utilité si l'information générée ne pouvait pas être captée et analysée par un récepteur. L'oreille est divisée en trois parties distinctes [4], cette division se faisait en fonction de la distance par rapport à l'environnement aérien, porteur des sons ; où une première partie, l'oreille externe, correspond à la partie visible de l'organe, pavillon et lobe, à laquelle est rattaché le conduit auditif externe qui permet de propager le son jusqu'au tympan ; le tympan marque la frontière entre l'oreille externe et l'oreille moyenne. Les organes de l'oreille moyenne permettent de transformer les sons en vibrations grâce au contact qu'ils ont avec le tympan. Ces vibrations, une fois générées, sont transmises à la cochlée qui constitue l'organe majeur de l'oreille interne. La cochlée permet de transformer les vibrations en influx nerveux par le biais de cellules ciliées qui captent les vibrations produites dans le fluide de la membrane basilaire par l'étrier, le dernier os de l'oreille moyenne. Cet influx nerveux est alors transmis au cerveau en charge du traitement.

Une description détaillée de l'oreille (figure 1.2) permettra au lecteur de mieux comprendre les différents organes la constituant, et de mieux visualiser leurs répartitions.

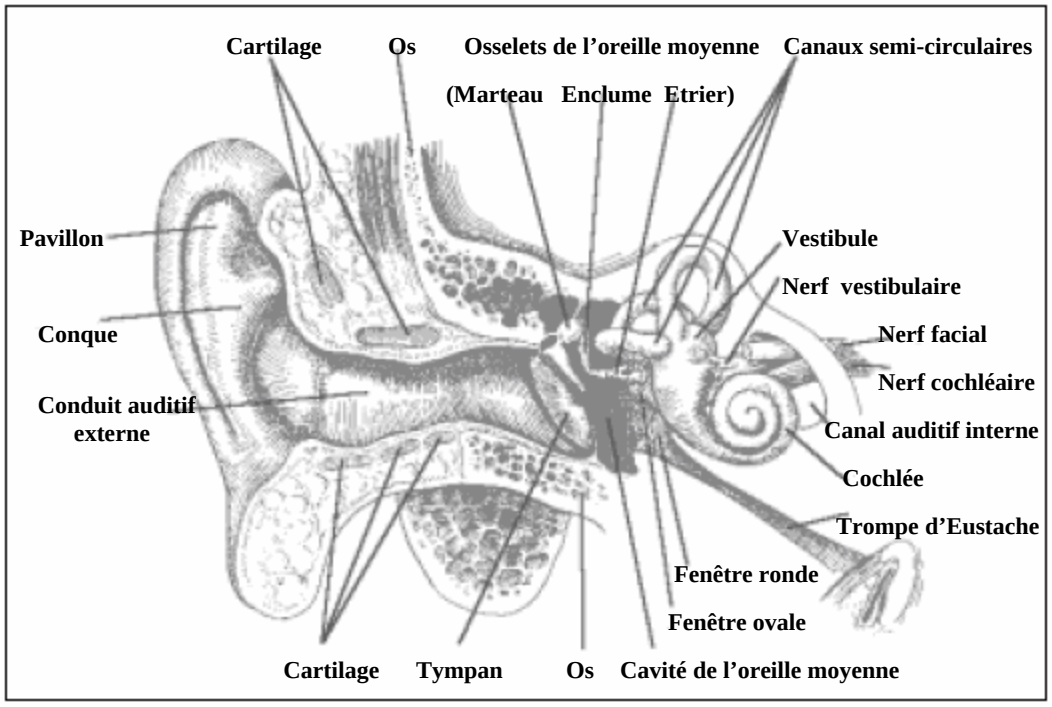


Fig. 1.2: Coupe de l'appareil auditif humain [4].

1.2.3. Sons voisés ou non voisés et fréquence fondamentale

On peut remarquer que, d'après ce que nous avons vu précédemment, le système phonatoire est composé de deux générateurs de sons, voisés, et non voisés, et d'un filtre (le conduit vocal) capable d'amplifier ou d'amortir certains sons.

En résumé, sans entrer dans les détails, un son voisé est un signal quasi périodique et un son non voisé peut être considéré comme un bruit blanc. La figure ci-dessous illustre un segment de parole qui contient une tranche voisée et une autre non-voisée. Dans le domaine spectral, les parties voisées du signal, apparaissent sous la forme de successions de pics spectraux, dont les fréquences centrales sont multiples de la fréquence fondamentale. Par contre, le spectre d'un signal non voisé ne présente aucune structure particulière.

La forme générale de ces spectres, appelée *enveloppe spectrale*, présente elle-même des pics et des creux qui correspondent aux résonances et aux anti-résonances du conduit vocal, et sont appelés *formants* et *anti-formants*. L'évolution temporelle de leur fréquence centrale et de leur largeur de bande détermine le timbre du son. Il apparaît en pratique que l'enveloppe spectrale des sons voisés est de type passe bas, avec environ un formant par 1kHz de bande passante, et dont seuls les trois ou quatre premiers contribuent de façon importante au timbre.

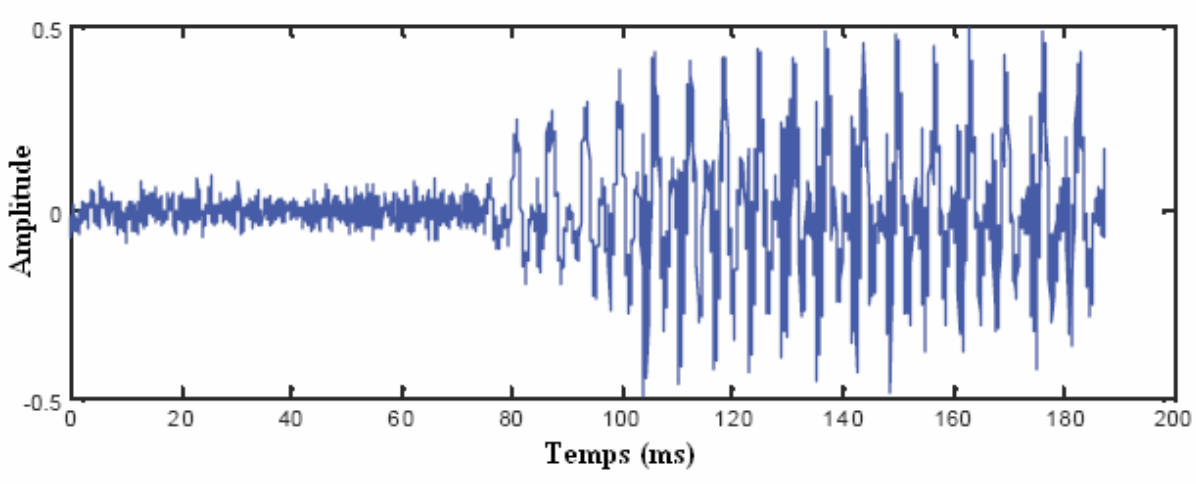


Fig. 1.3: Parties de son extraite à partir d'un audiogramme.
Segment non voisé (0-75 ms), et segment voisé (75-185 ms)

Cette manière de modéliser la parole est un peu sommaire mais permet de réaliser des modifications satisfaisantes des paramètres prosodiques. Ces derniers sont au nombre de trois :

- La fréquence fondamentale ou **pitch** dans les zones voisées.
- Le rythme d'élocution.
- L'intensité.

La fréquence fondamentale vient du fait que lorsque nous prononçons certains sons tel que [*b*], ou [*z*], on fait fonctionner les cordes vocales du larynx qui vibrent à une certaine fréquence quasi périodique. La fréquence fondamentale peut varier de : [4]

- 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'enfant.

Le rythme d'élocution correspond à la vitesse du débit de parole. On peut faire varier ce paramètre de manière à ce qu'une phrase prononcée trop rapidement puisse être ralentie pour la rendre plus compréhensible lors de l'apprentissage d'une langue étrangère par exemple. L'intensité du son émis est liée à la pression de l'air en amont du larynx.

1.3. Codeurs à bas et très bas débit

L'efficacité de la transmission se traduit par l'utilisation la plus restreinte possible des ressources pour émettre le signal, c'est-à-dire, par l'utilisation de peu de bits ou par l'occupation d'une bande de fréquences étroite.

De nombreux algorithmes ont été proposés pour diminuer ce débit, tout en essayant de conserver une qualité donnée en fonction des exigences de l'application à laquelle le codeur est destiné. On distingue en général trois plages de débits [6] :

- Les hauts débits, supérieurs à 16 kbit/s, correspondants à des algorithmes de codage de la forme d'onde non spécifiques à la parole.

- Les débits moyens, de 4 kbit/s à 16 kbit/s, correspondants à des techniques de codage hybrides utilisent des méthodes de codage de la forme d'onde et prennent en compte certaines propriétés de la parole ou de la perception auditive. Le principal représentant de cette classe est le codage CELP (Code Excited Linear Prediction).

- Les bas et très bas débits, de quelques dizaines de bits par seconde à 4 kbit/s, correspondants aux vocodeurs (Voice Coder) spécifiques au codage de la parole.

1.3.1. Codeurs à bas débit

Pour les bas débits, typiquement de 800 bps à 4000 bps, les techniques de codage de la forme d'onde ne donnent pas de bons résultats. Les codeurs doivent éliminer les informations sans pertinence à la perception. Les vocodeurs utilisent certaines caractéristiques de la perception et de la production de la parole, aussi sont-ils généralement très peu efficaces pour les signaux autres que la parole comme par exemple pour les signaux DTMF de numérotation téléphonique.

1.3.1.1. Codeur CELP

Le codage CELP a été introduit par Schroeder et Atal [7]. Il est très efficace pour les débits moyens de 4,8 kbit/s à 16 kbit/s, comme en témoignant les nombreuses normes qui l'utilisent. La figure 1.4 représente le principe du codage CELP.

Dans chaque trame, une analyse spectrale par prédiction linéaire détermine le filtre de synthèse $1/A(z)$. On découpe chaque trame en sous-trames plus courtes (durée typique 5 ms) sur lesquelles on effectue une quantification vectorielle du signal par une technique d'analyse par synthèse. On compare à l'aide d'un critère dit « perceptuel [1]» de type moindres carrés pondérés, le signal de parole original avec tous les signaux synthétiques possibles obtenus après quantification vectorielle. Ces signaux synthétiques sont générés en filtrant par le filtre

[1] Le terme perceptuel indique un critère ou un filtre, essayant de tenir compte de la perception auditive.

de synthèse un signal d'excitation choisi dans un dictionnaire de séquences d'excitation (on ajoute parfois la sortie de plusieurs dictionnaires), et en ajustant le signal résultant par le gain optimal.

Le codeur transmet le ou les index des segments qui minimisent le critère de perception, ainsi que le ou les gains associés, les paramètres spectraux et le pitch fractionnaire. Le critère perceptuel prend en compte la propriété de masquage du bruit de quantification par les formants en pondérant plus fortement l'erreur de quantification dans les zones de faible amplitude du spectre et plus faiblement dans les zones de formants. Cette pondération s'effectue en filtrant le signal d'erreur par un filtre de type $A(z)/A(z/\gamma)$ où γ est compris entre 0 et 1 (typiquement $\gamma = 0.85$).

Les dictionnaires utilisés sont appelés stochastiques ou adaptatifs selon qu'ils contiennent des séquences fixes de bruit ou bien les séquences d'excitation de trames précédentes. Le dictionnaire adaptatif permet de prendre en compte la redondance introduite par la quasi-périodicité des sons voisés.

La qualité subjective des codeurs CELP décroît rapidement lorsque le débit descend en dessous de 4 kbps [6]. En conséquence, le codage CELP effectue essentiellement une quantification vectorielle de la forme d'onde et pour un débit trop faible il n'est pas possible de coder cette forme d'une manière précise.

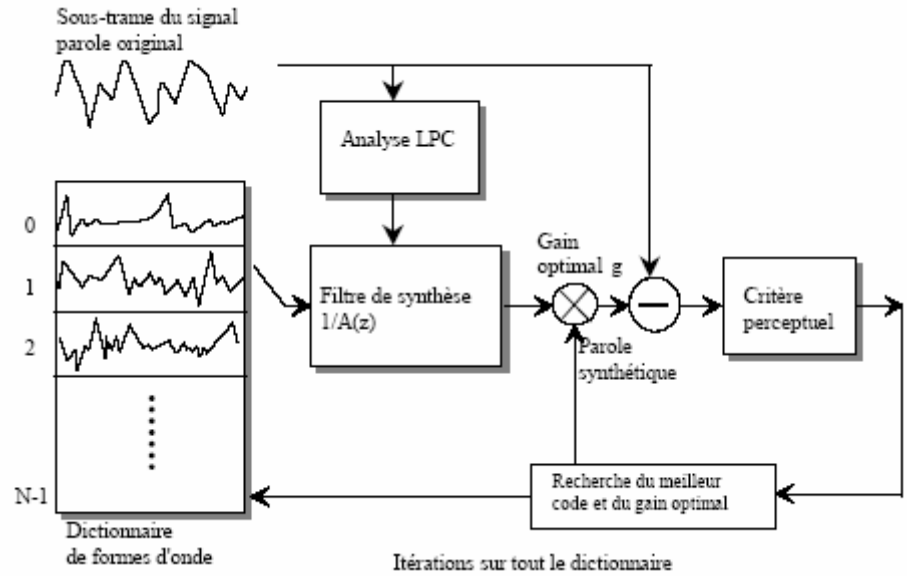


Fig. 1.4: Le codage CELP [6].

1.3.1.2. Vocodeurs classiques à deux états d'excitation

Dans les vocodeurs classiques, vocodeurs à canaux, vocodeurs à formants, ou vocodeurs LPC, les différentes trames de signal sont classées en trames voisées (V) et trames non voisées (NV). Ces vocodeurs classiques utilisent le modèle "source-filtre". La synthèse du signal décodé utilise un signal d'excitation reconstruit formé d'un bruit blanc pour les trames non-voisées, et d'un train périodique d'impulsions à la fréquence F_0 pour les trames voisées. La figure 1.5 représente le synthétiseur d'un vocodeur à deux états d'excitation.

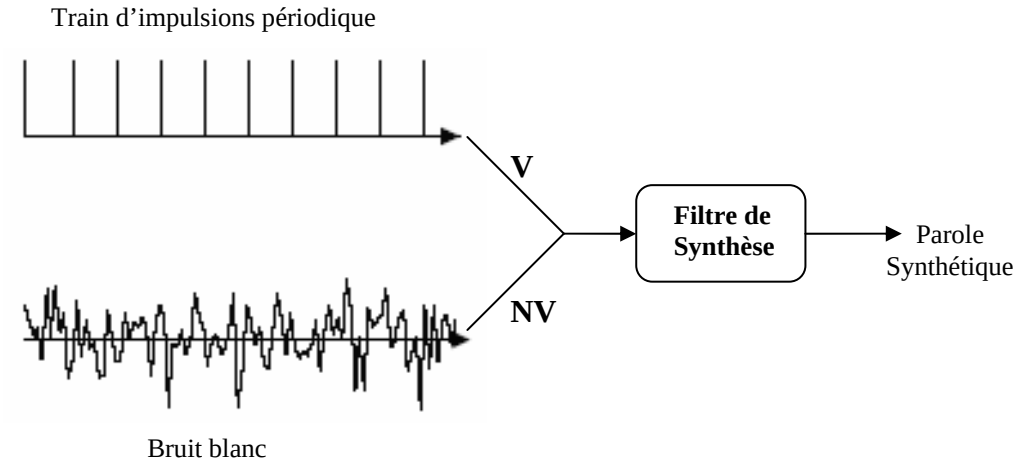


Fig. 1.5: Synthèse dans un vocodeur a 2 états d'excitation.

1.3.1.3. Nouveaux algorithmes de codage à bas débit

Dans les 20 dernières années, plusieurs algorithmes ont été proposés, qui permettent un codage à bas débit avec une qualité de type communications (MOS autour de 3.5). Ces nouveaux algorithmes ont en commun une meilleure représentation des parties voisées du signal et de l'évolution des paramètres de voisement, et aux transitions entre sons. La plupart du temps, les paramètres spectraux sont codés par quantification vectorielle sans distorsion audible pour un débit de 1200 bps [8,9]. Une pondération perceptuelle peut-être appliquée autour des formants, les paramètres LSF (Line Spectral Frequency) se prêtant bien à ce type de pondération.

Parmi les nouvelles méthodes de codage à bas débit, on peut distinguer les algorithmes de type harmoniques (MBE, STC), les algorithmes à interpolation de forme d'onde (WI, étudier en détail dans les sections suivantes), et les algorithmes à excitation mixte (MELP, HSX [10]). La complexité de ces nouvelles approches est nettement supérieure à celle des codeurs LPC classiques, mais il est possible de les implanter sur un seul circuit DSP.

1.3.2. Codeurs à très bas débit

Pour obtenir des débits inférieurs à quelques centaines de bits par seconde, il n'est plus possible de travailler sur des trames de longueur fixe. Une approche segmentale utilisant des segments de longueur variable est nécessaire. On peut considérer que les codeurs à très bas débit effectuent une reconnaissance de segments acoustiques dans la phase d'analyse et une synthèse de parole à partir d'une suite d'index de segments dans le décodeur. Le codeur réalise une transcription symbolique du signal de parole à partir d'un dictionnaire d'unités élémentaires de taille variable, qui peuvent être des unités linguistiques (comme des phonèmes, des transitions entre phonèmes, des syllabes,…), on parle alors de vocodeurs phonétiques, ou bien des unités acoustiques obtenues automatiquement de manière non supervisée sur un corpus d'apprentissage [6].

1.4. Qualité de la parole

Une considération importante dans tout codage de la parole est la qualité du signal reconstruit. Les recherches sur les différents types de codage essaient toujours de trouver un bon compromis entre la qualité du signal de parole restitué et le débit de transmission .Pour un débit fixé, le critère de qualité pourra alors être employé pour évaluer un système de codage.

Deux types de mesures, objective et subjective, peuvent permettre l’évaluation de la qualité de la parole.

1.4.1. Tests subjectifs

Lors d'un test subjectif, on demande à des participants de tester un système de télécommunications dans différentes conditions et de noter sur une échelle d'aptitude la qualité vocale de ce système, la notation s'effectue selon l'une des méthodes définies par l'Union Internationale des Télécommunications dans la Recommandation P.800 [11]. D'une manière générale, la qualité dépend de la personne qui la juge ; sa perception met en jeu l'expérience passée, les attentes et l'humeur de chacun.

La qualité vocale, dans le cadre des systèmes de communications, est elle aussi dépendante de celui qui l'évalue. Ainsi, les notes des participants pour une condition de test donnée, sont moyennées pour obtenir la note moyenne d'opinion MOS (Mean Opinion Score), qui permet de diminuer l'effet subjectif sur l'évaluation de la qualité vocale.

De plus, la perception de la qualité vocale dépend du contexte et de l'environnement dans lesquels est placée la personne qui juge. En effet, si elle est simplement en train d'écouter un message vocal (contexte d'écoute) ou si elle est impliquée dans une conversation avec un interlocuteur (contexte de conversation) ; les processus d'attention mis en jeu ne sont pas les mêmes, et le jugement de la qualité en est impacté. De même, l'environnement (bruit, informations visuelles ou sonores supplémentaires, etc.), influe sur le jugement de la qualité.

Niveau	Qualité de la parole	Niveau de distorsion
5	Excellente	Imperceptible
4	Bonne	a Peine perceptible mais pas gênante
3	Moyenne	Perceptible et un peu ennuyeux
2	Pauvre	Ennuyeux mais pas désagréable
1	Mauvaise	Très ennuyeux et désagréable

Tableau. 1.1: Description de l'échelle du MOS.

1.4.2. 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.

Parmi les mesures de distorsion objectives simples les plus couramment utilisées dans le domaine temporel mentionnant :

- Le rapport signal sur bruit (*SNR*) : si $s(n)$ est le signal de parole original, $\hat{s}(n)$ est le signal de parole reconstitué comportant N_τ échantillons, alors le *SNR* est définit comme suit [12]:

$$\boldsymbol{SNR(dB) = 10\log_{10}\frac{\sum_{n=0}^{N\tau-1} s^2(n)}{\sum_{n=0}^{N\tau-1}(s(n)-\hat{s}(n))^2}} \qquad \textbf{(1.1)}$$

D'après (1.1) le *SNR* ne peut prendre décision qu'après avoir écouter le fichier de parole entier.

- Le rapport signal sur bruit segmental (*SEGSNR*) : Le signal parole est découpé en N_F segments de N_S échantillons chacun, et on calcule une moyenne [12] ($s(n)$ est le signal de parole original et $\hat{s}(n)$ le signal synthétisé). Le *SEGSNR* est une meilleure mesure que le *SNR*, mais ce n'est pas toujours le cas quand la trame entière est presque silencieuse. Pour cela en fait appel à d'autres mesures objectives.

$$\boldsymbol{SEGSNR = \frac{1}{N_F}\sum_{i=0}^{N_F-1} 10\log_{10}\frac{\sum_{j=0}^{N_S-1} s^2(N_S*i+j)}{\sum_{j=0}^{N_S-1}(s(N_S*i+j)-\hat{s}(N_S*i+j))^2}} \qquad \textbf{(1.2)}$$

Ainsi, des méthodes objectives plus poussées que les mesures objectives simples telle que le rapport signal sur bruit, ont été développées.

En 2001, l'UIT-T a rendu public son système d'évaluation perceptuelle de la qualité vocale PESQ (Perceptuel Evaluation of Speech Quality) ou MOS estimé, normalisé sous la recommandation P862 [11] ; une méthode objective de prédiction de la qualité subjective pour la téléphonie avec bande passante réduite, et pour les codeurs vocaux.

Le processus clé de PESQ, comme pour les méthodes apparentées, est la transformation des signaux original et dégradé en représentations psychophysiques proches de celles des signaux auditifs du système auditif humain.

Le score PESQ est produit sur une échelle similaire au MOS, avec des valeurs situées entre -0,5 et 4,5. Les valeurs habituelles sont comprises entre 1,0 et 4,5.

La relation entre les scores PESQ et la qualité audio est la suivante [13] :

- Des scores PESQ entre 3 et 4,5 désignent une qualité perçue acceptable (avec 3,8 comme seuil de la qualité dans les systèmes téléphoniques traditionnels), niveau qu'on va appeler qualité « bonne ».
- Des valeurs entre 2 et 3 indiquent qu'un effort est nécessaire pour la compréhension de la parole, on va se référer à ceci comme qualité « basse ».
- Des scores inférieurs à 2 signifient que la dégradation a rendu la communication très difficile ou même impossible ; par conséquent la qualité est « inacceptable ».

1.5. Standardisation des codeurs de parole

Les algorithmes de codage de parole de base peuvent être divisés généralement en trois classes distinctes ; les codeurs de formes d'onde, les codeurs paramétriques, et les codeurs hybrides.

- **Les codeurs de formes d'onde :** très simples à mettre en oeuvre, utilisent des techniques de codage qui cherchent avant tout à préserver l'allure temporelle du signal de parole, ce qui les rend robustes aux différents types d'entrées. Le signal reconstruit avec ce type de codeur converge vers le signal original avec l'augmentation du débit de transmission. La qualité du signal synthétisé obtenue est excellente pour un débit relativement élevé.

Les premiers codeurs de formes d'onde sont de type PCM (Pulse Code Modulation), apparus dans les années 50, ces derniers reposent exclusivement sur le théorème d'échantillonnage de Shannon, et une quantification scalaire a pas fixe.

Etant donnée la distribution d'amplitudes des échantillons de parole, un quantificateur non uniforme apporterait une meilleure qualité pour le même débit. Ainsi, l'Union Internationale des Télécommunications a normalisé le codeur G.711 en 1972, un codeur logarithmique de parole de type PCM pour la transmission téléphonique avec un débit de 64 kbits/s.

Ce type de codage échantillonne le signal de parole à une fréquence de 8 kHz et opère une quantification sur 8 bits du signal de parole dans la bande de fréquences [300, 3400] Hz, avec une complexité plus élevée, le codage de parole peut être obtenu avec des débits inférieurs.

Ensuite, une technique, dont le principe est de quantifier non plus la valeur d'un échantillon à un instant donné, mais la différence avec une valeur prédite à partir d'échantillons précédents, a été introduite ; cette technique dite ADPCM (Adaptive Differential Pulse Code Modulation), émerge au début des années 70 et permet de réduire le débit de la moitié par rapport à la technique PCM sans détériorer la qualité de parole, l'ADPCM est utilisée aussi pour des services annexes (télécopie, télégraphie, transmission de données). Le codeur G.721, normalisé par l'UIT depuis 1984, est un exemple de système ADPCM qui fonctionne à 32 kbits/s. [14]

- **Les codeurs paramétriques :** basés sur des connaissances théoriques de production de la parole, ont permet des transmissions à moyen et bas débit (entre 5 et 16 kbits/s). La technique consiste à extraire du signal de parole les paramètres les plus pertinents permettant au décodeur de les synthétiser. Les performances des codeurs paramétriques, également connus sous le nom de vocodeurs, dépendent de la précision des modèles de production de parole. Ces codeurs ont été conçus pour des applications à bas débit et sont principalement prévus pour maintenir l'intelligibilité du signal vocal. La plupart des codeurs paramétriques sont basés sur le codage prédictif linéaire LP, ce dernier est détaillé dans le prochain chapitre.

La réduction de débit des codeurs de formes d'onde fait chuter rapidement la qualité d'écoute pour des débits inférieurs à 9.6 kbits/s. Une meilleure qualité pourra être observée pour des vocodeurs jusqu'à des débits de 4 kbits/s, mais ces applications restent réduites à cause d'une complexité accrue.

- **Les codeurs hybrides :** utilisent alors les deux méthodes, de formes d'onde et paramétriques. Tous les codeurs hybrides s'appuient eux aussi, sur une analyse LPC pour obtenir les modèles de synthèse de parole. Les deux techniques paramétrique et de formes d'onde modélisent respectivement le conduit vocal et le signal d'erreur résiduel.

Ce n'est qu'en 1985, qu'Atal et Schroeder définissent le codeur CELP, qui détermine une forme d'onde optimale du signal d'erreur en utilisant l'analyse par synthèse. Récemment, les codeurs CELP ont suscité beaucoup d'attention et servent de base à la plupart des algorithmes de codage de la parole. Un codeur LDCELP (Low-Delay CELP) ciblant un faible délai de codage/décodage, a été donné par la recommandation G.728 de l'UIT.

Ensuite, le codeur G.729, basé sur un codage de parole ACELP (Algebraic CELP) à 8 kbits/s, a été normalisé en 1996. Le tableau 1.2 récapitule les normes les plus utilisées pour les différents types de codage, et fournit leurs caractéristiques principales tel que : le débit, le délai de codage, la qualité estimée par une note subjective MOS et la complexité de calcul exprimée en million d'instructions par seconde MIPS.

Type Codage	**Type Codeur**	**Norme**	**Débit (kbit/s)**	**Qualité (MOS)**	**Délai De Codage (ms)**	**Complexité (MIPS)**
Formes d'onde	PCM	G.711	64	4.3	0.13	0.1
	ADPCM	G.726	16, 24,32	4.0	0.3	12.0
Paramétrique	LPC-10E	FS 1015	2.4	2.6	50	7.0
Hybrides	CELP	FS 1016	4.8	3.5	50	16.0
	LD-CELP	G.728	16	4.1	3	33.0
	CS-ACELP	G.729	8	4.0	30	20.0

Tableau. 1.2: Récapitulatif des contraintes des différentes techniques de codage. [13,14]

1.6. Conclusion

Les propriétés d'un signal de parole et les composantes essentielles du système auditif humain introduites dans ce premier chapitre, offrent des informations essentielles pour mieux comprendre le fonctionnement d'un codeur de parole ; ainsi nous avons illustré un brève historique des différents types de codeurs de paroles et des différentes techniques employées pour évaluer la qualité de ces derniers, alors on a constater que les techniques hybrides tel que le CELP, permet de diminuer le débit de quantification à des degrés considérablement réduits (8 Kbits/s) par rapport aux techniques classiques; pareillement, les nouvelles technologies des télécommunications exigent des débits encore plus limités ; donc de nouvelles approches de codages apparaissent. Nous étudierons dans les prochains chapitres l'une de ces techniques.

Chapitre 2 :

Prédiction Linéaire de la Parole

2.1. Introduction

La prédiction linéaire LP (Linear Prediction), exploite les redondances du signal de parole en le modélisant par un nombre restreint de paramètres sous la forme d'un filtre linéaire, est l'un des outils les plus importants de l'analyse, et du codage de la parole.

L'idée de base du codage LP est de considérer que tout échantillon de parole peut être exprimé comme une combinaison linéaire d'échantillons antérieurs. Un ensemble unique de coefficients prédicteurs peut alors être déterminé et utilisé pour supprimer les redondances à court terme du signal. Lors d'une analyse à court terme, la redondance proche entre les échantillons du signal de parole est supprimée par le filtre d'analyse LP, représentant le conduit vocal. Ce filtre permet d'extraire la structure des formants du signal d'entrée et d'obtenir un signal de sortie de faible énergie, correspondant à l'erreur de prédiction appelée signal résiduel, ou d'excitation si on ajoute une prédiction long terme.

Le filtre inverse d'analyse est le filtre de synthèse LP, dont la fonction de transfert décrit l'enveloppe spectrale du signal de parole. Chaque trame de parole est donc modélisée en sortie du système linéaire LP par un signal d'excitation. Un meilleur codage de celui-ci pourra être obtenu en utilisant un prédicteur à long terme qui prendra en compte la corrélation entre les échantillons éloignés du signal de parole ; alors que la période correspond à la valeur du pitch estimé. Notons que cette analyse n'aura aucun effet sur les sons non voisés de la parole dont le signal d'excitation est pratiquement un bruit blanc.

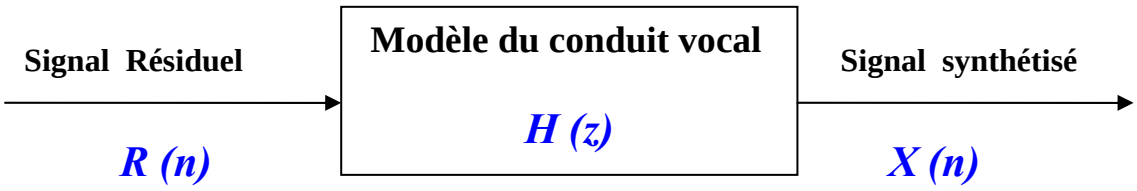


Fig. 2.1: Modélisation de la production de la parole.

2.2. Modèle LP

La relative simplicité de calcul, et la bonne estimation des paramètres de la technique de prédiction linéaire, fait une des méthodes les plus utilisées dans les processus de traitement de la parole, tels que les codeurs à bas débit, la synthèse ou la reconnaissance vocale.

Soit $H(z)$ la fonction de transfert du conduit vocal définit par :

$$H(z) = \frac{1}{1 - \sum_{k=1}^{N} a_k z^{-k}} = \frac{1}{A(z)} \qquad \textbf{(2.1)}$$

La modélisation complète d'un signal parole peut être décomposée en deux parties :

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

- Une partie analyse qui filtre le signal d'entrée avec la fonction de transfert $A(z)$. Ce filtre tout zéro est défini comme le filtre d'analyse LP, permet d'extraire l'information prédictible du signal, et de définir un signal d'erreur résiduel $r(n)$ entre le signal de parole d'entrée $x(n)$ et son estimation :

$$\begin{aligned} r(n) &= x(n) - \sum_{k=1}^{N} a_k . x(n-k) \\ x(n) &= \sum_{k=1}^{N} a_k . x(n-k) + r(n) \end{aligned} \qquad \textbf{(2.2)}$$

Le signal résiduel est l'excitation idéale du modèle LP du conduit vocal $H(z)$.
Une modélisation précise de ce signal permet d'obtenir un signal reconstruit naturel. Or l'estimation des paramètres LP du modèle vocal entraîne une approximation du signal d'excitation.

A mesure que l'ordre du modèle LP augmente, un meilleur ajustement au spectre de la voix est obtenu. Cependant, plus l'ordre N sera élevé, plus le nombre de paramètres à transmettre augmentera. L'ordre doit ainsi déterminer à partir d'un compromis entre une bonne représentation de la structure formantique du signal de parole, la complexité de calcul et le débit de transmission. En général, deux pôles sont nécessaires pour représenter chaque formant, et jusqu'à quatre autres sont employés pour approximer les vallées du spectre de parole.

2.3. Estimation des paramètres LP

Il y a deux approches pour l'estimation des cœfficients a court terme $\{a_k\}$ qui sont la méthode d'autocorrelation et la méthode de covariance. Les deux méthodes utilisent le principe classique des moindres carrés et donnent l'ensemble $\{a_k\}$ qui minimisent l'énergie du signal résiduel résultant.

2.3.1. Méthode d'autocorrection

Dans cette méthode, le signal $x(n)$ est multiplié par une fenêtre $w(n)$. On obtient ainsi le signal fenêtré $x_w(n)$:

$$\boldsymbol{x_w(n) = w(n).x(n)} \qquad \textbf{(2.3)}$$

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

$$\boldsymbol{E = \sum_{n=-\infty}^{\infty} r^2(n) = \sum_{n=-\infty}^{\infty} \left[x_w(n) - \sum_{k=1}^{N} a_k . x_w(n-k) \right]^2} \qquad \textbf{(2.4)}$$

La recherche des coefficients $\{a_k\}$ se fait, en minimisant E relativement aux coefficients $a_1 \ldots a_k$. En dérivant E par rapport aux coefficients $\{a_k\}$:

$$\frac{\partial E}{\partial a_k} = 0 \qquad \text{Pour } k=1,2,\ldots, N \qquad (2.5)$$

On obtient donc les N équations suivantes :

$$\sum_{n=-\infty}^{\infty} x_w(n).x_w(n-i) = \sum_{K=1}^{N} a_k \sum_{n=-\infty}^{\infty} x_w(n-i).x_w(n-k) \qquad \text{Pour } i=1,2,\ldots, N \qquad (2.6)$$

Dans les équations (2.6), on considère que les données sont nulles à l'extérieur de la fenêtre d'analyse $w(n)$[1].

En définissant la fonction d'autocorrelation du signal fenêtré $x_w(n)$ par:

$$R(i) = \sum_{n=-\infty}^{\infty} x(n).x(n-i) = \sum_{n=i}^{L_w-1} x_w(n)x_w(n-i) \qquad (2.7)$$

Où L_w représente la longueur de la fenêtre d'analyse, et en substituant les équations (2.7) aux équations (2.6), on obtient l'équations matricielle suivante :

$$\begin{bmatrix} R(0) & R(1) & \cdots & R(N-1) \\ R(1) & R(0) & \cdots & R(N-2) \\ \cdots & \cdots & \cdots & \cdots \\ R(N-1) & R(N-2) & \cdots & R(0) \end{bmatrix} \times \begin{bmatrix} a_1 \\ a_2 \\ \cdots \\ a_N \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ \cdots \\ R(N) \end{bmatrix} \qquad (2.8)$$

Ce système matriciel se résout en tenant compte du fait que la matrice d'autocorrélation est une matrice de Toeplitz. Cette propriété permet de résoudre efficacement le système, c'est-à-dire sans inversion de la matrice $R(i)$, par l'algorithme de Levinson-Durbin décrit dans [15]. Cette propriété assure également que le filtre $A(z)$ est à phase minimale.

[1] Généralement on utilise une fenêtre de pondération pour éliminer l'effet de Gibbs.

Dans le filtre de Synthèse $H(z) = 1/A(z)$, les zéros de $A(z)$ deviennent les pôles de $H(z)$ et le fait que $A(z)$ soit à phase minimale, garantit la stabilité du filtre de synthèse $H(z)$.

2.3.2. Méthode de la covariance

Dans la méthode de la covariance, on fenêtre le signal d'erreur, au contraire de la méthode de l'autocorrelation dans laquelle on fenêtre le signal $x(n)$.

L'énergie E du signal s'écrit alors :

$$E = \sum_{n=-\infty}^{\infty} r^2(n) w(n) \qquad \textbf{(2.9)}$$

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

$$\sum_{k=1}^{N} \phi(i,k).a_k = \phi(i,0) \qquad ,1 \le i \le N \qquad \textbf{(2.10)}$$

Où $\phi(i,k)$ est la fonction de covariance du signal $x(n)$ définie par :

$$\phi(i,k) = \sum_{n=-\infty}^{\infty} w(n).x(n-i).x(n-k) \qquad \textbf{(2.11)}$$

Les équations (2.10) peuvent s'écrire sous la forme matricielle suivante :

$$\begin{bmatrix} \phi(1,1) & \phi(1,2) & \cdots & \phi(1,N) \\ \phi(2,1) & \phi(2,2) & \cdots & \phi(2,N) \\ \cdots & \cdots & \cdots & \cdots \\ \phi(N,1) & \phi(N,2) & \cdots & \phi(N,N) \end{bmatrix} \times \begin{bmatrix} a_1 \\ a_2 \\ \cdots \\ a_N \end{bmatrix} = \begin{bmatrix} \phi(1,0) \\ \phi(2,0) \\ \cdots \\ \phi(N,0) \end{bmatrix} \qquad \textbf{(2.12)}$$

Ou $\phi(i) = \phi(i,0)$ $\boldsymbol{i = 1,2,\ldots,N}$

Cette matrice est symétrique mais les coefficients sur les diagonales ne sont pas égaux entre eux, à la différence de la matrice d'autocorrélation définie ci-dessus. La méthode de décomposition de Cholesky permet ainsi la résolution de ce système.

La méthode de covariance n'applique pas le fenêtrage au signal de parole d'entrée, ce qui la rend très avantageuse pour les applications d'estimation spectrale à haute résolution. Cependant elle ne garantie pas la stabilité de filtre tout-pole de synthèse LP ; les pôles estimés peuvent se retrouver à l'extérieur du cercle unité. La méthode de covariance impose donc l'utilisation d'un algorithme de stabilisation pour ramener les pôles à l'intérieur de cercle unité, ce qui augmente la complexité du codeur. Pour cette raison il est préférable d'utiliser la méthode d'autocorrelation dans l'implémentation d'un codeur WI [2].

2.4. Considérations pratiques

La meilleur performance de l'analyse LP se base essentiellement sur :

- Le choix de la fréquence d'échantillonnage en 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 $1/A(z)$. 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 respectivement *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*** :

$$w[k] = 0.54 - 0.46\cos\left(2\pi\frac{k-1}{n}\right) \qquad \textbf{(2.13)}$$

avec $k = 1,\ldots, n$

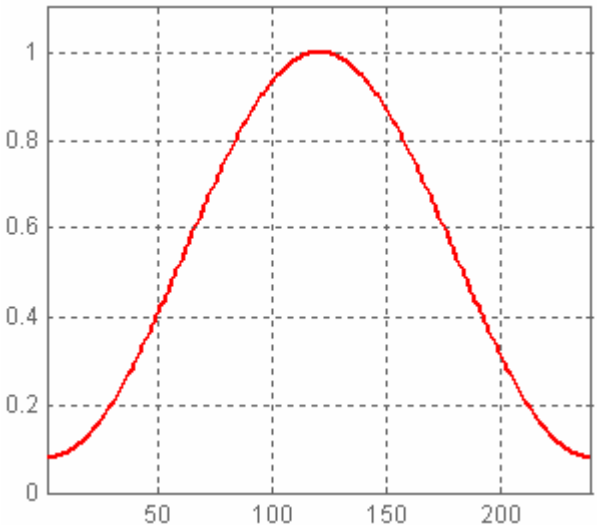


Fig. 2.2: Exemple d'une fenêtre de Hamming de 240 points, extraite à partir de Matlab.

On retiendra donc que l'analyse LP d'un signal de parole implique la résolution d'un système de (l'ordre de 10 par exemple) 10 d'équations à 10 inconnues toutes les 10 à 30ms.

2.5. 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.

Parmi les représentations qui se sont avérés efficaces [16], 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 suivante :

$$A(z)=\frac{P(z)+Q(z)}{2} \qquad \textbf{(2.14)}$$

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 [8,9].

La méthode utilisée pour l'extraction des LSP est celle de Kabal et Ramachadran, 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.

$$P_{N+1}(z) = A(z) + z^{-(N+1)} A(z^{-1}) \qquad \textbf{(2.15)}$$

$$Q_{N+1}(z) = A(z) - z^{-(N+1)} A(z^{-1}) \qquad \textbf{(2.16)}$$

On montre que si $A(z)$ est à phase minimale[2] , 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 (2.15) et (2.16)

$$P_{N+1}(z) = \left(1+z^{-1}\right) \prod_{i=1}^{N/2} \left(1 - 2\cos(w_{2i-1}) z^{-1} + z^{-2}\right) \qquad \textbf{(2.17)}$$

$$Q_{N+1}(z) = \left(1-z^{-1}\right) \prod_{i=1}^{N/2} \left(1 - 2\cos(w_{2i}) z^{-1} + z^{-2}\right) \qquad \textbf{(2.18)}$$

Avec w_i étant la fréquence de la raie spectrale. Elle a comme propriété principale d'être croissante. En effet, $0<w_1<\ldots\ldots<w_{2i}<\ldots\ldots<w_N<\pi$. [15]

[2] Autrement dit que ses racines sont toutes à l'intérieur du cercle unité

La figure 2.3 illustre un exemple de localisation des racines des deux polynômes pour un ordre pair. Notons que les racines en 0 et π ne sont pas représentées.

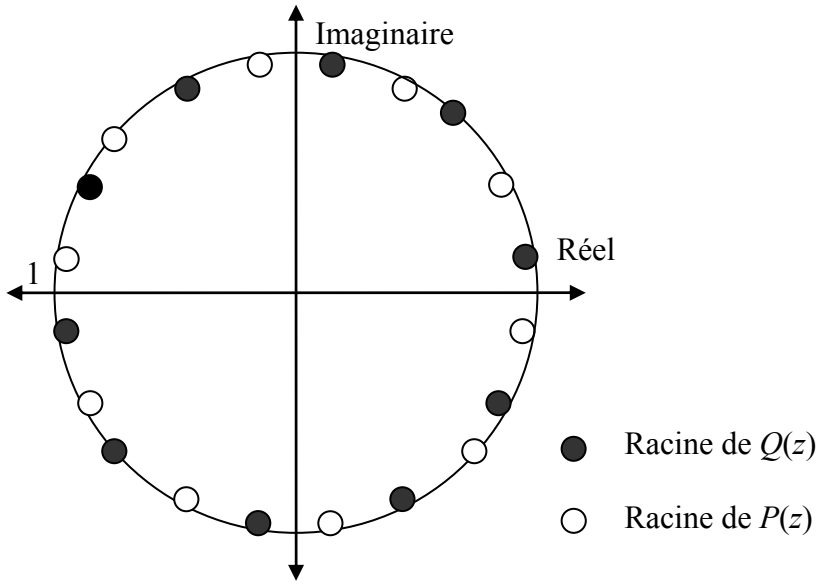


Fig. 2.3: Localisation possible des racines pour $P(z)$ et $Q(z)$ d'ordre pair [5].

2.5.1. 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. [12, 17]

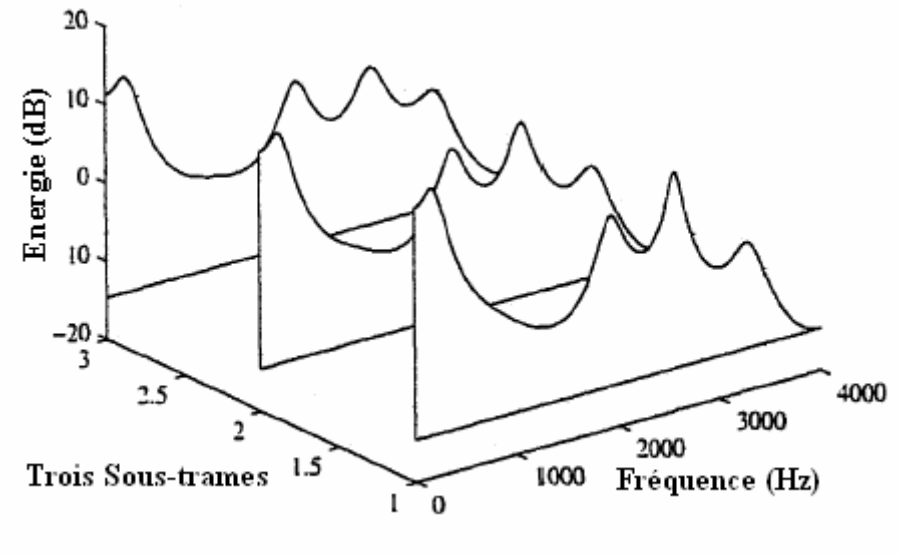


Fig. 2.4: Représentation spectrale de l'interpolation des cœfficients LP.
La deuxième sous trame est le résultat de l'interpolation entre la première et troisièmes sous trame. [12]

2.6. Principe de la prédiction à long terme

La mise en oeuvre 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 $P_A(z)$ d'ordre p_i. La fonction de transfert de ce filtre est défini par :

$$P_A(z) = 1 - \beta . z^{-p_i} \qquad \textbf{(2.19)}$$

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.

2.7. Expansion de la largeur de bande

Parfois, dans l'analyse LP, la représentation spectrale du filtre de synthèse possède des formant a forme aigue. Cela implique que les pôles du filtre sont assez près du cercle unité et d'où, le filtre est marginalement stable ; de telle stabilité marginale peut multiplier les chances d'avoir un croisement dans la quantification des cœfficients LSF. Cependant il existe deux méthodes permettant d'améliorer cette estimation.

La première consiste à appliquer une fenêtre de forme gaussienne, sur le signal d'autocorrélation .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 [12,16].

L'expansion de la largeur de bande peut être calculée comme suit [12]:

$$\Delta B = -\frac{1}{\pi T} Ln(\gamma) \qquad \textbf{(2.20)}$$

Tel que T représente la période d'échantillonnage.

2.8. La préaccentuation

La procédure de la conversion A/D sur un signal de 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'autocorrelation male 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 $T(z)$ tel que :

$$T(z) = 1 - \alpha . z^{-1} \qquad \textbf{(2.21)}$$

Où α détermine la fréquence de coupure de filtre tout zéros d'ordre 1 (α toujours positive inférieur a 1).

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 ; pour éliminer l'effet de la préaccentuation un filtre de désaccentuation est utilisé au décodeur.

$$G(z) = 1 / T(z) \qquad \textbf{(2.22)}$$

2.9. Conclusion

La prédiction linéaire est l'un des outils les plus largement répandu et utilisée en codage de parole. Cette technique a fait et fait encore aujourd'hui l'objet de nombreuses études. Les techniques d'estimation des filtres, de quantification des paramètres sont fiables, précises et robustes. Grâce à la technique de conversion des coefficients de filtre en coefficients LSP, ces derniers offrent de meilleures propriétés de codage et une quantification vectorielle de ces coefficients conduit à un taux de compression élevé [8,9].

Dans ce deuxième chapitre, on a décrit les principales méthodes de calcules des cœfficients LP ; ainsi pour avoir un bonne estimation des cœfficients du filtre et du signal résiduel, l'opération LP est toujours précédée par une opération de préaccentuation et d'une expansion de la largeur de bande qui s'exécute après l'estimation de ces derniers.

Chapitre 3 :

Interpolation de la Forme d'Onde

3.1. Introduction

La qualité des codeurs à forme d'onde, tels que les codeurs utilisant la technique CELP, dégrade rapidement aux taux au-dessous de 4.8 kbps. En effet le codage CELP se base essentiellement sur les mesures objectives tel que le SNR. Cependant, le SNR n'est pas une mesure idéale de la qualité perceptuelle du signal de la parole reconstruite. Donc, pour avoir une haute qualité perceptuelle du codage de la parole à 4 kbps, il est nécessaire de développer des algorithmes de codage qui maintiennent la périodicité du signal synthétisé. Plus loin, ces nouveaux algorithmes doivent exploiter la nature évolutive de la parole.

L'interpolation de la forme d'onde est parmi les méthodes qui permettent de réaliser des codeurs à un taux plus bas que 4.8 kbps en maintenant, et en même temps en améliorant la qualité perceptuelle de la parole. Cette technique a été introduite et développée par W. B. Kleijn [1], qui est considérée comme étant la première version et a été baptisée par PWI (Prototype Waveform Interpolation). Il existe plusieurs améliorations depuis lors [18, 19, 20, 21]. 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 [3].

Bien que la PWI travaille remarquablement 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 améliorée pour devenir la WI (Waveform Interpolation) qui est capable de prendre en charge les sons voisés et non voisés, la WI représente un signal de parole avec une séquence de formes d'ondes. Ces formes d'ondes sont simplement de longueurs égales à la période du pitch.

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 qui sera abrégé par CW (Characteristic Waveform) par la suite.

Dans la WI les formes d'ondes sont prélevées à une fréquence plus grande. Cependant, une augmentation de la fréquence de prélèvement des formes d'ondes, entraînera une augmentation du débit. Pour contourner ce problème, la WI décompose la CW en une forme d'onde à évolution lente appelée SEW (Slowly Evolving Waveform) qui représente la section voisée, et une forme d'onde à évolution rapide ou REW (Rapidly Evolving Waveform) qui représente la section non voisée. 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 3.1 fournit une vue d'ensemble du plan WI comme a été présentée par E. Choy [2].

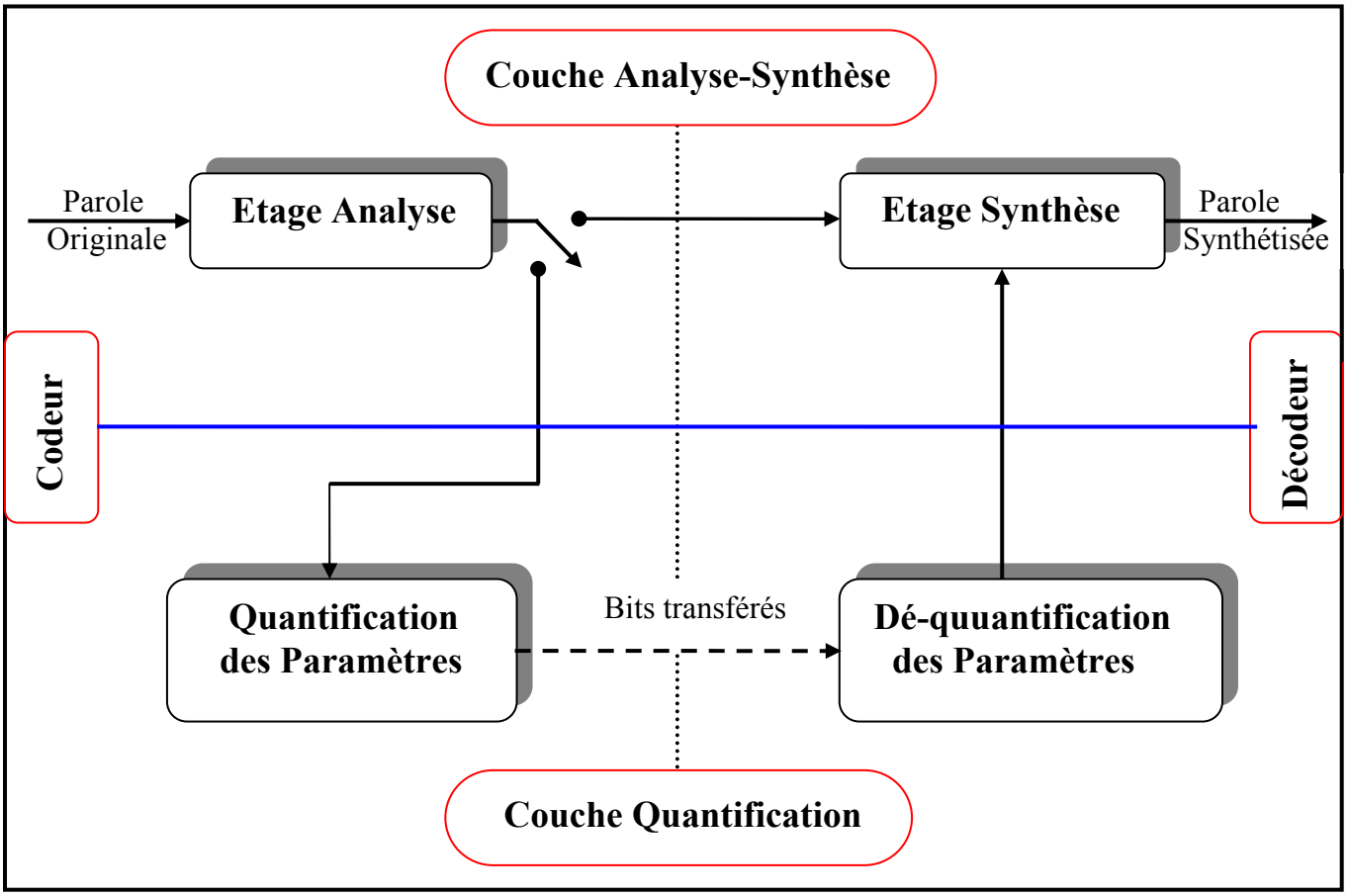


Fig. 3.1: Vue d'ensemble du codage WI.

3.2. Le codeur WI

Comme il est déjà mentionné, le but fondamental de la couche d'analyse est de décomposer le signal de parole en une série d'ondes qui sera alors convertit en surface bidimensionnelle, ainsi que d'extraire d'autres paramètres orthogonaux tels que les cœfficients LSF, l'énergie et le pitch. La figure 3.2 montre tous les processeurs que comprend la couche d'analyse.

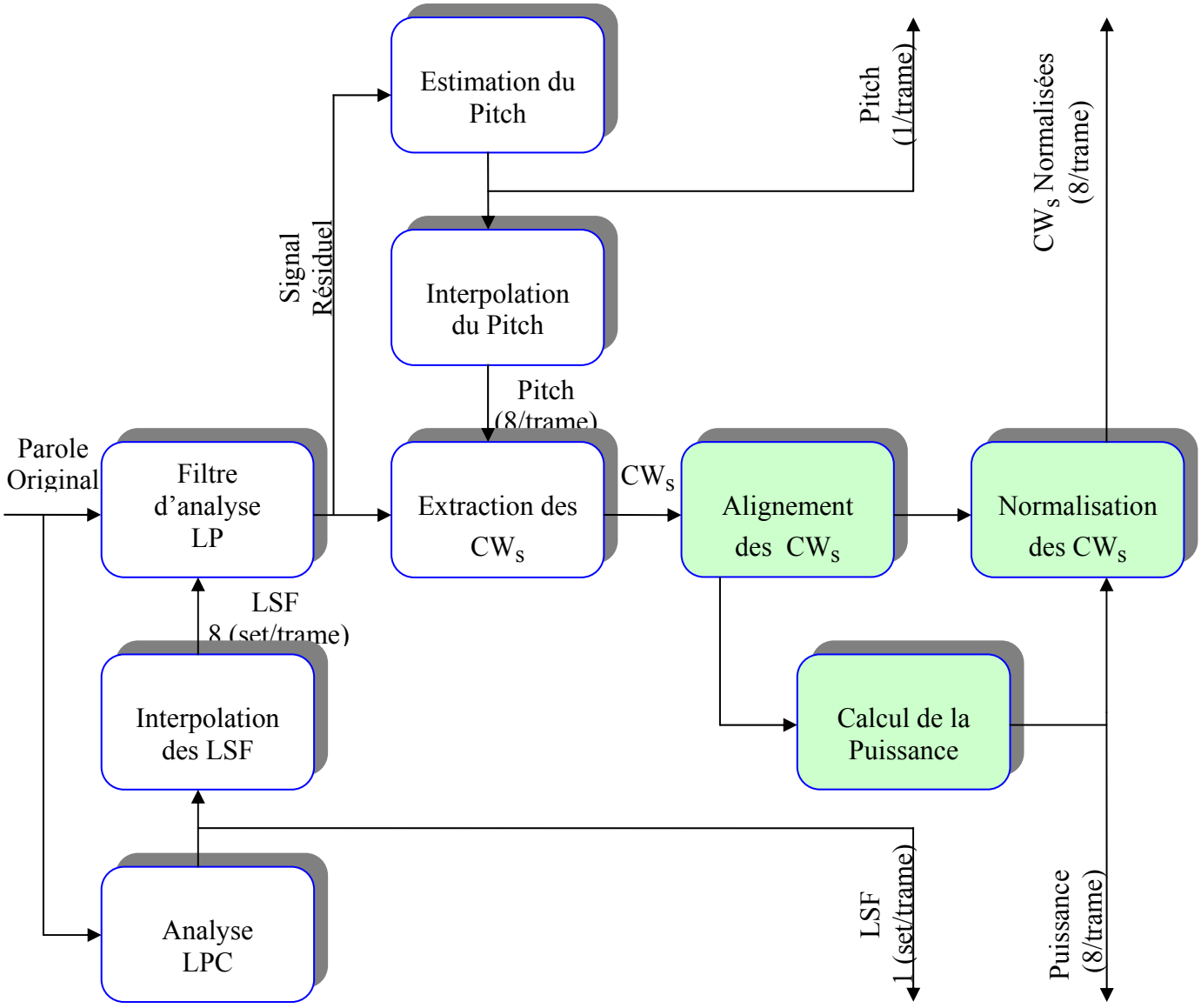


Fig. 3.2: Schéma bloc de l'étage d'analyse de la WI. Les processeurs colorés travaillent à la fréquence des sous-trames tandis que les autres travaillent à celle des trames [2].

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[1].

Comme la WI est basée sur le modèle de codage prédictif qui utilise une analyse par synthèse, le signal de parole est converti en signal résiduel en utilisant un filtre d'analyse d'ordre 10 [18] qui comporte les cœfficients de prédiction linéaire LP pour chaque trame d'analyse.

Avant cela, l'opération d'analyse LP est précédée par une préaccentuation avec un facteur $\alpha = 0.6$, et une pondération par une fenêtre de hamming de longueur $L_w = 240$. Le centre de la fenêtre coïncide avec l'extrémité droite de la trame courante [12]. En d'autres termes, la fenêtre couvre 120 échantillons de la trame courante et 120 de la trame future. 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.

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+1 [2, 22] et forment un ensemble intermédiaire pour chacune des 8 sous-trames de la trame à coder (Figure 3.3).

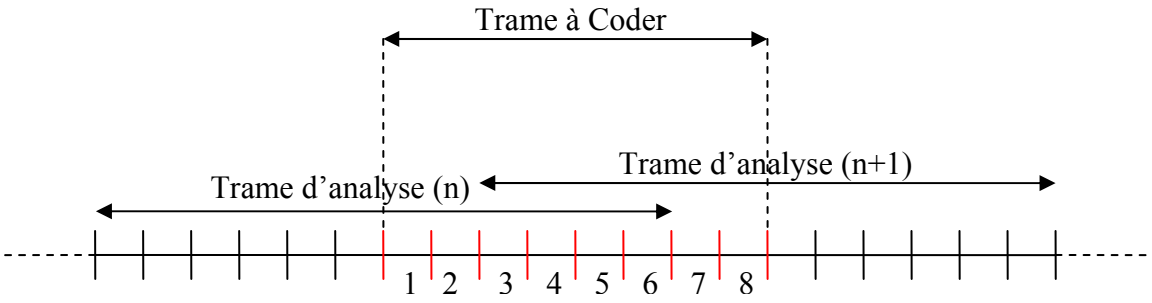


Fig. 3.3: Interpolation des cœfficients LSF.

[1] Il existe d'autres travaux où on utilise 16 échantillons comme longueur de sous trame

3.2.1. Détection 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 celle d'interpolation au décodeur reposent lourdement sur la valeur estimée du pitch.

II existe plusieurs procédures d'estimation du pitch. Quelques unes sont basées sur la localisation des marqueurs de pitch tandis que d'autres sont basées sur la recherche de la position du maximum d'autocorrélation ou du gain de prédiction pour une trame d'échantillons. Dans cette simulation de la WI, on a adopté un algorithme tiré du EVRC [22] qui se base sur le calcul de l'autocorrection normalisée du signal résiduel. Ce dernier est suivi de certaines modifications tirées de [23, 24, 25] qui peuvent influencer sur la précision et le temps de calcul du pitch. On donne une description brève de cet algorithme dans ce qui suit.

L'estimation du pitch est effectuée une fois par trame, et 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, ce dernier 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 [26].

La valeur du pitch estimée à premier niveau (d_{max}) et les limites inférieurs et supérieurs du pitch (P_{min} , P_{max}) sont définis par :

$$\boldsymbol{d}_{\max} = \arg\max\Big[\sum_{\boldsymbol{k}=\boldsymbol{0}}^{\boldsymbol{40-d-1}} \boldsymbol{r_{dec}(d) r_{dec}(k+d)}; \ldots\ldots\ldots\ldots \boldsymbol{5 \le d \le 30} \Big] \qquad \textbf{(3.1)}$$

$$\boldsymbol{P}\min = \boldsymbol{Max}(\boldsymbol{20}, \boldsymbol{\lambda} * \boldsymbol{d}_{\max} - \boldsymbol{3}) \text{ et } \boldsymbol{P}\max = \boldsymbol{Min}(\boldsymbol{120}, \boldsymbol{4} * \boldsymbol{d}_{\max} + \boldsymbol{3}) \qquad \textbf{(3.2)}$$

Sachant que $r_{dec}(.)$ représente le signal sous échantillonné du signal résiduel $r(.)$; selon les essais qu'on a effectués sur notre simulation, λ peut prendre les valeurs 1 ou 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 :

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

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

$$dd_{\max} = \arg\max\left[R(d)\right] \qquad P\min \le d \le P\max \qquad \textbf{(3.5)}$$

$$R_s(d) = sqrt\left(\left(\sum_{k=0}^{160-dd_{\max}-1}(r(k))^2\right) \times \left(\sum_{k=dd_{\max}}^{160-1}(r(k))^2\right)\right) \qquad \textbf{(3.6)}$$

Où $R(d)$ et $Rs(d)$ représentent respectivement la fonction d'autocorrection et l'énergie du signal résiduel, d'après [24, 25] on peut mettre $Rs(d)$ sous la forme :

$$R_s(d) = sqrt\left(\sum_{k=dd_{\max}}^{160-1}(r(k))^2\right) \qquad \textbf{(3.7)}$$

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

Remarques

- Il existe d'autres méthodes [3] qui font l'opération inverse (sur-échantillonnage), avant le calcul de l'autocorrection normalisée, or l'augmentation de la résolution temporelle diminue considérablement les cas de doublement ou triplement du pitch qui est un phénomène indésirable pour le bon fonctionnement de la WI. Cette procédure présente deux inconvénients qui consistent à exécuter plus d'opérations et à occuper un plus grand espace mémoire.

- L'équation (3.3) utilise deux paramètres P_{min} *et* P_{max} qui sont les valeurs minimale et maximale de la période du pitch, compris entre 20 et 120. On pouvais étendre cet intervalle [25] de 20 à 147 échantillons, puisque de toute manière, on alloue 7 bits pour quantifier le pitch (147 - 20 + 1 = 128 = 2^7). Cependant, un intervalle plus large de valeurs du pitch peut mener à plusieurs apparitions de doublement ou triplement du pitch [2]. Par conséquent ; l'opération de quantification des CW sera plus compliquée.

Maintenant, après avoir trouver 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 l'algorithme suivant [22]:

si ($\beta_0 > \beta_1 + 0.4$) {
  si ($|\ d_0 - d_1\ | > 15$) {
    $\boldsymbol{d_{opt}} = d_0$
  }
  sinon {
    $\boldsymbol{d_{opt}} = (d_0 + d_1)/2$
  }
sinon si ($\beta_0 > \beta_1 + 0.4$ **et** $|\ d_0 - dernier\ pitch\ | < 7$) {
  si ($|\ d_0 - d_1\ | > 15$) {
    $\boldsymbol{d_{opt}} = d_0$
  }
  sinon {
    $\boldsymbol{d_{opt}} = \text{argmax}\left[\sum_{k=40}^{160-d} \boldsymbol{r}(k)\boldsymbol{r}(k+\boldsymbol{d})\right]$ Avec $((d_0 + d_1)/2)-1 \le d \le ((d_0 + d_1)/2) + 1$
  }
sinon {
  $\boldsymbol{d_{opt}} = d_1$
}

3.2.2. 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[2] 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 :

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

Comme le risque d'apparition d'un doublement où 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 $P(n_1)$:

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

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

[2] On a huit point d'extraction par trame

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

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

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

La figure 3.4 illustre un exemple d'une telle interpolation dans le cas d'un doublement du pitch et dans celui d'une diminution de moitié.

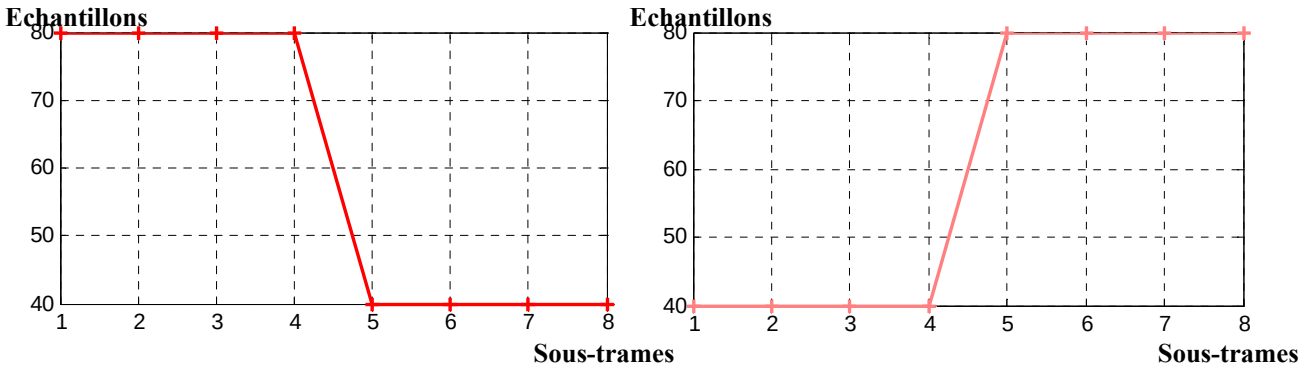


Fig. 3.4: Interpolation du pitch dans le cas d'un doublement de sa valeur (a droite), et dans le cas de diminution de la moitie (a gauche), entre 40 et 80.

3.2.3. Extraction des CW

L'opération d'extraction est effectuée une fois par sous-trame on aura donc huit extractions par trame avec un débit d'extraction égal à R_{extr}. Dans le processus d'extraction, on commence par diviser la trame courante en huit[3] intervalles de même longueur. Le point situé sur l'extrémité droite de chaque intervalle sera un point d'extraction comme illustré dans la figure 3.5a. Donc, deux points d'extraction adjacents seront séparés de 20 échantillons. Cet intervalle définit la longueur L_{sf} de notre sous-trame.

[3] On peut aussi deviser en dix intervalle, si on veut avoir R_{extr}=10 c-à-d L_{sf} =16 échantillons.

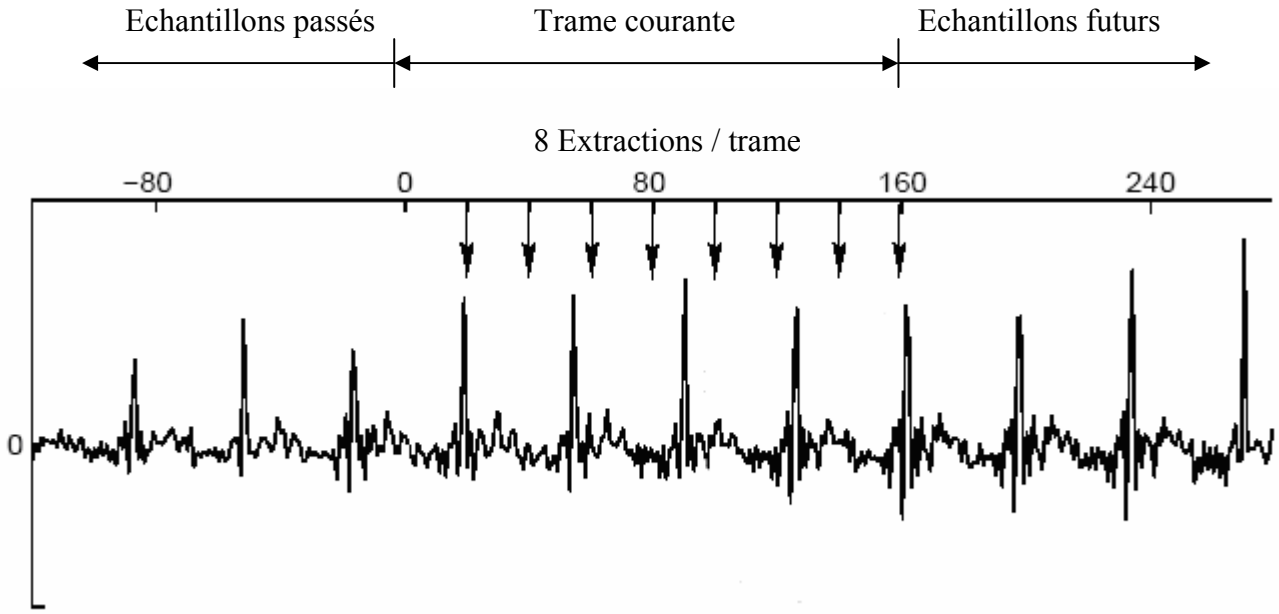
A chaque point d'extraction, on prend le pitch interpolé à ce point et on forme une fenêtre d'extraction de cette longueur. La fenêtre d'extraction est centrée au point d'extraction et le signal résiduel contenu dans cette fenêtre formera donc notre CW extraite. Par conséquent, la CW extraite a toujours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

Les CW sont étendues périodiquement pendant la conversion au domaine DTFS (voir Annexe A). Alors, 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 [18], tel que ε peut varié entre -16 et +16 [2]. La position qui donnera la plus petite énergie du signal autour des deux extrémités de la fenêtre d'extraction est choisie. La figure 3.5 montre un exemple de l'opération d'extraction.

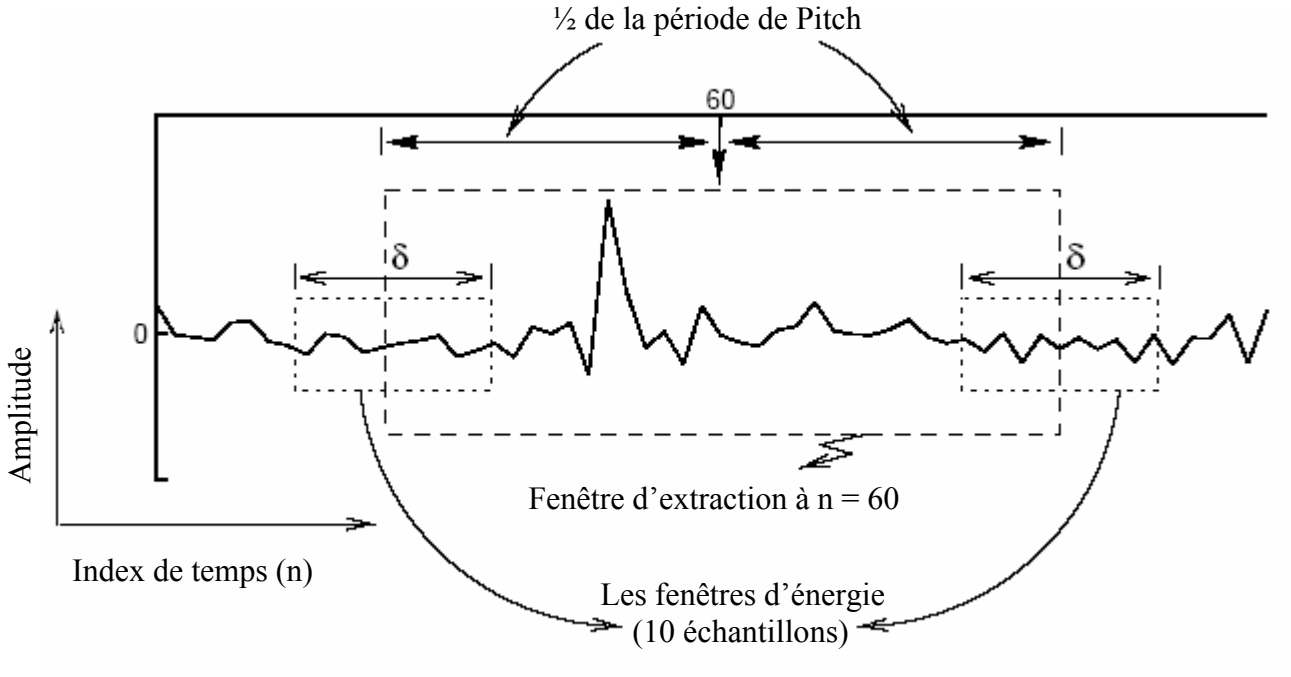
3.2.4. Alignement des CW

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 chaque 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 [3], ce décalage temporel circulaire est en réalité, équivalent à l'adition d'une phase linéaire aux coefficients DTFS.

La figure 3.6 montre un schéma bloc de la procédure d'alignement.



(a). Le segment d'un signal résiduel par analyse LP



(b). Illustration de l'opération d'extraction

Fig. 3.5: Exemple de l'opération d'extraction, (a) les positions originales des points d'extraction des 8 CW. (b) illustration détaillée pour le point d'extraction a n = 60. [27]

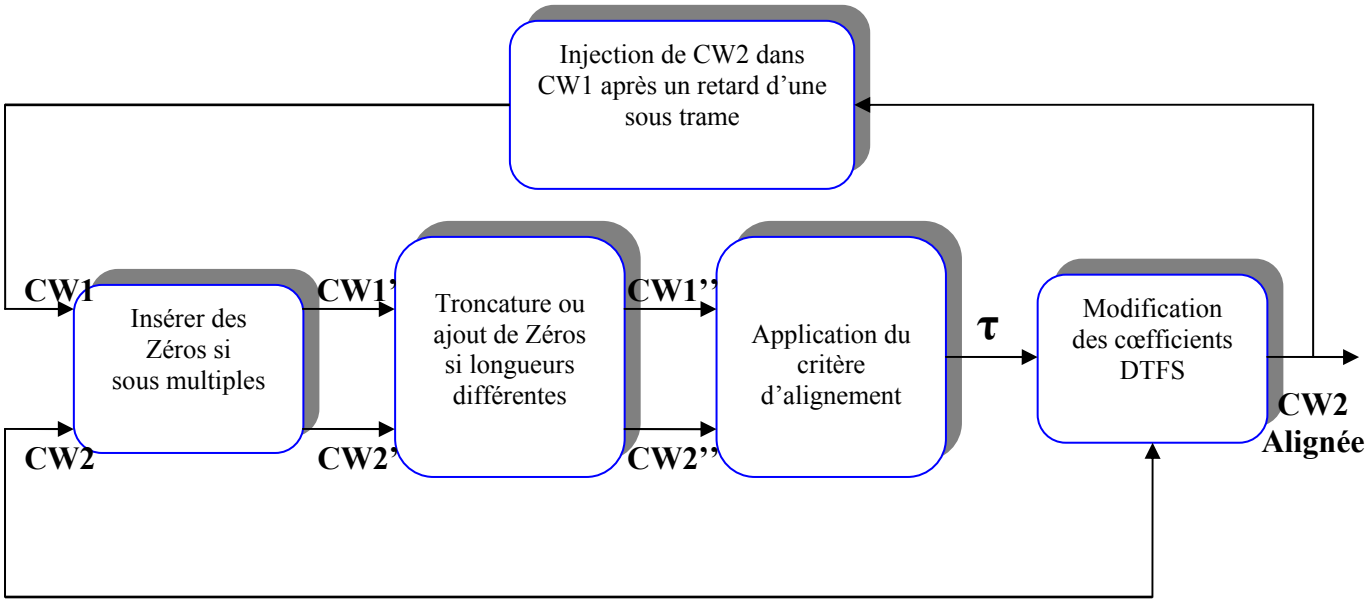


Fig. 3.6: Schéma bloc de la procédure d'alignement.

Puisque les CW n'ont pas toujours la même longueur, ce qui nous mène à étudier chaque un des cas possibles [2] :

Premier Cas : Les deux CW ont la même longueur

Les représentations en DTFS[4] des deux CW successives sont :

$$s(n_0,m)=\sum_{k=0}^{M}\left[A_k(n_0)\cos\left(\frac{2\pi km}{P}\right)+B(n_0)\sin\left(\frac{2\pi km}{P}\right)\right]$$
$$s(n_1,m)=\sum_{k=0}^{M}\left[A_k(n_1)\cos\left(\frac{2\pi km}{P}\right)+B(n_1)\sin\left(\frac{2\pi km}{P}\right)\right] \qquad (3.11)$$

Où n_0 et n_1 sont les positions dans le temps, respectivement des CW précédente et présente et $n_1 - n_0 = L_{sf}$. En plus, pour une meilleure commodité de notation ;

$$P=P(n)=P(n-1)$$
$$M=\lfloor P(n)/2\rfloor=\lfloor P(n-1)/2\rfloor \qquad (3.12)$$

[4] Pour donner une meilleure précision ; dans tous les calculs qui se suivent, la composante continue est prise en considération.

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

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

$$s(n_1,m-T)=\sum_{k=0}^{M}\left[A_k(n_1)\cos\left(\frac{2\pi k(m-T)}{P}\right)+B_k(n_1)\sin\left(\frac{2\pi k(m-T)}{P}\right)\right] \qquad \textbf{(3.13)}$$

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 définit par :

$$T=\underset{0\le T'\le P}{\arg\max}\sum_{k=0}^{M}\left\{\begin{array}{l}\left[A_k(n_0)A_k(n_1)+B_k(n_0)B_k(n_1)\right]\cos\left(\frac{2\pi kT'}{P}\right)+\\ \left[B_k(n_0)A_k(n_1)-B_k(n_1)A_k(n_0)\right]\sin\left(\frac{2\pi kT'}{P}\right)\end{array}\right\} \qquad \textbf{(3.14)}$$

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

$$\tau=\underset{0\le \tau'\le P}{\arg\max}\sum_{k=0}^{M}\left\{\begin{array}{l}\left[A_k(n_0)A_k(n_1)+B_k(n_0)B_k(n_1)\right]\cos(k\tau')+\\ \left[B_k(n_0)A_k(n_1)-B_k(n_1)A_k(n_0)\right]\sin(k\tau')\end{array}\right\} \qquad \textbf{(3.15)}$$

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-échantillonnage et sous-échantillonnage conventionnels. Cet alignement fractionnel se fait à n'importe quelle résolution désirée. Cependant une résolution de ¼ d'un échantillon [2] 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 (3.13) en utilisant les identités trigonométriques fondamentales. On obtient un nouveau ensemble de cœfficients DTFS de la CW décalée.

$$\left.\begin{aligned} A'_k(n_1) &= A_k(n_1)\cos(k\tau) - B_k(n_1)\sin(k\tau) \\ B'_k(n_1) &= A_k(n_1)\sin(k\tau) + B_k(n_1)\cos(k\tau) \end{aligned}\right\} \quad \text{Pour } k = 1, 2, \ldots, M \qquad (3.16)$$

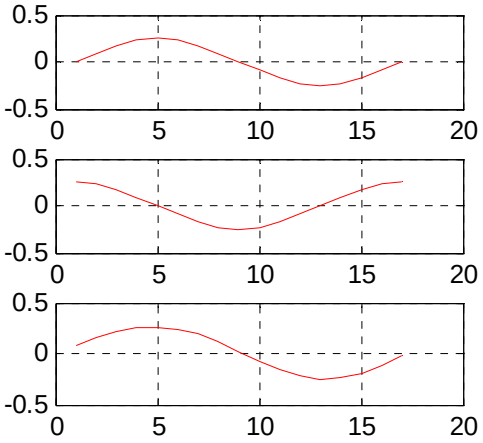


Fig. 3.7: Exemple d'alignement de la fonction sin(x)[5] avec la fonction cos(x).

Deuxième Cas : Les deux CW ont des longueurs différentes

Dans ce cas le critère d'alignement (3.15), qui est basé sur la supposition d'égalité de dimension, n'est plus applicable directement. Donc on précède l'application de ce critère 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.

[5] On a choisi les fonctions usuelles comme exemple, pour mieux comprendre la procédure d'alignement

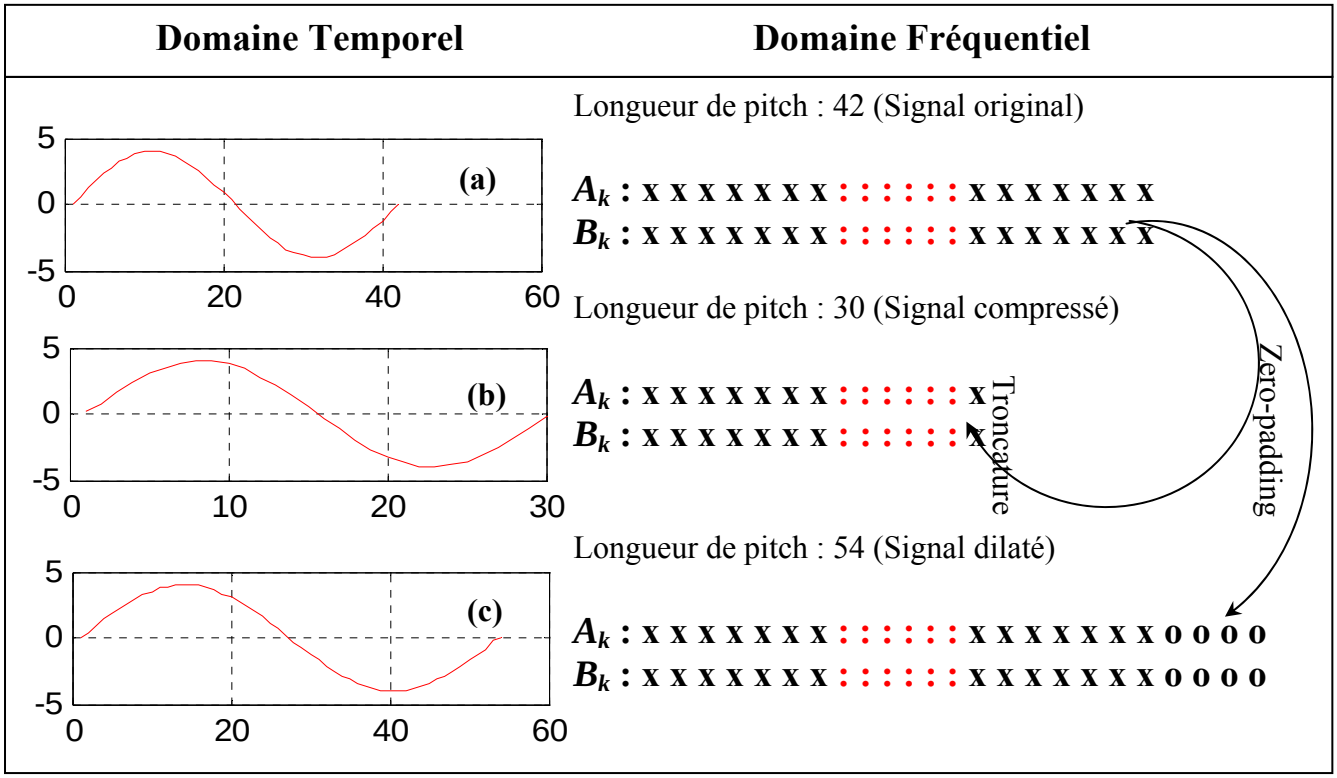


Fig. 3.8: Echelonnage temporel des CW, de la fonction sin(x).

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 3.9 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 3.2.2 en utilisant l'indicateur *C*. Si cet indicateur est différent de l'unité, alors, on insère *C*-1 zéro entre chaque deux coefficients DTFS.

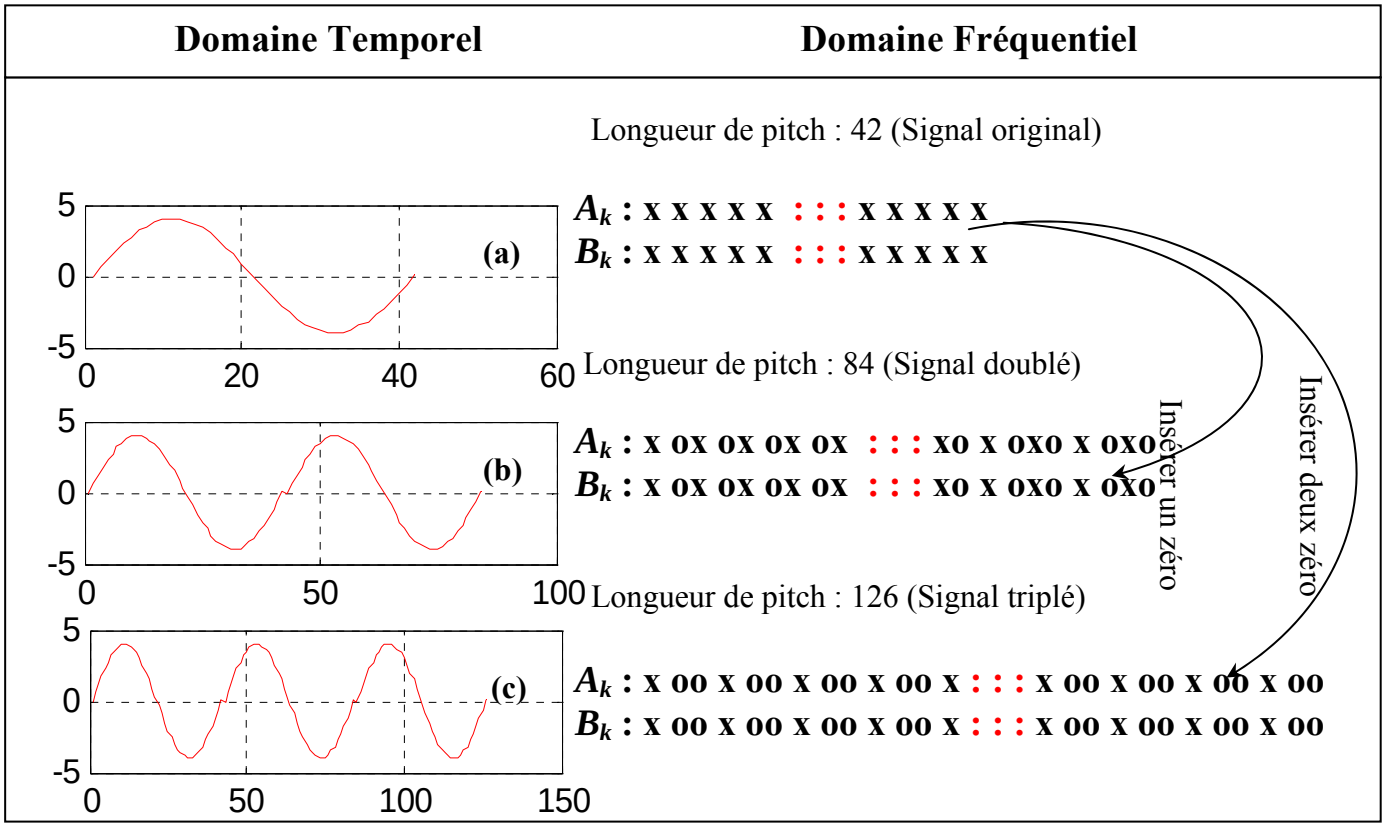


Fig. 3.9: Illustration de l'insertion de zéros entre les composantes spectrales.

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

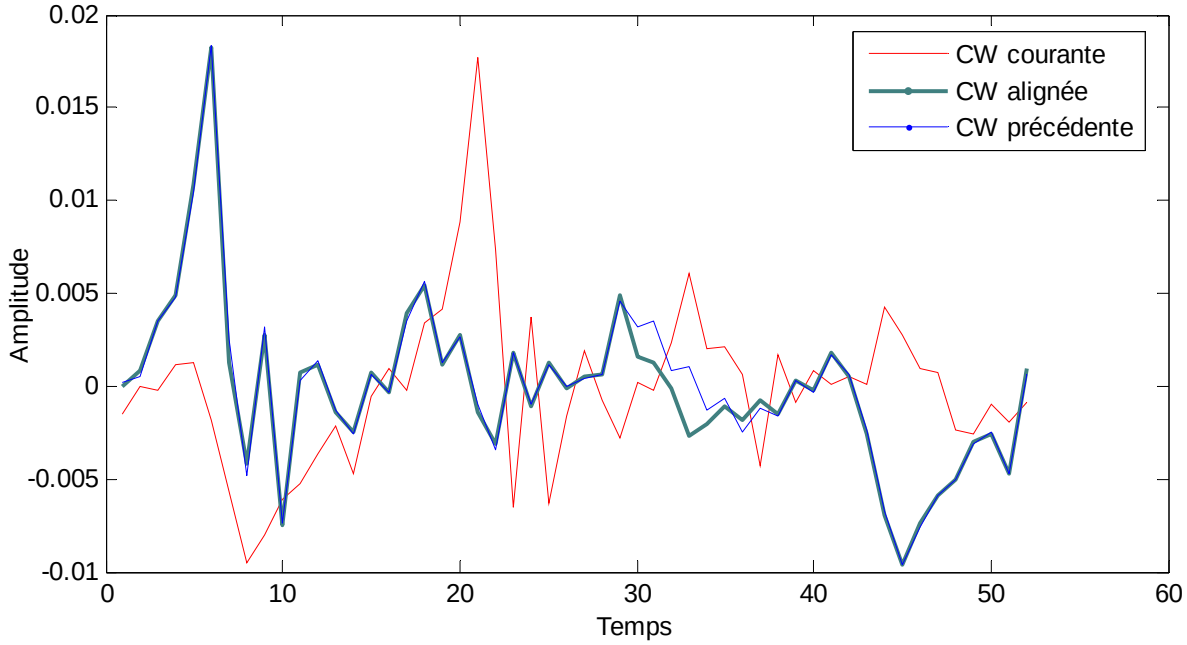


Fig. 3.10: Exemple du processus d'alignement pour deux CW adjacentes

3.2.5. 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 terme 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 et la normalisation sont réalisés sur les coefficients DTFS $\{A_k$ et $B_k\}$ [18].

La puissance moyenne d'une CW à l'instant n, notée $\Psi(n)$ Peut être exprimée par :

$$\Psi(n)=\frac{1}{P(n)}\sum_{m=0}^{P(n)-1}\left|s(n,m)\right|^2 \qquad \textbf{(3.17)}$$

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

$$\begin{aligned}\Psi(n)&=\frac{1}{P(n)}\sum_{m=0}^{P(n)-1}s(n,m).s^*(n,m)\\&=\frac{1}{P(n)}\sum_{m=0}^{P(n)-1}s(n,m)\sum_{k=0}^{P(n)/2}\left[A_k^*\cos\left(\frac{2\pi km}{P(n)}\right)+B_k^*\sin\left(\frac{2\pi km}{P(n)}\right)\right]\end{aligned} \qquad \textbf{(3.18)}$$

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

$$\Psi(n)=\left[\frac{1}{P}\sum_{k=0}^{P/2}A_k\sum_{m=0}^{P-1}s(m)\cos\left(\frac{2\pi km}{P}\right)\right]+\left[\frac{1}{P}\sum_{k=0}^{P/2}B_k\sum_{m=0}^{P-1}s(m)\cos\left(\frac{2\pi km}{P}\right)\right] \qquad \textbf{(3.19)}$$

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

$$\Psi(n)=\begin{cases}\dfrac{1}{2}\displaystyle\sum_{k=0}^{(P/2)-1}\left(A_k^2+B_k^2\right)+A_{P/2}^2+B_{P/2}^2 & si \quad P\ pair\\ \dfrac{1}{2}\displaystyle\sum_{k=0}^{(P/2)}\left(A_k^2+B_k^2\right) & si \quad P\ impair\end{cases} \qquad (3.20)$$

L'équation (3.20) est la formule utilisée, pour déterminer la puissance de la CW à partir de ses coefficients DTFS.

Donc, la normalisation consiste à faire diviser chaque coefficient DTFS, par la racine carrée de la puissance moyenne.

3.3. Le décodeur WI

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 3.11. Similaire aux processeurs du codeur, la fréquence d'exécution varie d'un bloc à un autre dans la couche de synthèse.

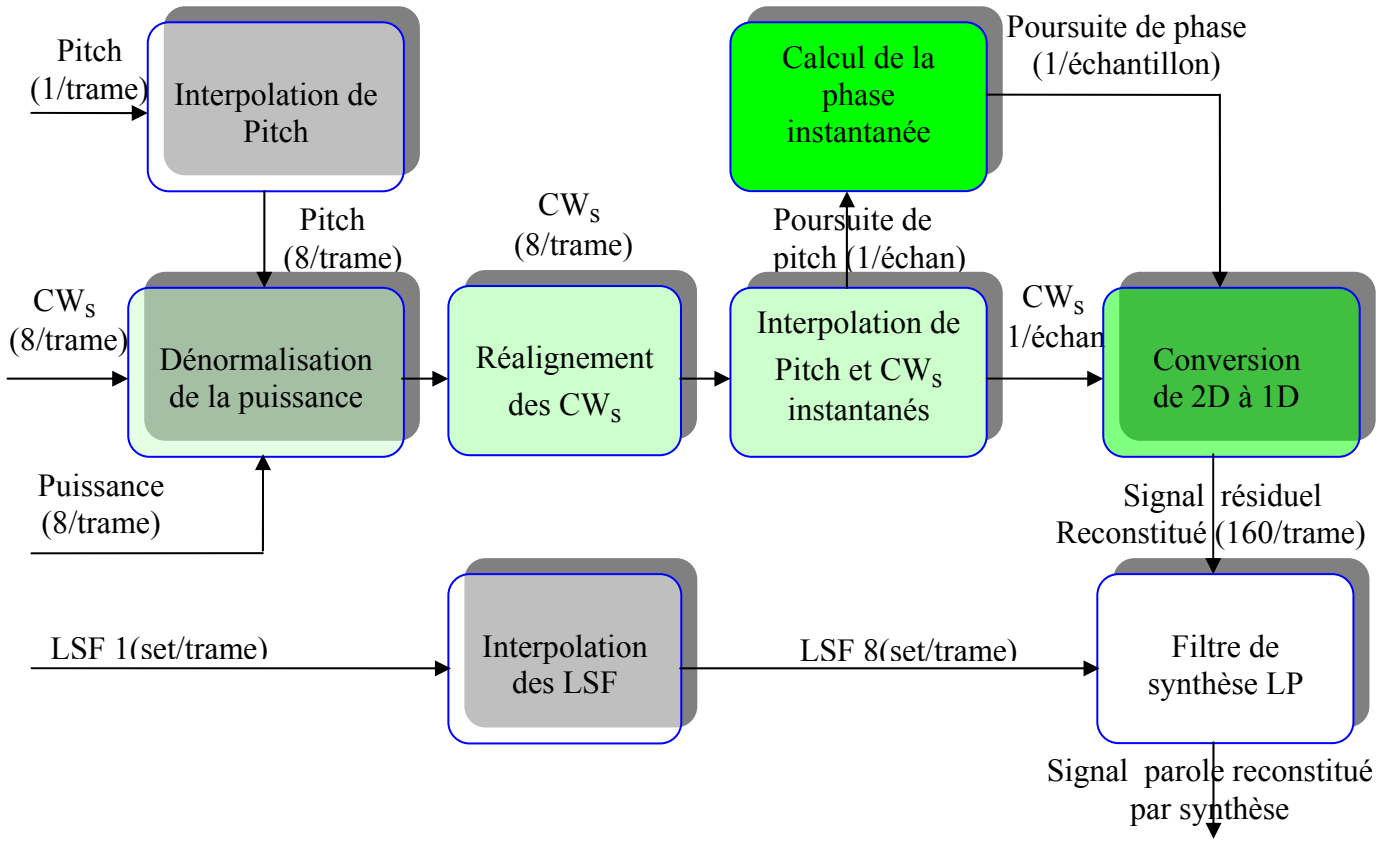


Fig. 3.11: Schéma bloc d'un décodeur WI. Les processeurs colorés en vert clair sont exécutés une fois par sous- trame tandis que ceux colorés en vert foncé sont exécutés à la fréquence des échantillons. Les autres sont exécutés une fois par trame [2].

3.3.1. Génération des pitchs et CW instantanés

Après la dé-normalisation 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 dé-quantifiées (si on inclut la couche de quantification), 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 [2].

La figure 3.12 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.

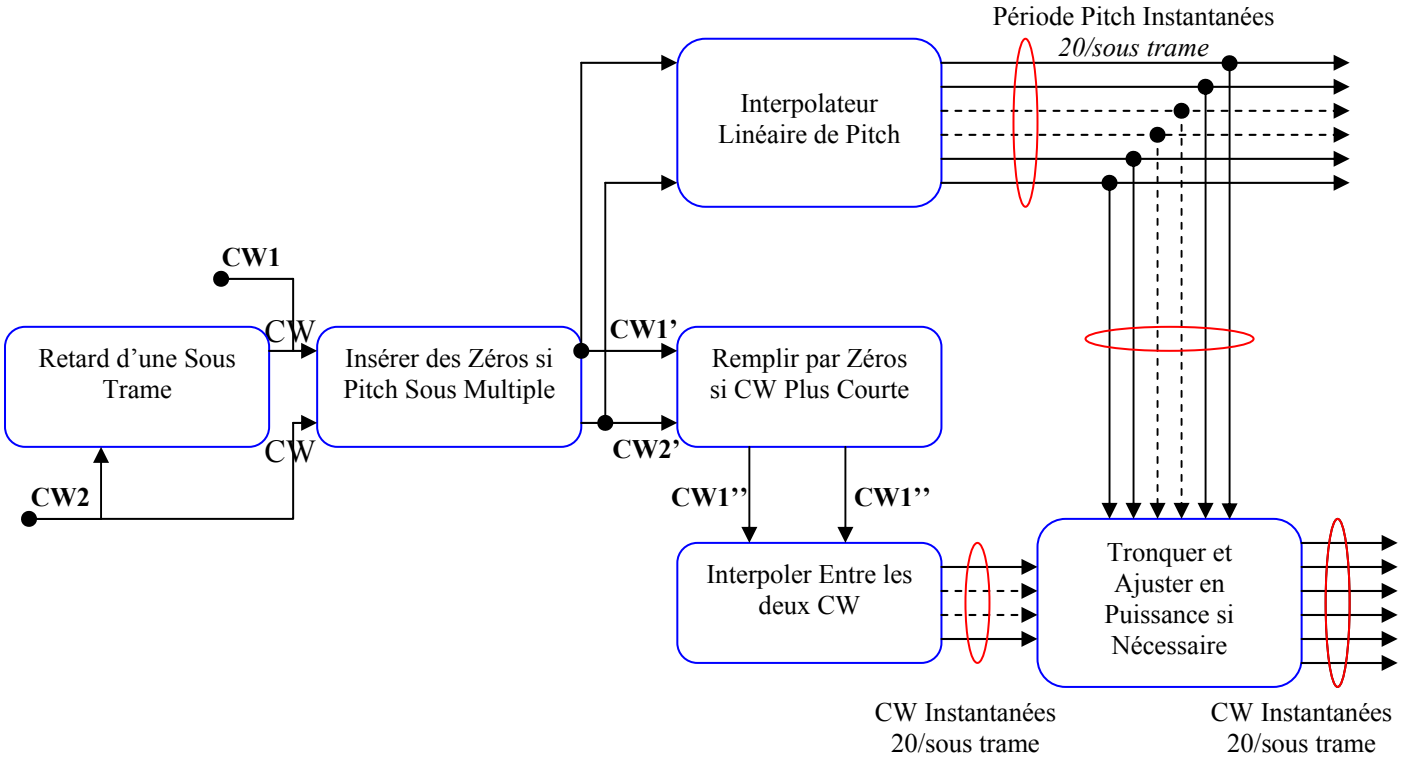


Fig. 3.12: Schéma bloc du processeur d'interpolation.

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 :

$$s(n,m)=\left(\frac{n_1-n}{n_1-n_0}\right)s(n_0,m)+\left(\frac{n-n_0}{n_1-n_0}\right)s(n_1,m) \qquad n_0 \le n \le n_1,\ 0 \le m \le P \quad \textbf{(3.21)}$$

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

$$\left.\begin{aligned} A_k(n) &= \left(\frac{n_1 - n}{n_1 - n_0}\right) A_k(n_0) + \left(\frac{n - n_0}{n_1 - n_0}\right) A_k(n_1) \\ B_k(n) &= \left(\frac{n_1 - n}{n_1 - n_0}\right) B_k(n_0) + \left(\frac{n - n_0}{n_1 - n_0}\right) B_k(n_1) \end{aligned}\right\} \quad \text{Pour } \mathbf{k = 1, 2, \ldots, [P/2]} \qquad \mathbf{(3.22)}$$

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 [3].

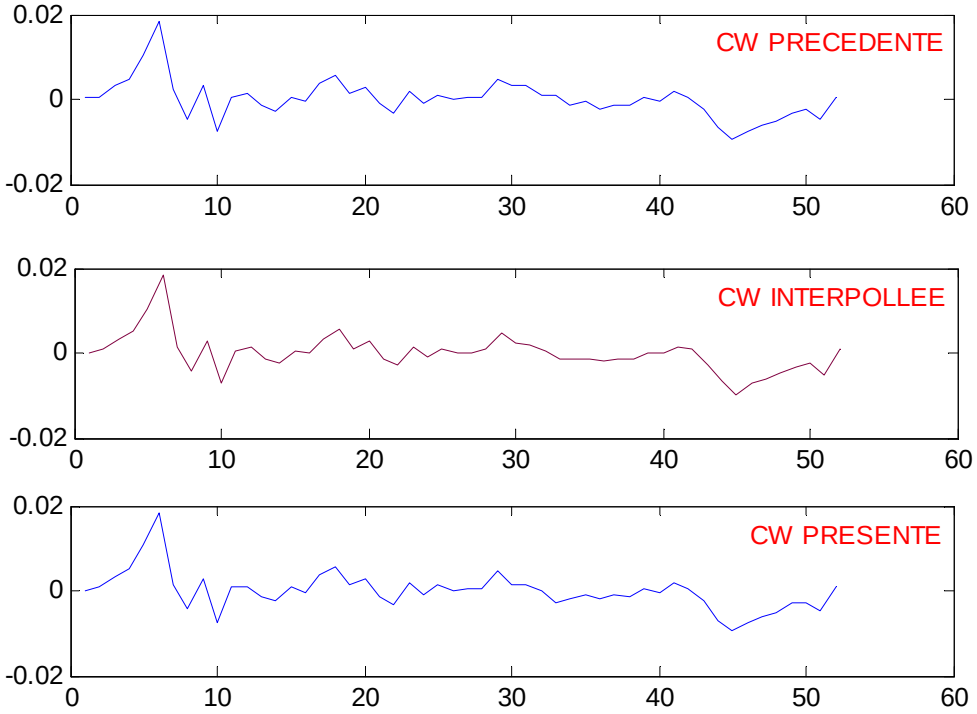


Fig. 3.13: Illustration du processus d'interpolation d'une CW au douzième échantillon.

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 3.2.4

Ainsi, l'équation d'interpolation linéaire conventionnelle (3.8) peut être utilisée pour sur-échantillonner le pitch. Cependant, les valeurs du pitch sur-échantillonnées, 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 3.2.2 on utilise l'indicateur *C* comme détecteur de (sous-) multiple de pitch.

Les *C*-1 zéros sont insérés entre les cœfficients DTFS, afin d'avoir compenser l'écart de longueur entre les deux CW ; ensuite les CW sont traitées de la même manière que dans le premier cas. La figure 3.14 montre un exemple d'interpolation des CW sur un intervalle d'une sous trame.

3.3.2. 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 [18] :

$$\phi(n) = \phi(n-1) + \int_{n-1}^{n} \frac{2\pi}{P(n')}\, dn' \qquad \textbf{(3.23)}$$

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

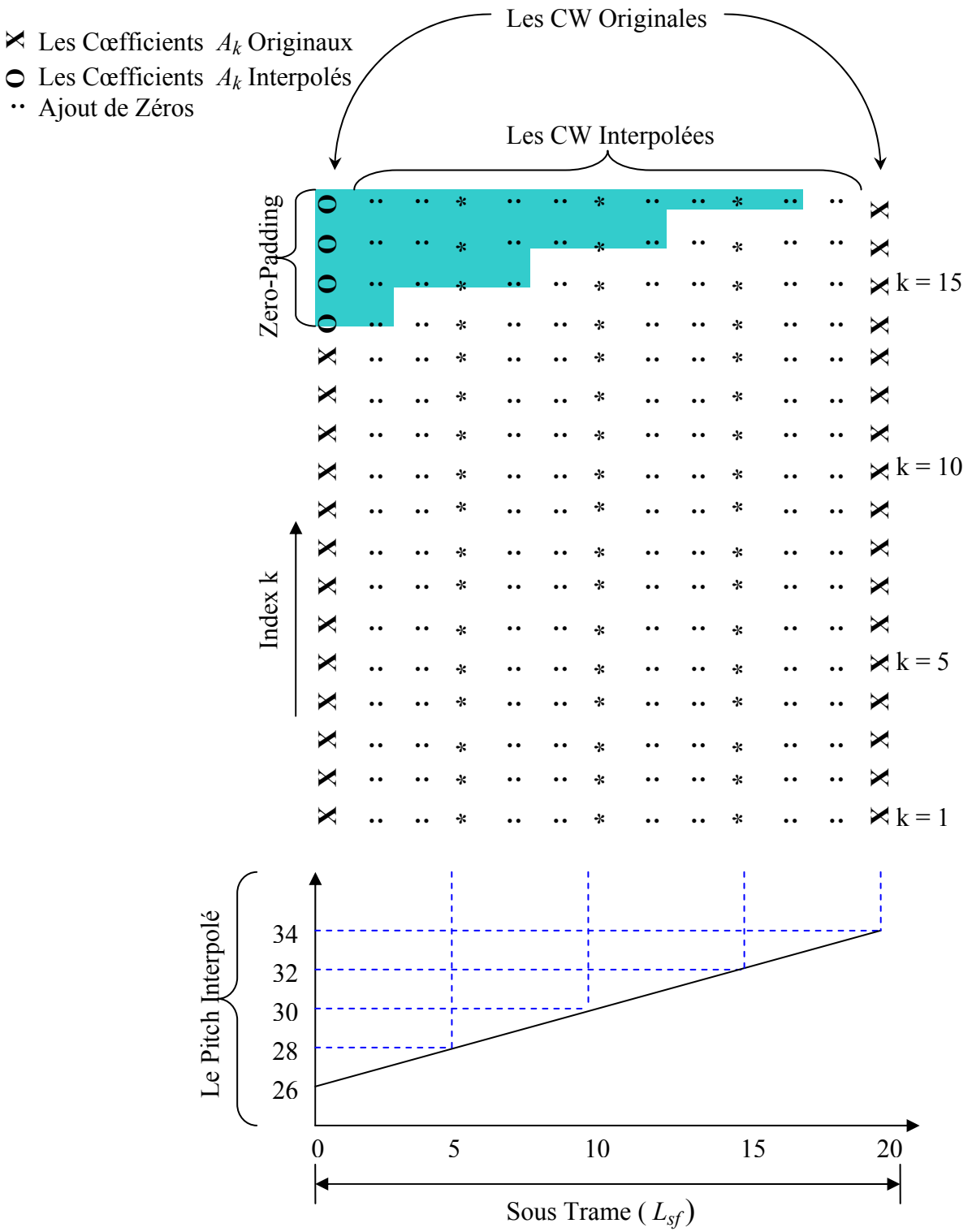


Fig. 3.14: Un exemple d'interpolation des CW sur un intervalle d'une sous-trame.

En supposant que le pitch évolue linéairement sur l'intervalle d'intégration, alors (3.23) peut être écrite sous la forme :

$$\phi(n)=\phi(n-1)+\int_{n-1}^{n}\frac{2\pi}{(n-n')P(n-1)+(n'-n+1)P(n)}\,dn' \qquad \textbf{(3.24)}$$

Une évaluation rapide de cette intégrale mène à :

$$\phi(n)=\begin{cases}\phi(n-1)+\dfrac{2\pi}{P(n)-P(n-1)}\ln\left[\dfrac{P(n)}{P(n-1)}\right] & si\quad P(n)\neq P(n-1)\\ \phi(n-1)+\dfrac{2\pi}{P(n)} & si\quad P(n)=P(n-1)\end{cases} \qquad \textbf{(3.25)}$$

Pour une implémentation en pratique, et afin de réduire la complexité de calcul, la relation (3.26) est une approximation fiable de (3.25) [3, 27].

$$\phi(n)=\phi(n-1)+\pi\left(\frac{1}{P(n-1)}+\frac{1}{P(n)}\right) \qquad \textbf{(3.26)}$$

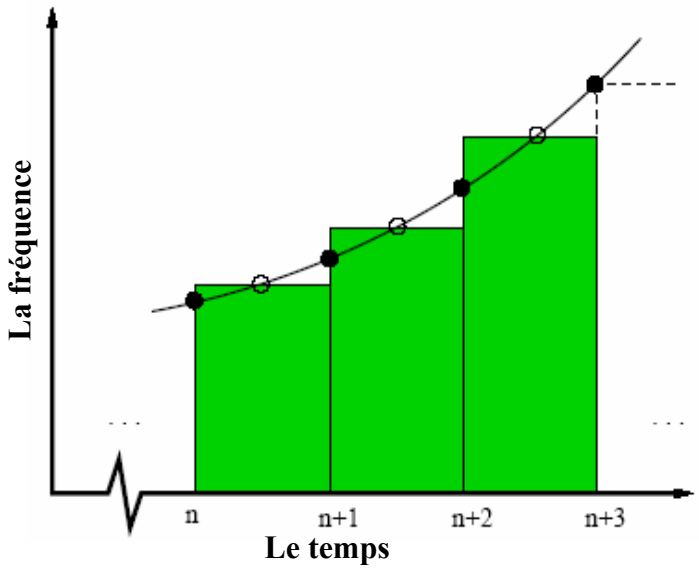


Fig. 3.15: Comparaison entre les deux approches de calcul de phase. [2]

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 initial.

3.3.3. 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 3.16 qui montre le processus de reconstitution [27], 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 [18, 19]. Ainsi $r(n)$ est exprimé par (voir Annexe A) :

$$\boldsymbol{r(n)=s(n,\phi(n))=\sum_{k=0}^{[P(n)/2]}\left[A_k(n)\cos(k\phi(n))+B_k(n)\sin(k\phi(n))\right]} \qquad \boldsymbol{0\le\phi(\cdot)<2\pi} \qquad \textbf{(3.27)}$$

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 la figure 2.1, et les coefficients du filtre sont donnés par la conversion des coefficients LSF en coefficients LP après interpolation [12].

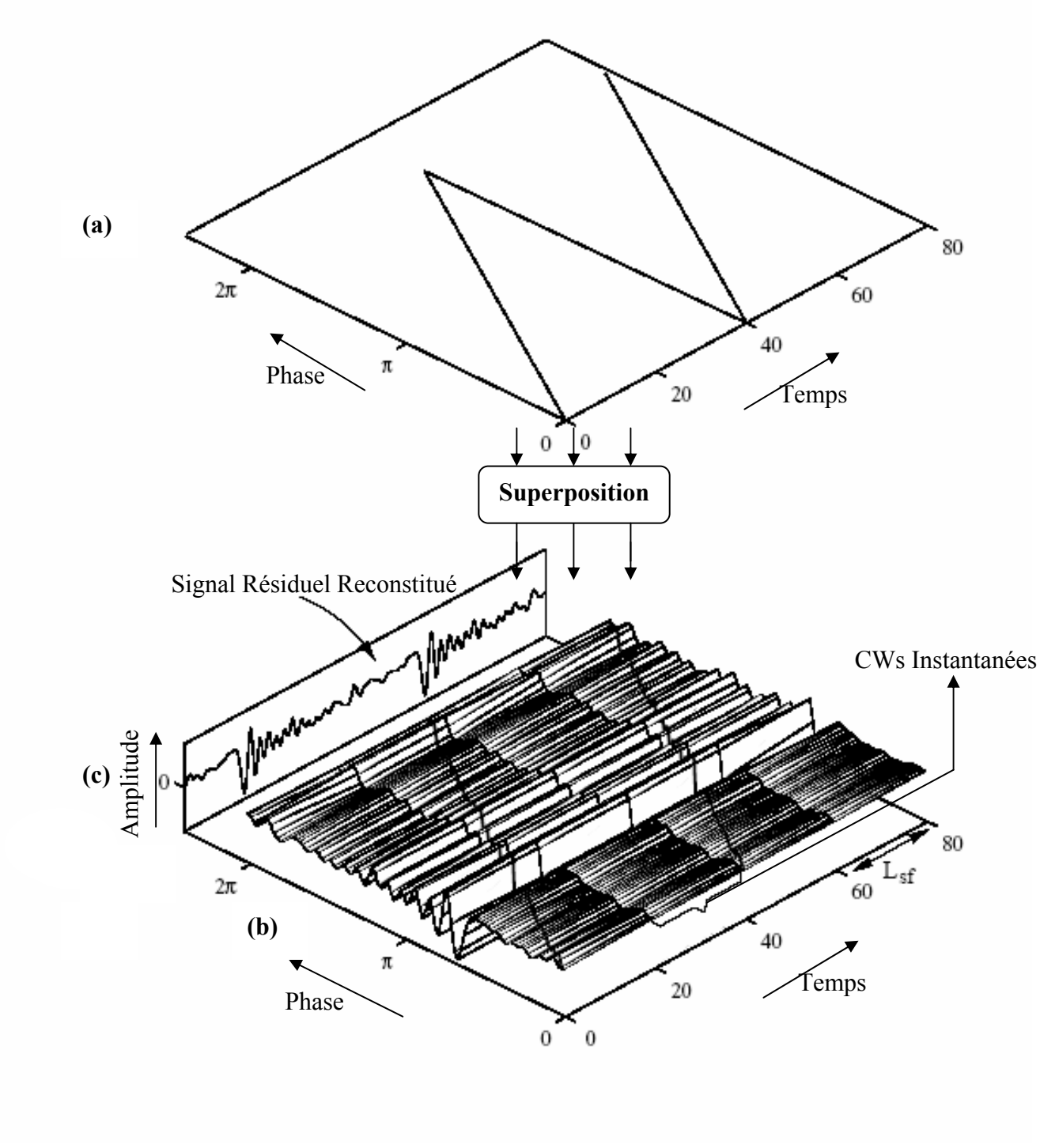


Fig. 3.16: Transformation de la surface 2D à 1D des CW [27].

3.4. Résultats d'évaluation de la qualité

D'après [18], une parole presque transparente peut être générée par le schéma non quantifié de la WI. Pour vérifier la précision de notre système d'analyse-synthèse, on a procédé à un test d'écoute, comparant des séquences de parole originales à celles reconstituées (sans quantification). Le test consistait à faire écouter les versions originales puis reconstituées de 12 phrases, dont 6 voix masculines, et 6 voix féminines ; les tests sont effectués pour les deux fréquences d'extraction des CW (400 Hz et 500 Hz).

La moyenne des évaluations subjectives était d'environ 4/5 pour les deux fréquences d'extractions, ce qui correspond à une bonne qualité de la parole reconstituée (à peine perceptible mais pas gênante, selon l'échelle MOS). La figure 3.17 donne un résultat des tests dans lequel on voit les représentations temporelles des phrases originale et reconstituées.

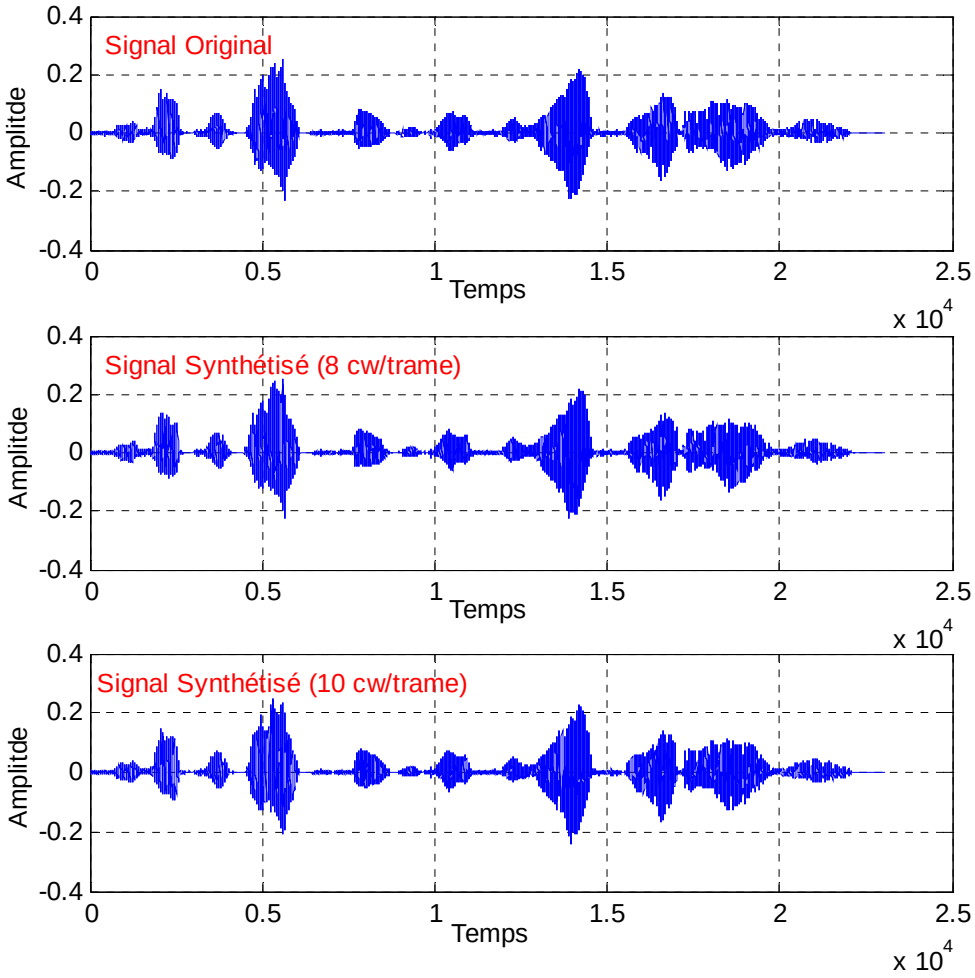


Fig. 3.17: Graphes original et reconstituées, à différentes fréquences d'extraction d'une voix masculine extraite du corpus TIMIT "She had your dark suit in greasy wash water all year".

Les épreuves précédents ont été exécutées à R_{extr} = 8 et R_{extr} = 10. On a remarqué qu'avec les deux débits d'extraction, on a obtenue approximativement les mêmes résultats, il y avait seulement une mineure amélioration dans la qualité de la parole résultante à R_{extr} = 10. Donc on préféra travailler avec R_{extr} =8 au lieu de 10, à raison de diminuer la complexité du codeur.

Concernant l'évaluation objective, la moyenne d'évaluation perceptuelle de la qualité vocale PESQ était de 3.6, ce qui désigne une qualité perçue acceptable. Mais généralement les tests objectifs ne conduisent pas à une bonne interprétation du résultat; par exemple dans le cas de la phrase précédente la mesure du SNR entre la parole originale et celle codée vaut -2.6391 dB et celle de SEGSNR vaut -2.4688 dB.

Ces valeurs négatives des SNR sont essentiellement dues à la désynchronisation de la WI dans le temps, entre le signal original et reconstitué [2]; 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.

3.5. Application de la 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 indésirables effets apparaissent due au filtrage adaptatif, ainsi que la parole reconstruite peut exhiber une enveloppe indésirable, dont le résultat est un sifflement audible, pour certains sons [3, 28].

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 sifflements dans la parole reconstruite, plusieurs tentatives étaient réalisées, tel qu'une compensation d'énergie pour lisser l'enveloppe temporelle de la parole reconstruite [3, 28] ; 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 [3]. 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 (voir figure 3.18), 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 a partir du signal résiduel, pour que sa valeur soit plus précise, et les forme d'ondes sont extraite à partir du signal original, ensuite les mêmes procédures précédentes sont appliquées.

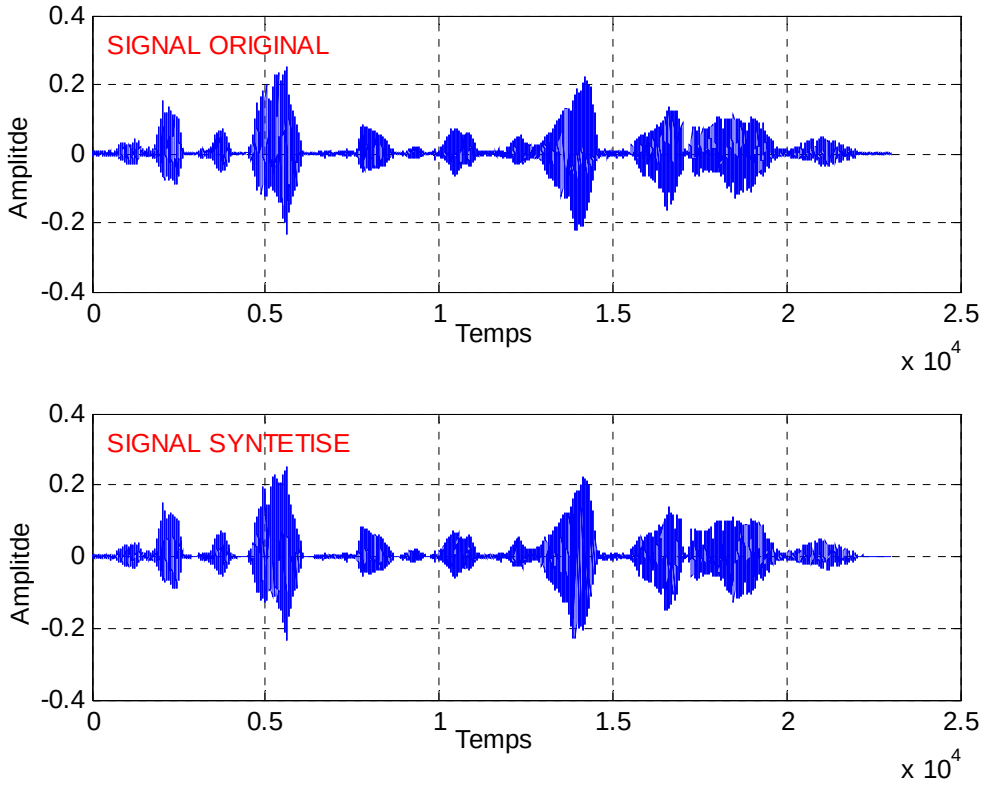


Fig. 3.18: Application de la WI sur le signal originale : graphes original (en haut) et reconstitué (en bas) de la même phrase précédente.

Les tests d'écoute ont montrés que la parole reconstituée contenu quelque rugosités (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 parait efficace ; Celle-ci 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) [21] ; dans cette dernière la détection du pitch est exécutée 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 seront étendues par le processus de sur-échantillonnage, donc le nombre d'opérations sera plus élevé.

3.6. 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 voisé / non voisé, une nouvelle technique de décomposition [18], dont 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 périodiques et non périodiques du signal parole. En exploitant la différence dans la perception humaine de ces deux formes d'ondes, une meilleure efficacité du 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 des impulsions de forme périodique 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.

3.6.1. Conception du filtre passe-bas

C'est un filtre Anti-repliement[6] non causal à phase linéaire. Cependant, ce filtre nécessite une fréquence de coupure de 20 Hz [18] équivalente à la fréquence normalisée 0.1, où 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 où 25 Hz) avec une fenêtre de Hamming de longueur 17 échantillons. Finalement, on peut obtenir $h_{CW}(i)$ en normalisant la réponse fenêtrée :

$$\sum_{i=-8}^{8} h_{CW}(i) = 1 \qquad \textbf{(3.28)}$$

La figure 3.19 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 [19] pour avoir une meilleure précision du filtre, mais aux dépends d'un retard algorithmique plus important.

3.6.2. 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 suivante [2] :

$$\left.\begin{aligned} \widetilde{A}_k(n) &= \sum_{i=-8}^{8} A_k(n - iL_{sf})\, h_{CW}(i) \\ \widetilde{B}_k(n) &= \sum_{i=-8}^{8} B_k(n - iL_{sf})\, h_{CW}(i) \end{aligned}\right\} \qquad \textbf{(3.21)}$$

[6] Ce filtre évite l'effet dénommé « aliasing » postulé par le théorème de Nyquist-Shannon.

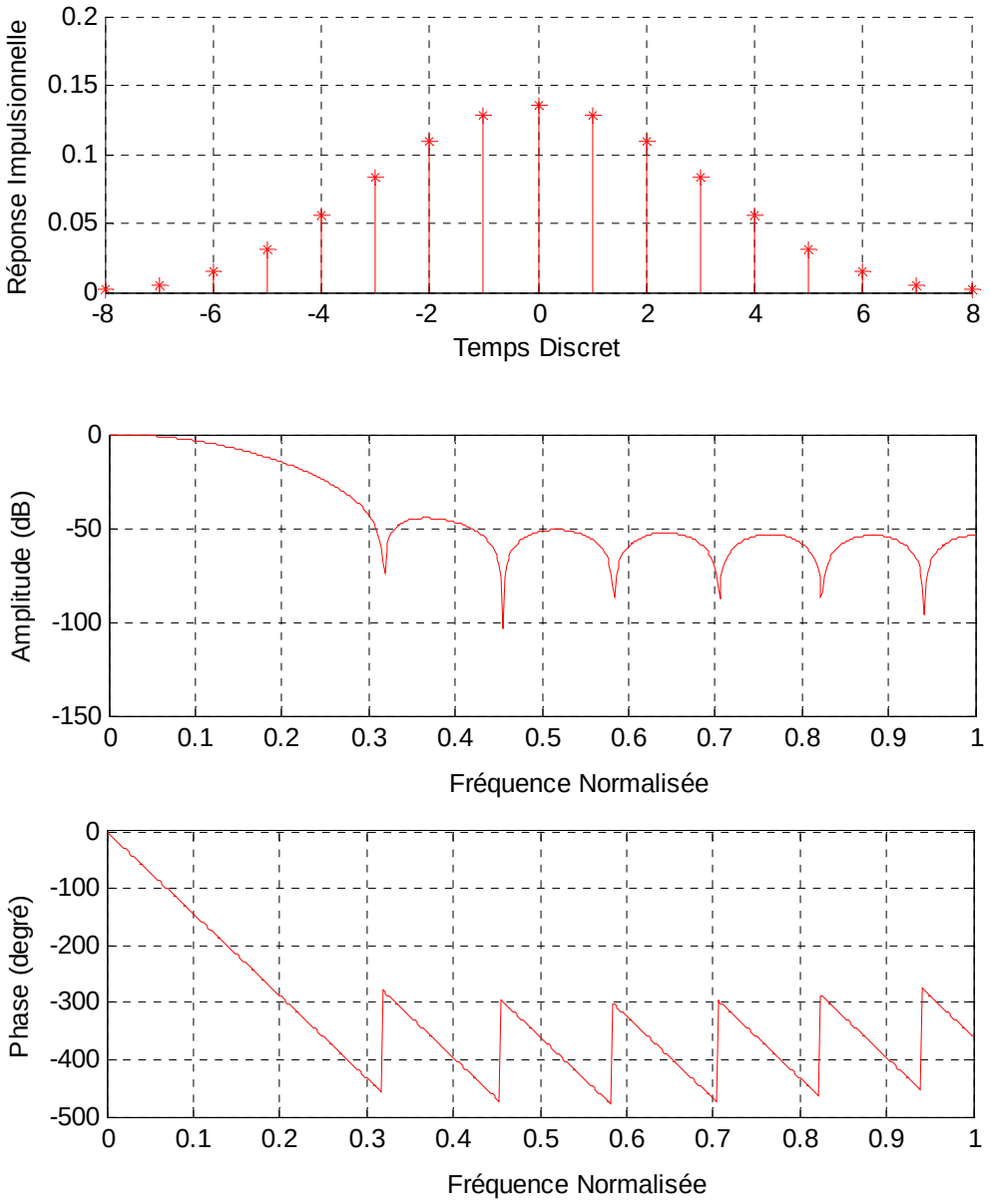


Fig. 3.19: Caractéristiques du filtre passe-bas de décomposition en SEW-REW. La fréquence de coupure normalisée égale à 0.1 (20Hz).

Or, la dimension des CW varie avec le pitch. Pour faciliter le filtrage, les mêmes techniques des paragraphes précédents 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 3.20 décrit l'opération d'ajustements des CW appliquée avant le filtrage.

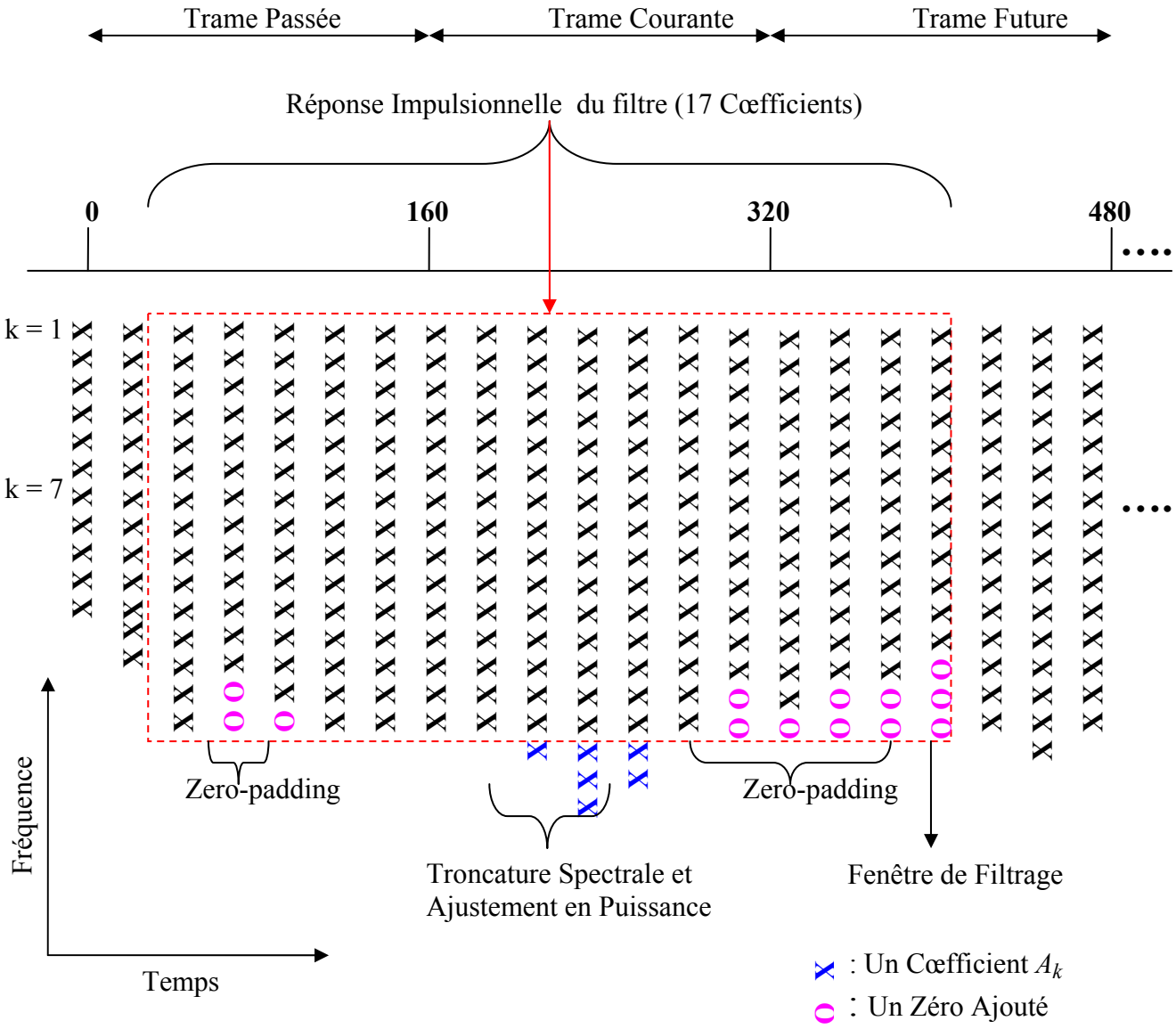


Fig. 3.20: Opération de filtrage passe-bas pour la décomposition en SEW-REW. Le Schéma montre 17 CW successives couvrant trois trames.

Ensuite, la procédure de filtrage est exécutée *k* par *k* (ligne par ligne). La figure 3.21 illustre un exemple de décomposition en deux surfaces SEW et REW de 17 formes d'ondes successives.

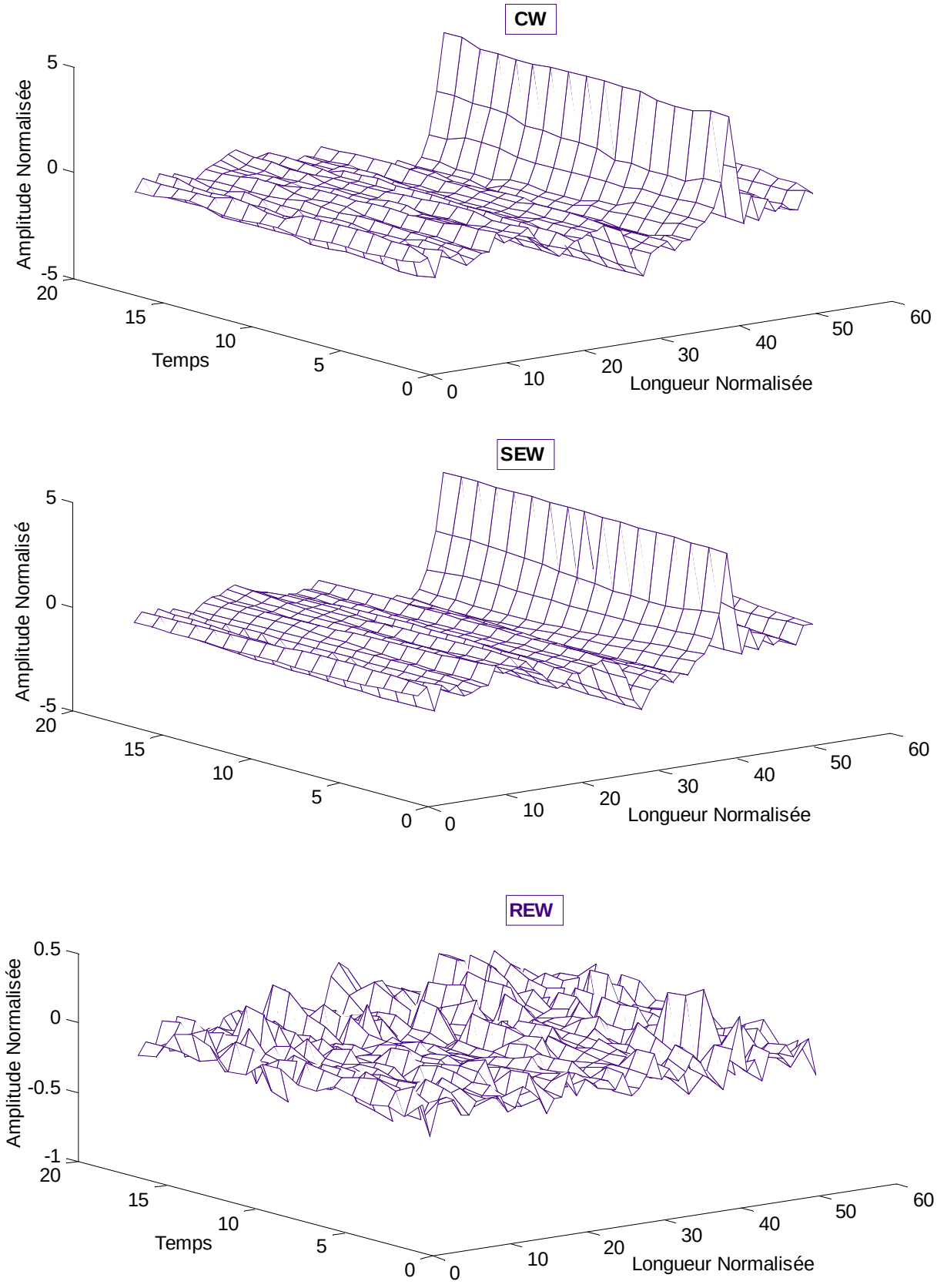


Fig. 3.21: Décomposition d'un segment de longueur 40 ms (17 CW) en surfaces SEW et REW. La fréquence de coupure du filtre passe-bas est de 20 Hz.

3.7. Conclusion

Les travaux réalisés dans ce troisième chapitre nous on permis de mieux voir le fonctionnement d'un codeur WI avec ces différentes difficultés de calculs, ainsi on a abouti a faire la simulation de la couche d'analyse et celle de synthèse ; la nouvelle méthode de détection de pitch que nous avons employé nous a permis d'avoir un meilleur alignement des formes d'onde, donc la qualité de reconstitution sera plus élevée. La simulation a été exécutée pour deux fréquences d'extraction des CW, c-à-d pour 8 et 10 extractions par trame d'analyse, le but d'employer cette diversité de fréquence d'extraction est de visualiser l'effet de la fréquence sur la qualité de reconstitution. Par ailleurs, sur le point de vue perception la différence n'était pas vraiment remarquable ; par conséquent on a préféré d'utiliser la fréquence de 8 extractions par trame pour les prochains calculs, à raisons de diminuer le nombre d'opérations et d'espace mémoire.

Pour évaluer la qualité du codeur en l'absence de la couche de quantification, on a employé des de tests d'écoutes (MOS) qui ont donné des résultats satisfaisants, mais pour certains sons on a observé quelques sifflements dans la parole reconstituée, pour remédier ce problème on a essayé d'appliquer la méthode proposée par M. Leong [3, 28], qui consiste à appliquer l'algorithme de la WI directement sur le signal de parole. On a observé que l'enveloppe temporelle est améliorée par rapport au cas précèdent, mais cette modification a conduit à l'apparition d'un caractère bruyant dans la parole reconstituée, qui est dû principalement à la haute limite d'énergie aux extrémités des formes d'onde extraites

Finalement, on a abouti à faire la décomposition des formes d'onde en une composante périodique SEW et une autre composante purement aléatoire qui est la REW, cette décomposition nous permet de faire la quantification de chaque composante séparément de l'autre, ainsi l'efficacité du codeur WI se centralise dans la meilleure qualité de codage de ces deux composantes, donc il est important de concevoir le filtre adéquat durant le processus de décomposition des formes d'onde.

Chapitre 4 :

Quantification des REW

4.1. Introduction

La couche d'analyse-synthèse dans la WI (en l'absence de la 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 [8, 9] 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 technique de 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) [21, 29].

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 [18], 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) [18, 30]. Les valeurs sous-échantillonnées sont codées par la technique de quantification vectorielle par analyse et synthèse sur 6 bits [30]. 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 méthode 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.

Dans ce quatrième chapitre on pressente et on étudie une méthode de quantification de la composante REW, qui pressente la partie aléatoire ou non périodique de la forme d'onde d'onde CW. Cette méthode fait partie de plusieurs tentatives pour la quantification de la REW et de certaines transformations mathématiques tel que la DCT et l'ajustement polynomiale. Avant d'entamer la section de quantification des CW, particulièrement les REW on parlera tout d'abord de la théorie de la quantification.

4.2. Notions sur la quantification

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. Cette opération consiste à représenter les amplitudes du signal analogique, à des instants discrets dans le temps, par une valeur choisie parmi un ensemble fini. Outre la nécessité de la quantification pour numériser les données, elle est aussi un moyen de compression.

Dans cette partie, nous présenterons rapidement les notions de base de la quantification scalaire (QS) qui traite chaque échantillon indépendamment des précédents, et introduirons la quantification multidimensionnelle ou vectorielle (QV).

4.2.1. Quantification scalaire

Posons S le paramètre à quantifier, pouvant représenter tout paramètre extrait du signal, tel que l'amplitude du signal échantillonné. Notons f_S (s) la densité de probabilité de la variable S et b sa résolution, c-à-d le nombre de bits pour la représenter. L'opération de quantification consiste à discrétiser le paramètre S pour obtenir une représentation numérique i (n) de l'information qu'il représente ; son domaine de définition est alors partitionné en $L = 2^b$ intervalles distincts, et un représentant est défini pour chacun de ces intervalles.

La procédure d'encodage Q décide à quel intervalle appartient le paramètre s (n) et lui associe le numéro i (n) ∈ {1,...,L} correspondant (figure 4.1). C'est ce numéro d'intervalle qu'il faudra transmettre au processus de décodage Q^{-1} qui effectuera la procédure inverse en associant à i (n) son représentant $\hat{s}^{i}(n)$ [5].

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

$$\boldsymbol{e(n) = s(n) - \hat{s}^{i}(n)} \qquad \textbf{(4.1)}$$

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

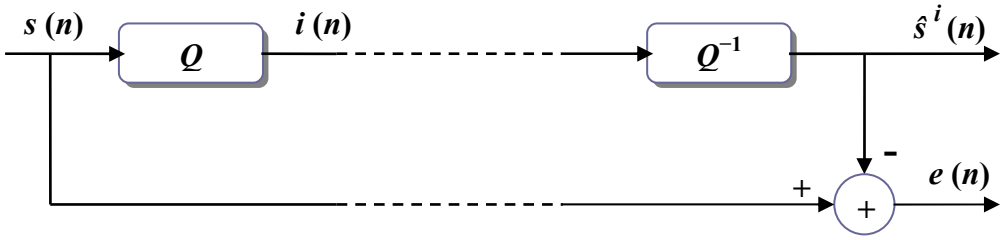


Fig. 4.1: Quantificateur scalaire.

4.2.1.1. Quantification uniforme

Le processus de quantification uniforme est tout à fait irréaliste, mais permet de présenter de façon simple la quantification scalaire. Cette technique consiste à partitionner l'intervalle [−A, A], où le signal à temps discret s (n) prend ses valeurs avec une loi uniforme, en L intervalles distincts { P^1, ..., P^L } de même longueur $\Delta = 2A/L$.

Le représentant $\hat{s}^{i}(n)$ est défini généralement par le milieu de l'intervalle (figure 4.2), pour répondre au mieux, au critère de minimisation de l'erreur quadratique moyenne σ_s^2 :

$$\sigma_s^2 = E\left\{\left(s(n) - \hat{s}^i(n)\right)\right\}^2 \qquad \textbf{(4.2)}$$

Un critère de maximisation du rapport signal sur bruit de quantification SNR peut aussi être choisi.

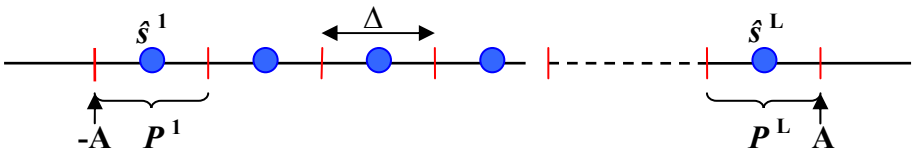


Fig. 4.2: Partition uniforme d'un intervalle [5].

4.2.1.2. Quantification non uniforme

Les hypothèses précédentes sont très mal adaptées à un processus réel, tel que le signal de parole, dont les propriétés statistiques ne sont pas stationnaires. Leur loi de probabilité n'étant pas uniforme, le quantificateur précédent n'est plus optimal. Donc la partition $\{ P^1, ..., P^L \}$ ne pourra plus être composée d'éléments de longueur constante. La longueur de chaque intervalle devra être d'autant plus petite que la densité de probabilité $f_S(s)$ correspondante sera grande (figure 4.3). De plus, les représentants $\hat{s}^{i}(n)$ ne seront plus forcément les milieux des intervalles mais seront déterminés par la moyenne de la variable aléatoire S sur l'intervalle considéré.

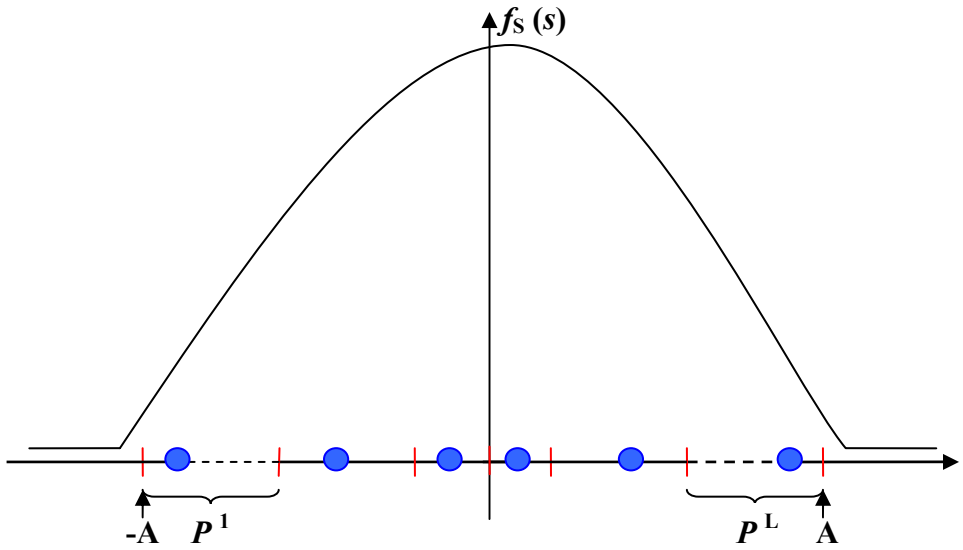


Fig. 4.3: Quantificateur scalaire non uniforme [5].

Dans la pratique, on ne connaît pas la densité de probabilité. Pour construire un quantificateur, des données empiriques sont utilisées en associant à chaque échantillon le même poids. La base d'apprentissage devra être composée d'un grand nombre d'échantillons représentatifs de la source. L'algorithme de Lloyd-Max [31] permet ainsi de construire un quantificateur optimal.

4.2.2. Principe de la quantification vectorielle

Contrairement à la quantification scalaire (QS) qui traite chaque échantillon indépendamment des précédents, la quantification vectorielle permet de prendre en compte la dépendance entre les différentes composantes du signal .La propriété fondamentale de ce quantificateur est la recherche de la corrélation qui peut exister entre les échantillons successifs du signal.

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

L'algorithme de Lloyd-Max proposé par Lloyd en 1957 et par Max en 1960 est un moyen pour optimiser une paire de codeur- décodeur de façon itérative. Chaque étape de l'algorithme optimise soit le codeur soit le décodeur, en gardant l'autre partie fixe.

Principalement il est caractérisé par les deux entités suivantes :

Le Plus Proche Voisin: la première étape est basée sur le principe du plus proche voisin selon lequel chaque entrée est codée par le niveau de quantification le plus proche.

Le Centroide: la deuxième étape est basée sur le principe du centroide selon lequel le représentant le plus apte pour un intervalle de quantification est la moyenne de toutes les entrées qui sont codées dans cet intervalle.

En1980, Linde, Buzo et Gray ont proposé la version vectorielle de cet algorithme surnommée LBG (Linde Buzo and Gray) [32]. Le LBG reste pratiquement identique à l'algorithme de Lloyd-Max, sauf que les scalaires sont remplacés par des vecteurs et les intervalles de quantification par les régions de Voronoi. Un exemple d'une quantification vectorielle à deux dimensions est montré dans la figure ci-dessous.

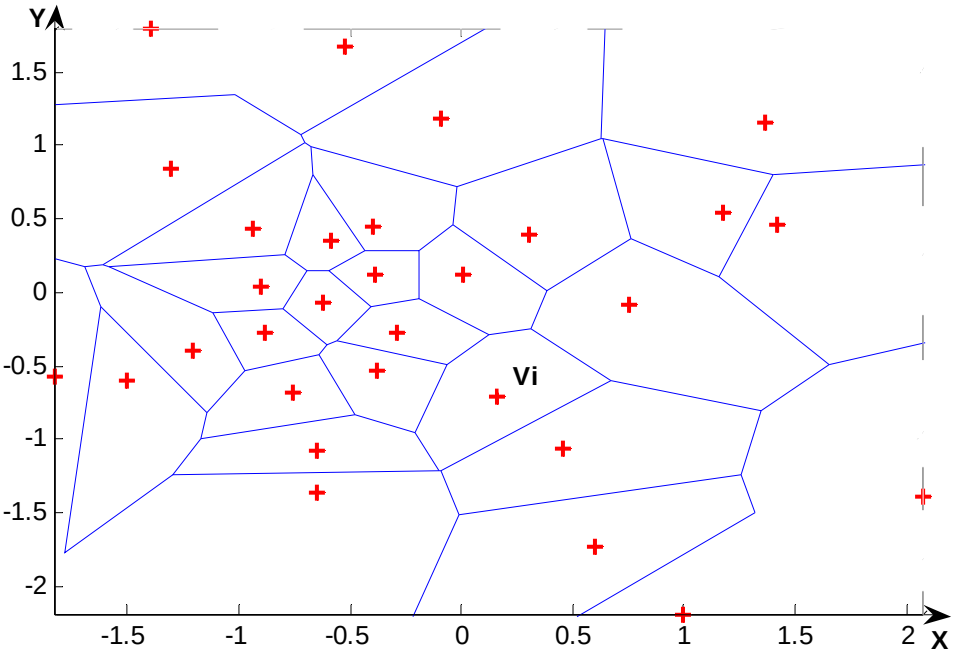


Fig. 4.4: Quantification vectorielle a deux dimensions.

Un quantificateur vectoriel Q' est dit globalement optimal s'il minimise une distorsion donnée D. Q' est optimal si pour tout autre quantificateur Q composé comme Q' d'un dictionnaire a N_c vecteurs, on a $D(Q') \leq D(Q)$.

Q est localement optimal si $D(Q)$ est un minimum local, une faible perturbation sur Q augmente la distorsion $D(Q)$.

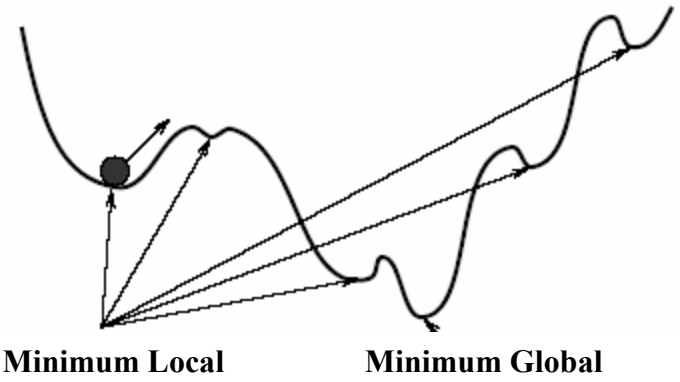


Fig. 4.5: Convergence d'un quantificateur vectoriel.

4.2.2.1. Détail d'un quantificateur vectoriel

Un quantificateur vectoriel peut être vu comme une application Q associant à chaque vecteur d'entrée $X_i = (x_j ; j = 1 \ldots k)$ un vecteur $Y_i = (y_j ; j = 1 \ldots k) = Q(X_i)$ choisi parmi un dictionnaire de taille finie $C = (\tilde{X_l} , l = 1 \ldots N_c)$, C peut être vu comme un catalogue de formes.

Le quantificateur est complètement décrit par :

- le dictionnaire (code book) C.
- le partitionnement $S = (S_i , i = 1,\ldots, N_c)$ qui divise l'espace d'entrée en N_c vecteurs X_i , et qui leur fait correspondre un vecteur $Q(X_i) = \tilde{X_i}$

La quantification vectorielle peut ainsi être vue, comme une combinaison de deux opérations (figure 4.6) :

- Un codeur qui reçoit une entrée X_i et qui recherche dans un dictionnaire l'adresse du vecteur qui lui ressemble le plus.
- Un décodeur qui reçoit l'adresse et génère le vecteur $\tilde{X}_i$ correspondant du dictionnaire, qui constitue une approximation du vecteur initial.

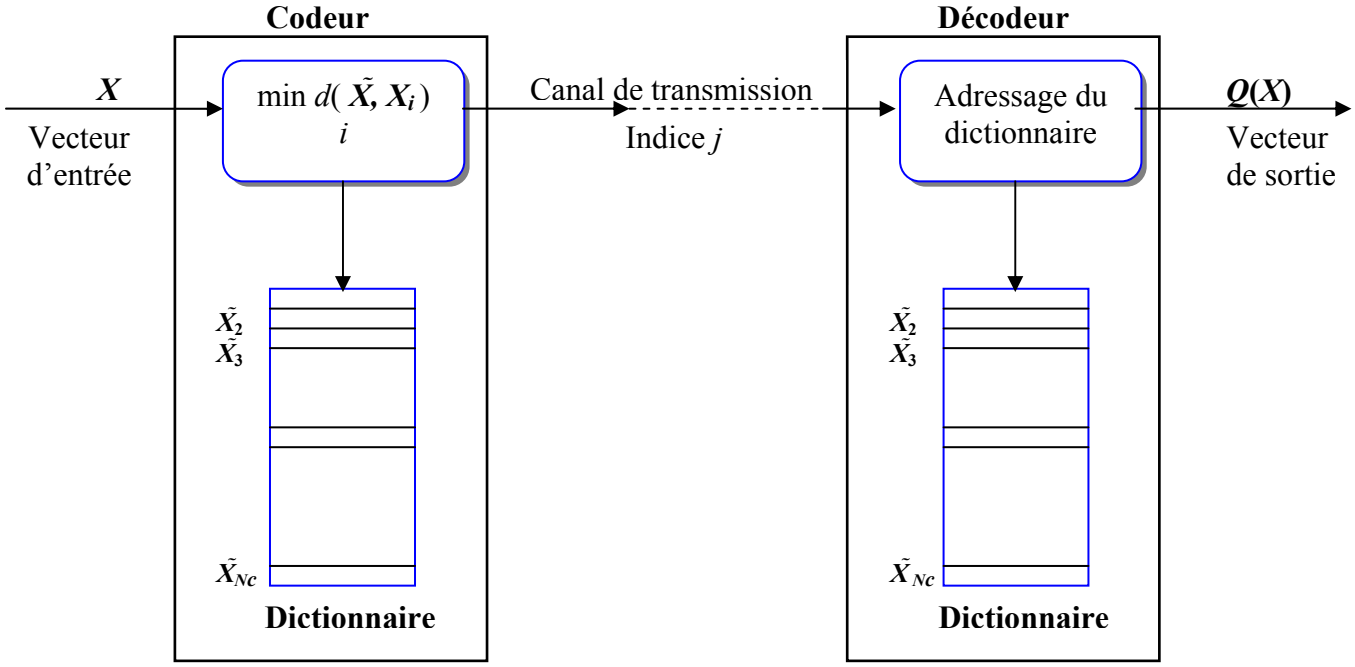


Fig. 4.6: Principe de la quantification vectorielle.

Le débit est donné par R = $\log_2 N_c / k$ pour chaque échantillon; ce débit est lié à la dimension k des vecteurs à coder de la séquence d'apprentissage, et à la taille du dictionnaire.

4.2.2.2. Mesure de la distorsion

Différentes mesures de distorsion ont été proposées dans la littérature, La mesure idéale doit évaluer la qualité subjective de la parole reconstituée.

L'erreur quadratique moyenne est la mesure la plus souvent utilisée en raison de sa simplicité. Elle s'exprime par :

$$d(X,Y)=\sum_{i=1}^{k}\left|x_i-y_i\right|^2 \qquad \textbf{(4.3)}$$

Dans un quantificateur on cherche à construire un dictionnaire optimal de N_c vecteurs, au sens où il minimise une distorsion moyenne donnée par :

$$D(X,Q(X))=\frac{1}{N_c}\sum_{i=1}^{N_c} d\left(X_i,\hat{X}_i\right) \qquad \textbf{(4.4)}$$

Le premier algorithme de classification a été proposé par Linde Buzo et Gray (voir paragraphe suivant). On peut aussi obtenir un dictionnaire à l'aide des cartes topologiques de Kohonen, ou de techniques d'optimisation stochastiques comme le recuit simulé, ce dernier permettant d'atteindre un optimum global au prix d'un accroissement du coût de calculs.

4.2.2.3. Détail de l'algorithme LBG

L'algorithme LBG a pour but de générer une partition sur un signal (séquence d'apprentissage), partant d'un dictionnaire initial composé des vecteurs les plus éloignés possible.

Ces vecteurs doivent être représentatifs des vecteurs rencontrés parmi les signaux à coder. L'algorithme itératif converge vers un dictionnaire localement optimal.

Le déroulement d'un algorithme LBG est décrit comme suit :

L'étape (1): On se donne un dictionnaire initial[1] C^0 composé de N_c vecteurs ($\tilde{X}_i$, i =1,…, N_c), une mesure de distorsion d, un seuil $\varepsilon \geq 0$, un compteur d'itération $l = 0$, et une distorsion moyenne D^{l-1} initialisée avec une très grande valeur et une séquence d'apprentissage composée de n vecteurs (X_j , j =1,…, n).

L'étape (2): Partant du dictionnaire $C^l = (\tilde{X}_i , i = 1,\ldots, N_c)$, trouver la partition $S^l = (S_i , i = 1,\ldots, N_c)$ de la séquence d'apprentissage minimisant la distorsion :

[1] Le dictionnaire initial est créé par une méthode aléatoire, où par méthode de divisions successives.

$X_j \in S_i$ si $d(\tilde{X}_i, X_j) \leq d(\tilde{X}_m, X_j)\ \forall m$. Autrement dit :

- Pour tous les vecteurs X_j de la séquence d'apprentissage ($j = 1,\ldots, n$).
- Pour tous les vecteurs $\tilde{X}_i$ du dictionnaire ($i = 1,\ldots, N_c$).
- Si $d(\tilde{X}_i, X_j) \leq d(\tilde{X}_l, X_j)\ \forall\ l$ alors $X_j \in S_i$.

La recherche de la distorsion minimale définit une région de décision S_i pour chaque vecteur $\tilde{X}_i$ du dictionnaire. Chaque vecteur X_j de la séquence d'apprentissage inclus dans la région de décision S_i est approximé par le vecteur $\tilde{X}_i$ associé.

L'étape (3): Calculer la distorsion moyenne D^l définit par :

$$\boldsymbol{D^l = D\left(C^l, S^l\right) = \frac{1}{n}\sum_{j=1}^{n} \min_{\hat{X} \in C^l}\ d\left(X_j, \hat{X}\right)} \qquad \textbf{(4.5)}$$

Cette expression permet de calculer la moyenne des distorsions minimales entre les n vecteurs X_j de la séquence d'apprentissage, et les vecteurs d'approximation correspondants à ce niveau d'itération de l'algorithme.

L'étape (4): Si $(D^{l-1} - D^l)/D^l \leq (\varepsilon = 0.0001)$, le dictionnaire C^l est conservé. La procédure s'arrête. Sinon, continuer.

Rechercher l'ensemble optimal des vecteurs d'approximation $\tilde{X}(S^l) = \tilde{X}(S_i)$,
($i = 1,\ldots, N_c$). Tel que $\tilde{X}(S_i)$ est définit par :

$$\boldsymbol{\tilde{X}(S_i) = \frac{1}{\|S_i\|}\sum_{j: X_j \in S_i} X_j} \qquad \textbf{(4.6)}$$

Cette formule représente le barycentre (centroide) de la partition S_i. $\| S_i \|$ est le nombre de vecteurs d'apprentissage X_j inclus dans la cellule S_i .L'utilisation de l'erreur quadratique moyenne pour le calcul des vecteurs $\tilde{X}(S_i)$ reviendrait à calculer la moyenne des vecteurs de la séquence d'apprentissage à l'intérieur des régions S_i.

Actualiser le dictionnaire $C^{l+1} = \tilde{X}(S^l)$, incrémenter l et aller à l'étape (2).

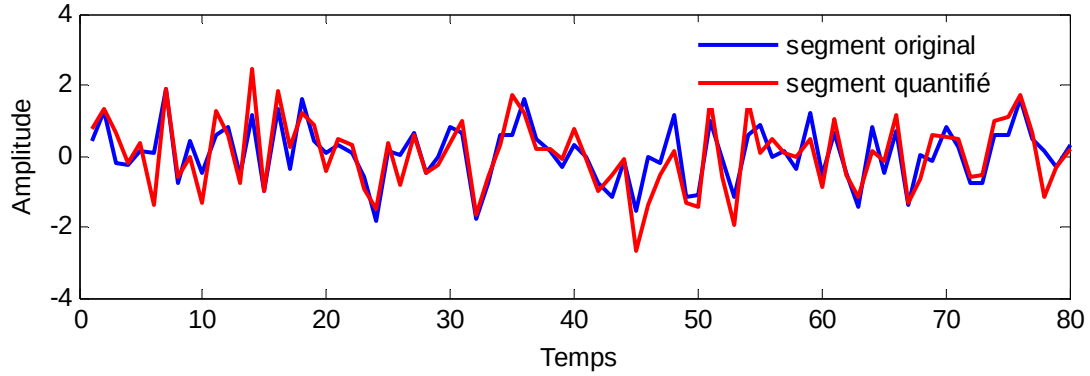


Fig. 4.7: Exemple d'une quantification vectorielle, pour une source aléatoire uniformément distribuée sur l'intervalle (0,1) de moyenne μ=0, et de variance σ^2=1. La séquence d'apprentissage contienne 15000 vecteurs, le dictionnaire contient 128 vecteurs, chaque vecteur est constitué de 8 échantillons.

4.2.2.4. Quantification vectorielle a 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'*.

4.3. Quantification conventionnelle des CW

Dans cette section, on va parler de la quantification et la dé-quantification conventionnelle des CW, La figure 4.8 donne les schémas blocs des deux processeurs ; les CW comme la puissance, nécessitent un traitement supplémentaire avant la quantification. Plus précisément, comme déjà vu au chapitre précèdent, chaque CW est décomposée en deux formes d'ondes (SEW et REW) qui seront quantifiées séparément, avec un traitement spécifique pour chaque composante.

4.3.1. Quantification des REW

Commençons tout d'abord par lister trois conclusions importantes tirées de [2, 34]:

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.

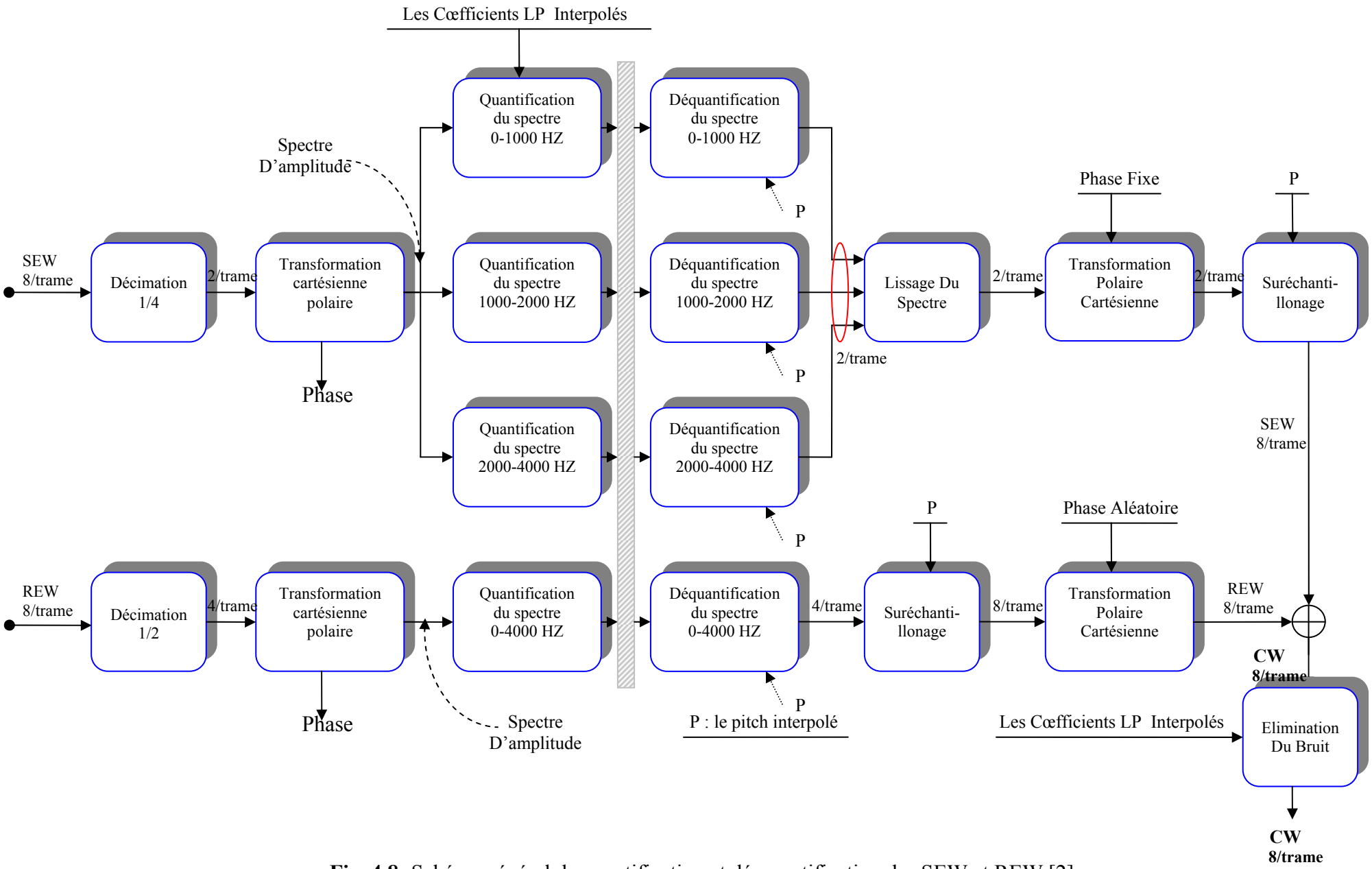


Fig. 4.8: Schéma général de quantification et dé-quantification des SEW et REW [2].

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 $[-\pi, \pi]$. Il est à noter que les spectres de phases aléatoires sont ajoutés aux REW à la fréquence des sous - trames.

4.3.2. Quantification des SEW

Puisque la fréquence de coupure du filtre de décomposition, est généralement comprise 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 (deux SEW par trame) 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. [35]

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.a.d. à 1000 Hz et à 2000 Hz), car 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) [2]. 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.

$$W(z)=\frac{1-\sum_{k=1}^{N} a_k \gamma_2{}^k z^{-k}}{1-\sum_{k=1}^{N} a_k \gamma_1{}^k z^{-k}} \qquad 0<\gamma_w \le 1 \qquad \textbf{(4.7)}$$

Où γ_1, γ_2 sont typiquement compris entre 0 et 1 ; pour le CELP on a γ_1=0.9, γ_2=0.5 [6].

4.4. Nouvelle technique de quantification des REW

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 conséquente.

Pour remédier ce problème, une technique simple et pratique, consiste à transformer les longueurs variables des REW en une seul 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 $\psi_i(\omega)$, avec une longueur L choisie [19, 21].

$$R(\omega)=\sum_{i=0}^{L-1} \gamma_i \psi_i(\omega) \qquad 0 \le \omega \le 2\pi \qquad \textbf{(4.8)}$$

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 [36]. Dans [37, 38] ils ont évalué cinq techniques de conversion et ont conclu que l'approche basée sur la transformé en cosinus discret DCT (Discrete Cosine Transform ; voir Annexe B) était capable de remplacer la quantification à dimension variable, avec la plus haute exactitude [39].

Il a été constaté que typiquement il y a des ressemblances entre les spectres adjacents des amplitudes des REW [36]. Cela suggère que cette corrélation peut être exploiter dans la quantification, et qu'une représentation simplifiée des REW pourrait être possible avec un taux de mise à jour approximativement 200 Hz. Cependant, puisque la propriété essentielle des REW est qu'elles évolue 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.

Dans cette section, nous présentons une méthode alternative, dont en exploitant la corrélation entre les spectres des REW consécutifs. La figure ci-dessous fournie un schéma détaillé de la couche quantification et dé-quantification du spectre d'amplitude des REW, basé sur la transformé DCT avec un taux de décimation des REW égale à 2.

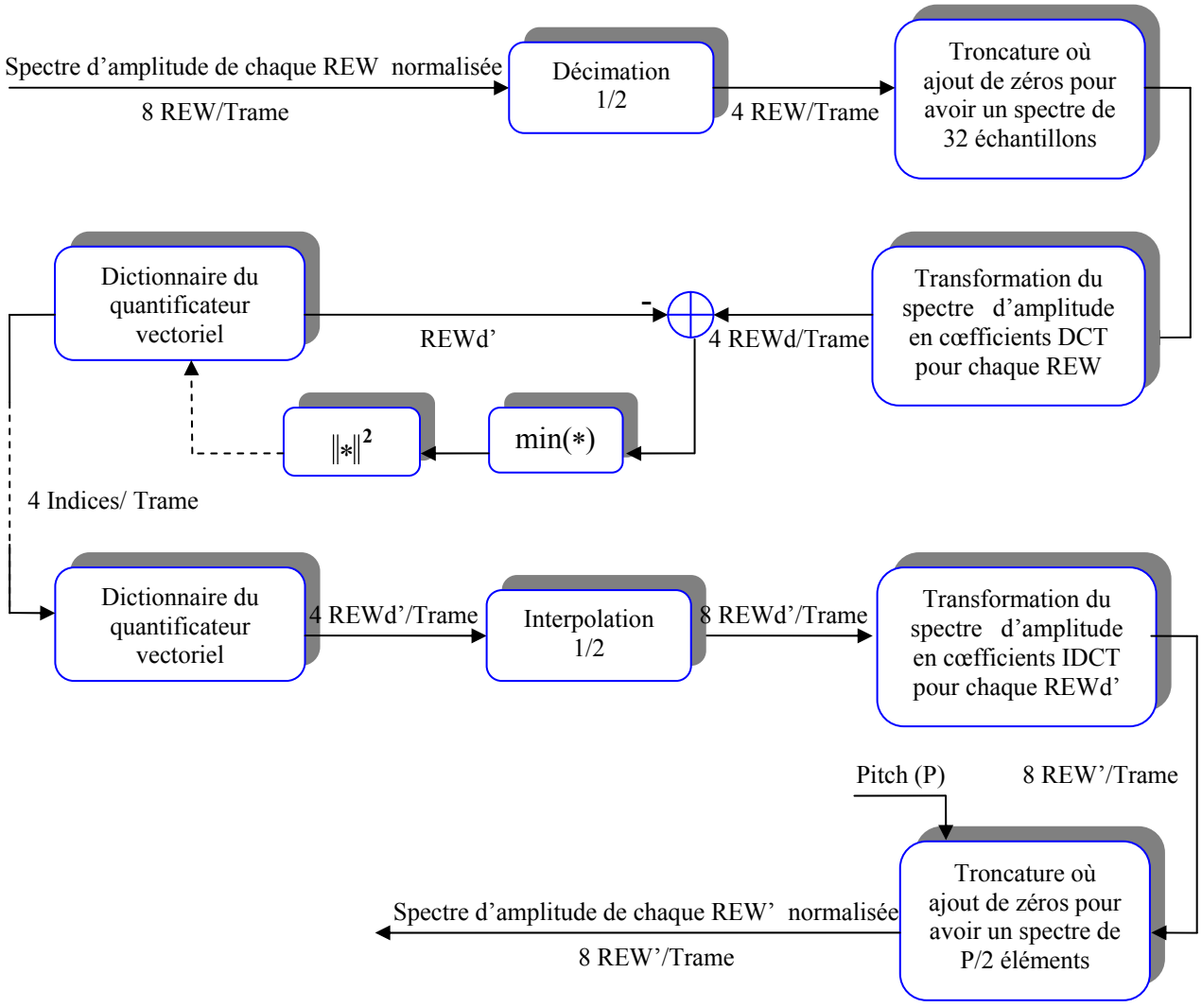


Fig. 4.9: Schéma bloc de la quantification et dé-quantification des REW.

4.4.1. Quantification des REW

Comme déjà citer au paragraphe précèdent, après la transformation du spectre d'amplitude des REW au domaine DCT avec une longueur choisi égale à 32, car d'après [39], choisir la dimension L=32 offre une meilleure performance, au-delà de cette valeur le SNR reste pratiquement stable ; alors, on applique la technique GLA conventionnelle pour la conception du dictionnaire avec une séquence d'apprentissage constituée de 20264 vecteurs DCT, le vecteur le plus proche est retrouvé à l'aide du critère MSE et son indice dans le dictionnaire est transmis au récepteur.

Pour bien approfondir notre étude, la quantification a été réalisée pour deux débits de quantification, ce qui nous conduit à faire la conception de deux dictionnaires :

-Dans le premier cas, chaque spectre DCT est codé par 5 bits, ce qui se traduit par un débit de 20 bits/trame, donc le dictionnaire est constitué de 32 vecteurs DCT.
-Dans le deuxième cas, chaque spectre DCT est codé par 4 bits, ce qui se traduit par un débit de 16 bits/trame, donc le dictionnaire est constitué de 16 vecteurs DCT.

Les figures 4.10 et 4.11 représentent les designs des deux dictionnaires avec la mesure de distorsion correspondante pour chaque dictionnaire, lors de la phase d'apprentissage (training).

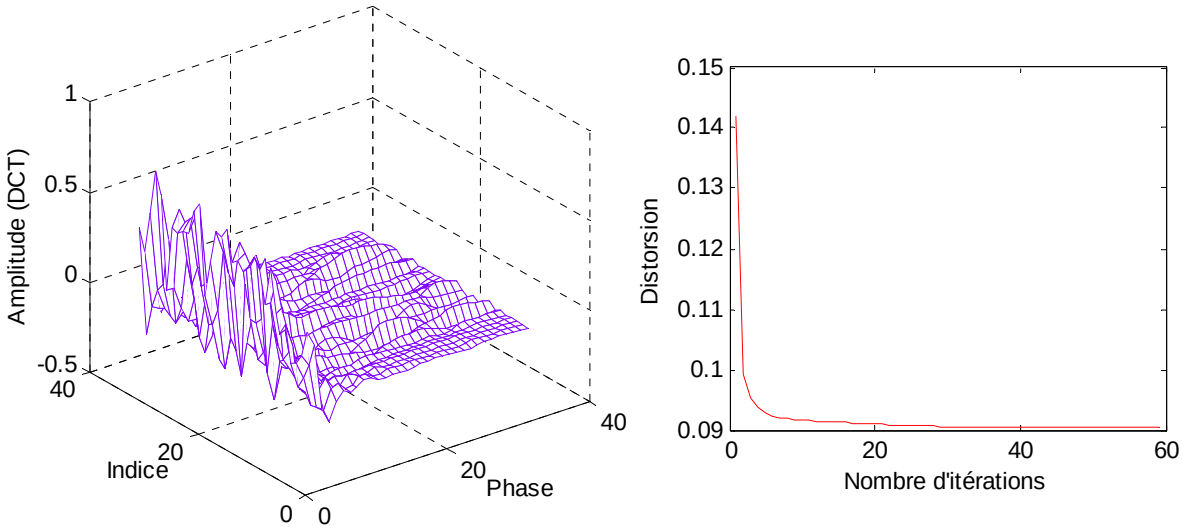


Fig. 4.10: Design du dictionnaire et mesure de distorsion associée, pour un codage de 20 bits/trame.

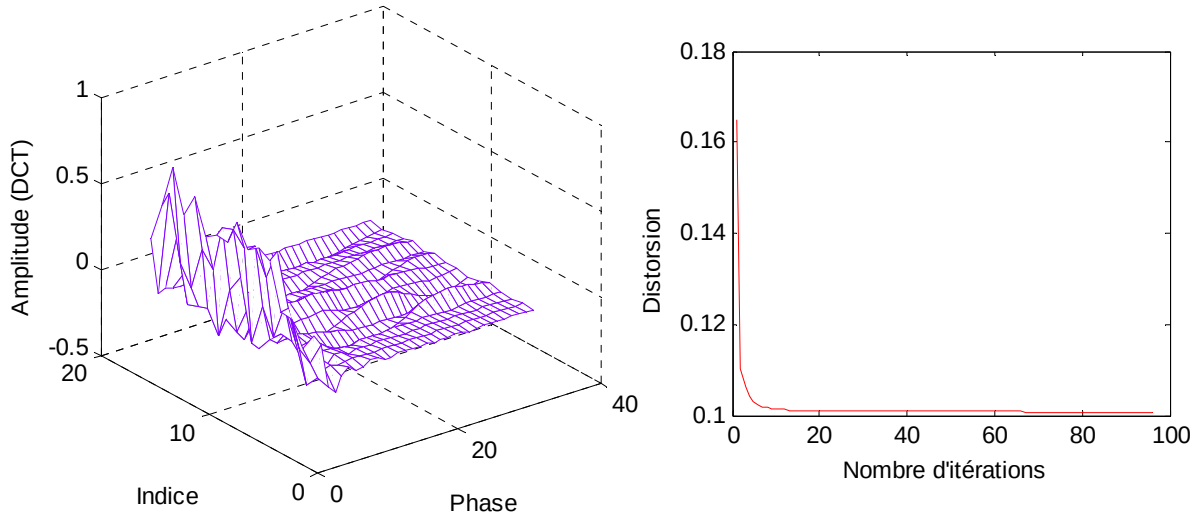


Fig. 4.11: Design du dictionnaire et mesure de distorsion associée, pour un codage de 16 bits/trame.

Le critère d'évaluation que nous avons adopté est la mesure du rapport signal sur bruit, pour les REW originales et quantifiées. Le tableau ci-dessous représente les moyennes des résultats d'évaluation de 12 phrases, dont 6 voix masculines, et 6 voix féminines.

Débit de quantification	**SNR (voix masculines) dB**	**SNR (voix féminines) dB**	**La moyenne dB**
REW quantifier sur 20 bits/Trame	6.7328	7.2595	6.9962
REW quantifier sur 16 bits/Trame	6.3854	6.8724	6.6289

Tableau. 4.1: L'évaluation objective pour les deux débits de quantification.

Suite aux résultats d'évaluation décrient dans le tableau précèdent, les rapports signal sur bruits pour les deux débits de quantification sont pratiquement les mêmes, avec une légère variation. Également, les tests subjectifs ont montré que la parole reconstituée contient une faible rugosité mais qui n'est pas nuisible.

La figure suivante montre la représentation temporelle de quatre REW successives avec leurs versions quantifiées.

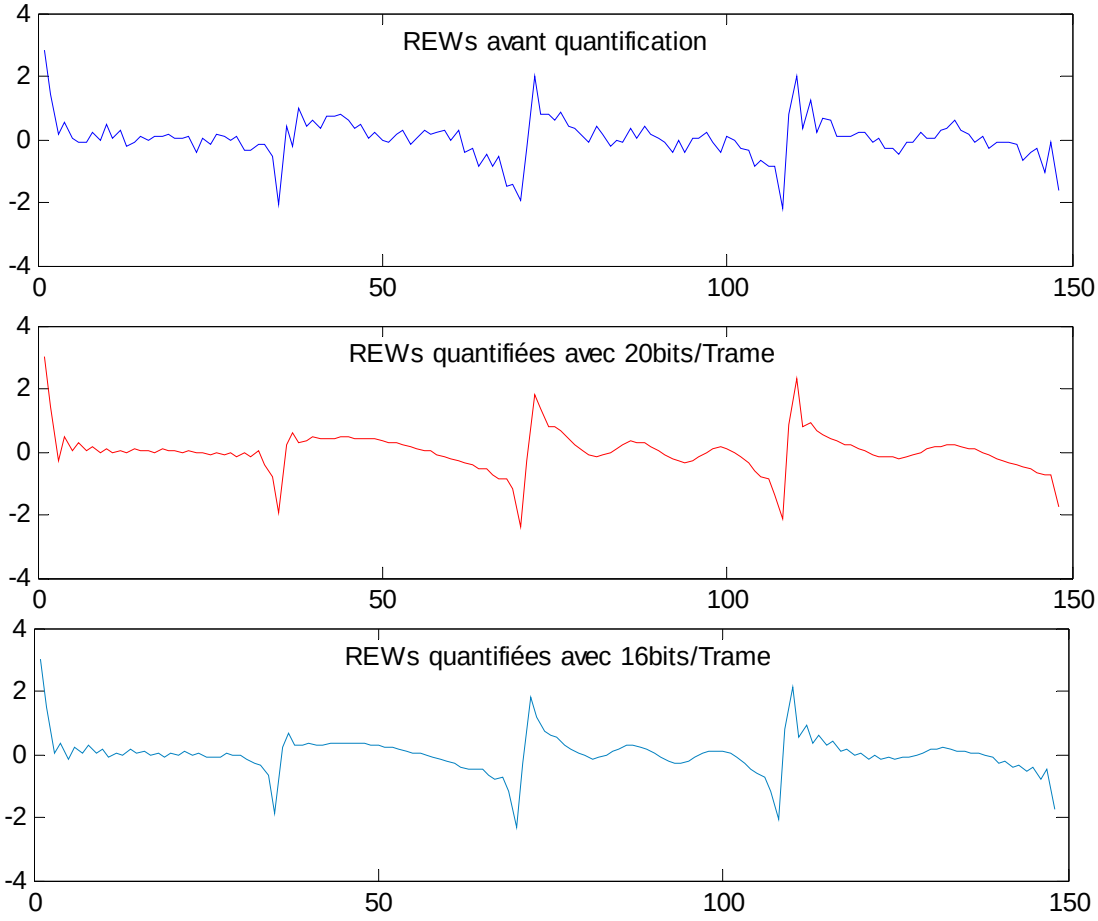


Fig. 4.12: Un exemple de quantification pour 4 REW successives.

D'après la figure précédente, la version quantifiée des REW fournie une structure plus fine que la REW originale, qui est due principalement aux raisons suivante :

-Faible débit de quantification ; c-à-d un dictionnaire qui contient un minime nombre de formes, par conséquent, on aura une distorsion importante par rapport au plus proche voisin.
-La structure d'une REW est a forme aléatoire, donc l'opération de quantification donne une approximation ou un ajustement de la forme originale

4.4.2. Ajustement des dictionnaires

Dans un système de traitement de signal tel que les codeurs de parole, la performance du codeur ne se caractérise pas seulement par la qualité de la parole reconstruite ou par la rapidité du codeur, mais aussi par l'espace mémoire occupé par celui-ci. L'objective de cette partie 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 (voir Annexe C), ainsi un signal de L échantillons peut être représenter par N échantillons tel que $N<L$, et N représente l'ordre du polynôme ajusté. La figure 4.13 montre l'influence de l'ordre d'ajustement sur une forme d'onde extraite des dictionnaires précédents.

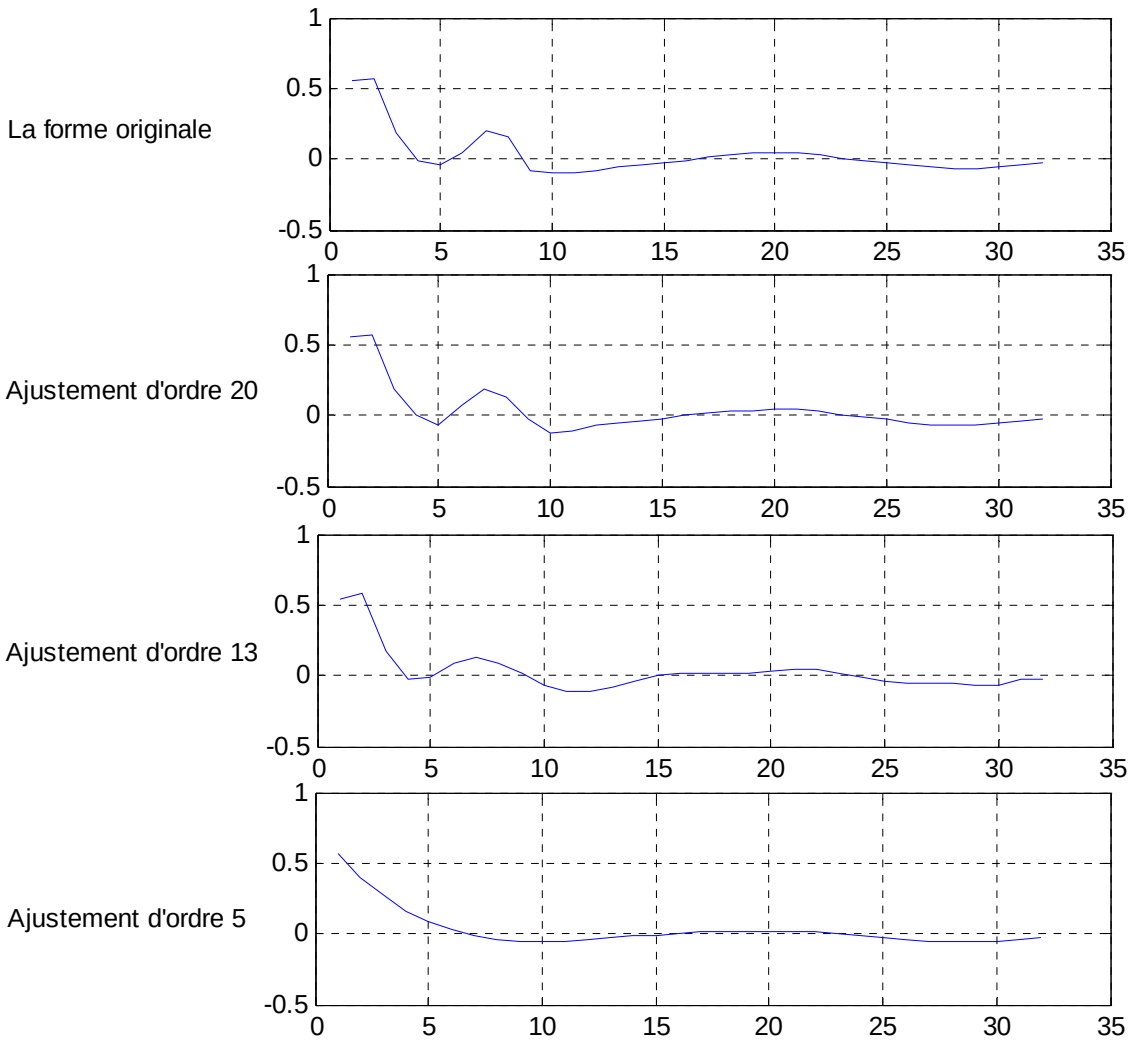


Fig. 4.13: Influence de l'ordre d'ajustement sur la forme originale.

Dans la nouvelle représentation ajustée du dictionnaire, le problème posé est, quelle serait la forme ajustée adéquate, qui donnerait une représentation fidèle de la forme originale au sens perceptuel avec un plus faible degré d'ajustement, ce qui nous mène à faire plusieurs essais, pour finalement obtenir le degré d'ajustement optimal. Les tableaux et les figures suivants représentent les SNR moyens pour chaque degré d'ajustement.

Ordre d'ajustement	**SNR (voix masculines) dB**	**SNR (voix féminines) dB**	**La moyenne dB**
REW ajustées à l'ordre 20	6.7340	7.2507	6.9924
REW ajustées à l'ordre 17	6.7344	7.2150	6.9747
REW ajustées à l'ordre 13	6.7429	7.0725	6.9077
REW ajustées à l'ordre 10	6.6875	6.9348	6.8111
REW ajustées à l'ordre 7	6.5270	6.8307	6.6788
REW ajustées à l'ordre 5	5.5035	5.9520	5.7278

Tableau. 4.2: REW quantifiées sur 20 bits/trame, pour différents degrés d'ajustement.

Ordre d'ajustement	**SNR (voix masculines) dB**	**SNR (voix féminines) dB**	**La moyenne dB**
REW ajustées à l'ordre 20	6.3896	6.8725	6.6311
REW ajustées à l'ordre 17	6.3921	6.8477	6.6199
REW ajustées à l'ordre 13	6.4012	6.7244	6.5628
REW ajustées à l'ordre 10	6.3879	6.6142	6.5011
REW ajustées à l'ordre 7	6.2743	6.5463	6.4103
REW ajustées à l'ordre 5	5.3376	5.7709	5.5542

Tableau. 4.3: REW quantifiées sur 16 bits/trame, pour différents degrés d'ajustement.

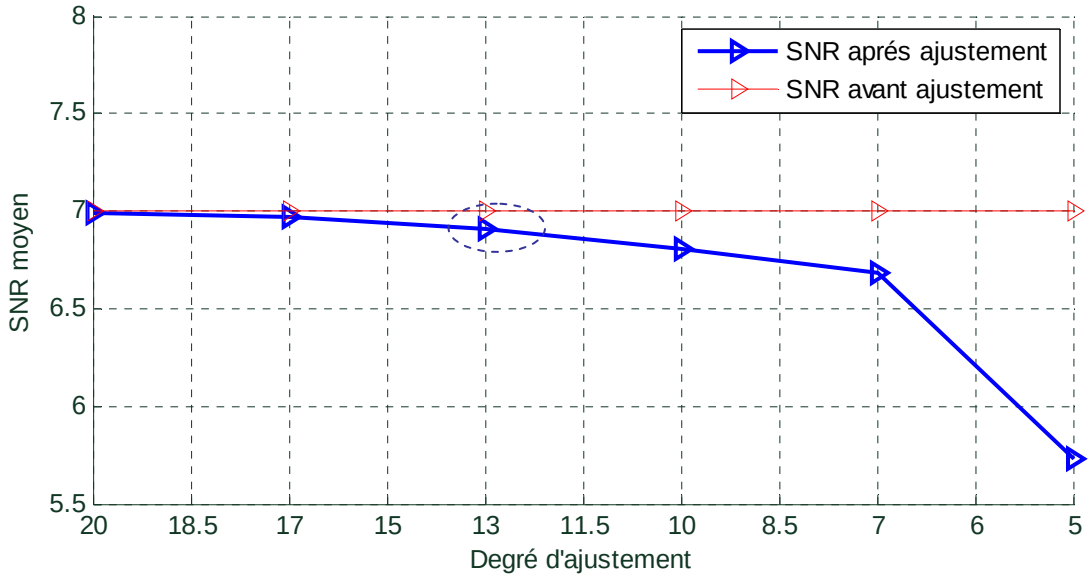


Fig. 4.14: Représentation graphique du SNR en fonction du degré d'ajustement dans le cas d'une quantification de 20 bits/trame.

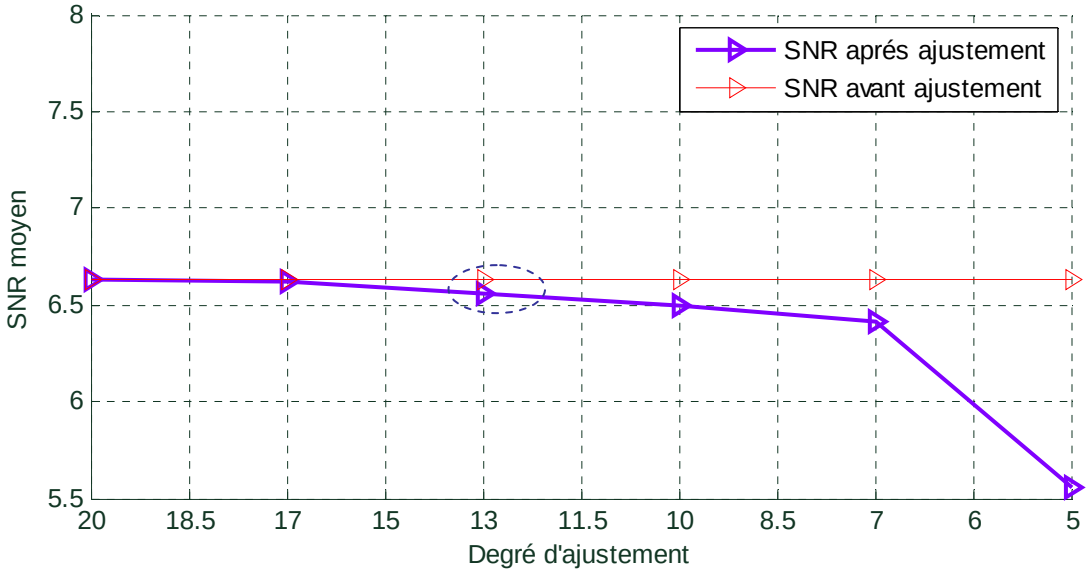


Fig. 4.15: Représentation graphique du SNR en fonction du degré d'ajustement dans le cas d'une quantification de 16 bits/trame.

Suite aux résultats précédents, on remarque que le SNR après ajustement commence a s'éloigné du SNR avant ajustement a partir de la valeur 13, donc le degré d'ajustement adéquat égal à 13, c-à-d chaque forme du dictionnaire de 32 échantillons peut être remplacée par 14 échantillons (nombre de cœfficients = degré d'ajustement+1) pour les deux cas de quantification, comme il est montré dans les figures suivantes.

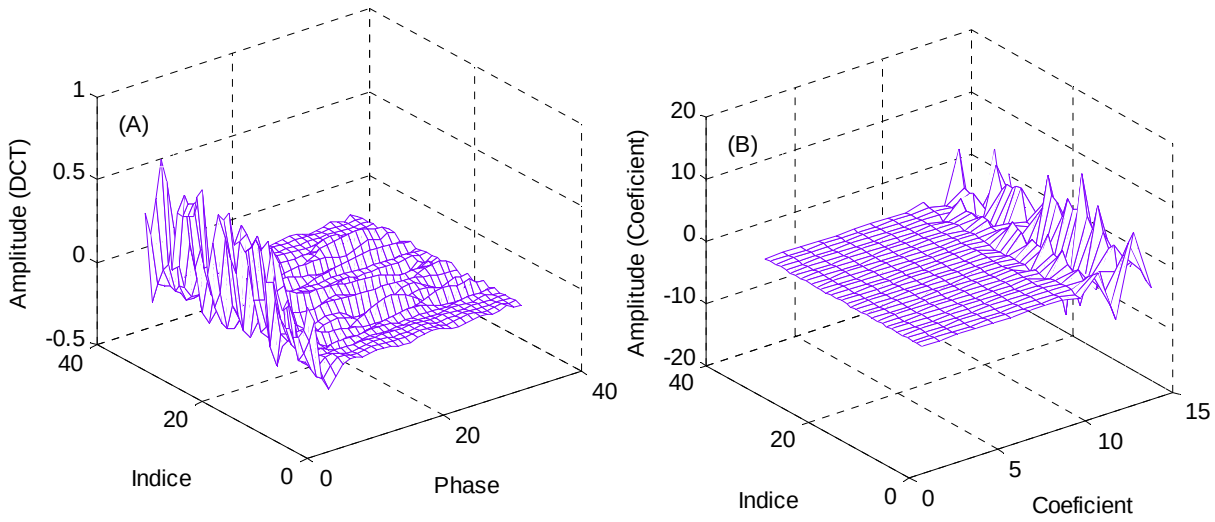


Fig. 4.16: Forme du dictionnaire de 32 éléments après ajustement d'ordre 13.
(A)Représentation des cœfficients DCT.
(B) Représentation des cœfficients polynomiaux.

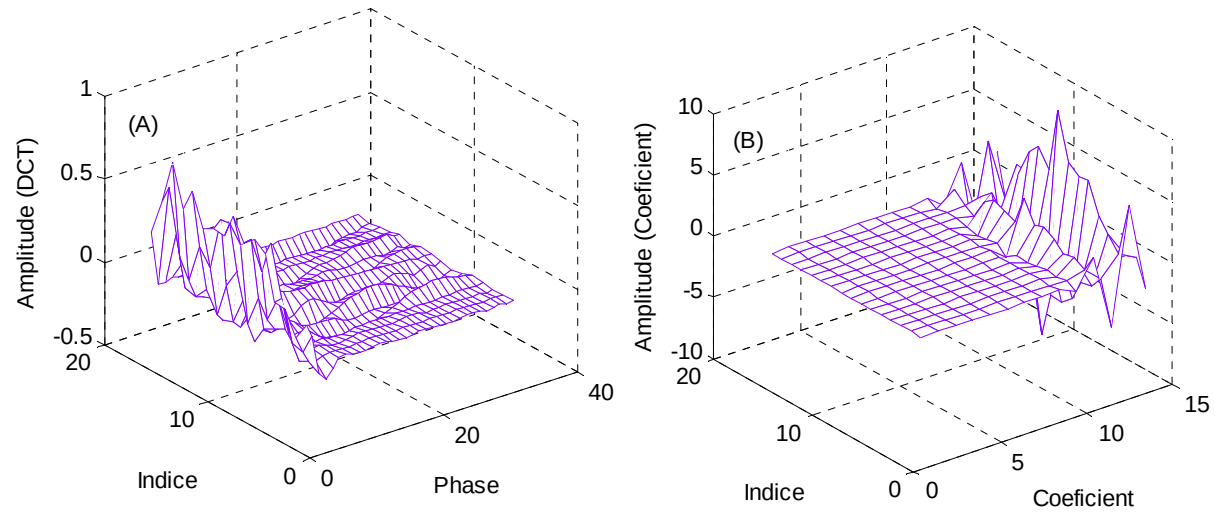


Fig. 4.17: Forme du dictionnaire de 16 éléments après ajustement d'ordre 13.
(A)Représentation des cœfficients DCT.
(B) Représentation des cœfficients polynomiaux.

4.4.3. Effet du rapport de décimation sur les REW reconstituées

Dans la simulation précédente avec un taux de décimation qui vaut deux, le débit de compression des REW était acceptable ; même avec un débit de 20 bits/trame, la conception d'un codeur WI autours de 4 Kbits/s est concevable. Dans ce qui suit, nous essayant de réduire le débit de quantification en modifiant le degré de décimation, les figures ci-dessous montrent l'effet de l'interpolation des cœfficients DCT après l'exécution d'une décimation d'ordre 4 comparé au résultat d'une décimation d'ordre 2 (le cas précèdent), pour 4 REW dans une trame d'analyse.

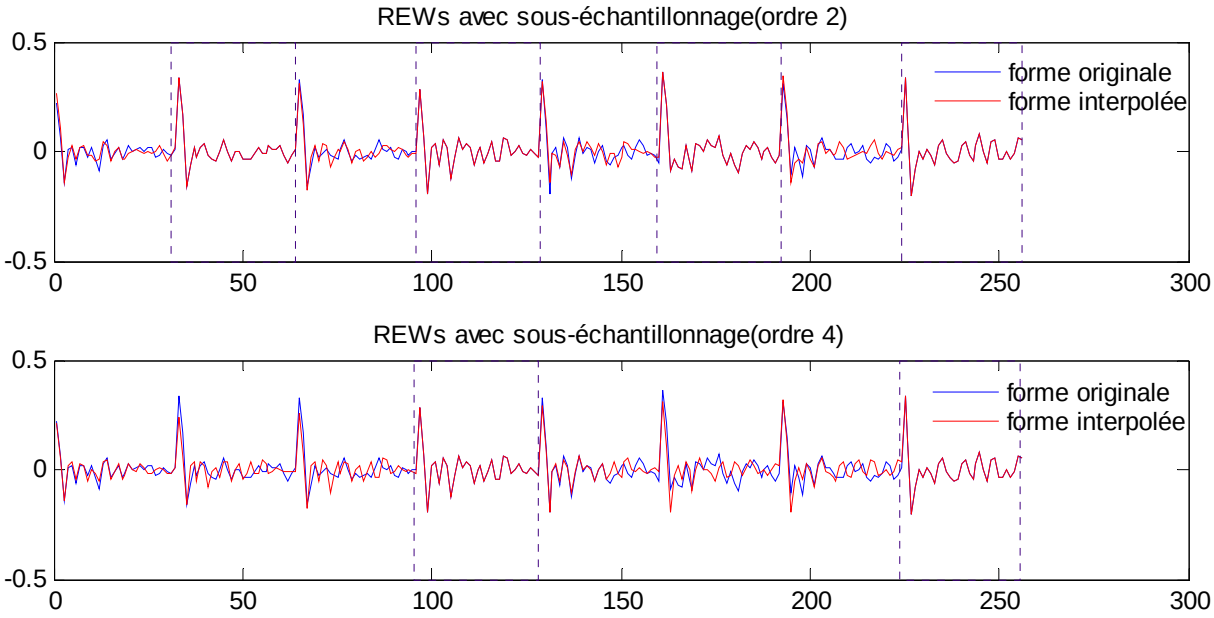


Fig. 4.18: Effet de l'interpolation sur les cœfficients DCT

D'après des essais effectués sur des REW sous échantillonnées avec un rapport de quatre, la valeur du SNR a subit une faible dégradation (voir les tableaux 4.4 et 4.5) ; mais pour le point de vue perception, la dégradation du signal reconstruit est plus importante que celle du cas précèdent, c.-à-d. on n'aura pas un codeur de qualité communication (Toll Quality), Ainsi ces résultats vérifies les conclusions tirées de [34, 36].

Ordre d'ajustement	SNR (voix masculines) dB	SNR (voix féminines) dB	La moyenne dB
Sans Ajustement	6.7648	7.1520	6.9584
Ajustement d'Ordre 13	6.4127	6.9755	6.6941

Tableau. 4.4: L'évaluation objective pour une quantification sur 10 bits/trame.

Ordre d'ajustement	SNR (voix masculines) dB	SNR (voix féminines) dB	La moyenne dB
Sans Ajustement	6.3977	6.7376	6.5677
Ajustement d'Ordre 13	6.3708	6.6035	6.4871

Tableau. 4.5: L'évaluation objective pour une quantification sur 8 bits/trame.

4.4.4. Evaluation de la performance

Dans cette section, nous donnons une présentation de l'allocation de bits pour un codeur WI travaillant à 3.85 Kbps à travers le tableau 4.6 ; notons que les paramètres sont quantifiés selon la description présentée dans ce chapitre avec le débit de quantification des SEW est extrait à partir de [39].

Paramètre	Bits/trame	Bits/seconde
Cœfficients 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

Tableau. 4.6: Allocation de bits d'un codeur WI à 3.85 Kbps.

4.5. Conclusion

La qualité de codage d'un codeur WI est reliée par l'efficacité de la quantification des formes d'onde CW ; la technique de décomposions de ces dernières en une forme a évolution rapide REW et une forme a évolution lente SEW a fait beaucoup de progrès pour améliorer la qualité de la parole reconstituée. Dans ce dernier chapitre on a choisi de développer une méthode de la quantification d'une de ces composante qui est la REW ; cette méthode repose sur des conclusions extraites à partir des travaux précédents, la transformée DCT qui a la propriété de répartir l'énergie du signal sur une bande de fréquences plus basse donc on utilise moins de cœfficients pour représenter un signal et la technique de représentation polynomiale qui permet de diminuer la taille du dictionnaire en représentant chaque spectre à coder par les cœfficients d'un polynôme, par conséquent à réduire l'espace mémoire.

La quantification a été réalisée pour des débits de 20 bits/trame et 16 bits/trame. Après la dequantification des REW et la reconstitution du signal de parole, la comparaison des tests d'écoute avec les résultats des tests précédents (couche analyse synthèse, sans quantification), a montrée que ces derniers étaient presque les mêmes avec une légère variation ; tendit que les mesures objectives sur les REW dequantifiées étaient du même ordre que celles trouvées dans [39], on a également remarqué que les résultats étaient pratiquement similaire pour les deux débits de quantification pour les points de vue objectif et subjectif.

Comme dernière tentative on a essayé de diminuer le débit de quantification des REW à des degrés considérablement réduits, en doublant le degré de décimation des REW pour les deux débits de quantification ; les testes d'écoutes ont montré l'existence de grandes détériorations qui sont audibles, donc les conclusions tirées à partir de [34, 36] ont été vérifiées (voir paragraphe 4.3.1).

Conclusion Générale

Le but initial de ce travail était de mettre en œuvre un algorithme de codage de type WI. Ce qui consistait à décomposer le signal de parole et le reconstituer. Ce qui nécessite l'apprentissage de méthodes de traitement spécifiques au signal parole. L'objectif principal a été atteint. Nous avons été au-delà pour nous initier à la compression appropriée pour ce type de traitement, qui impose un certain niveau de connaissances en traitement numérique du signal et aux différentes méthodes d'analyse mathématique.

Nous avons travaillé sous un environnement Matlab. Dont l'étape analyse-synthèse ou la décomposition-reconstitution du signal WI a été vérifiée par des tests d'évaluation subjective. Les performances ont été jugées satisfaisantes. En effet, la qualité de la parole reconstruite a été jugée bonne en moyenne pour plusieurs phrases (voix masculines et féminines). Également, Ce travail de recherche a été pour moi une occasion pour me familiariser à apprendre des méthodes et plusieurs aspects sur diverses techniques de codage d'un signal de la parole.

Comme déjà vu dans le dernier chapitre, des essais sur la quantification vectorielle et sur les différentes techniques de conception des dictionnaires de quantification ont été faits. Ainsi une méthode de quantification de la composante REW est développée. Cette solution est le résultat d'un mixage de diverses méthodes de calculs extraites à partir de plusieurs tentatives des codeurs WI. En profitant de la caractéristique perceptuelle de cette composante, le principal objectif est d'atteindre un débit plus faible, sous entendus en respectant les limites de perception humaine, un nombre d'opérations plus réduit et un espace mémoire restreint. Finalement, on a essayé de diminuer le débit de quantification, en augmentant le degré de sur-échantillonnage, mais les résultats ne sont pas vraiment adéquats aux normes de la bonne qualité de communication.

Notons enfin que le travail réalisé dans le domaine du codage de la parole dans un codeur WI au cours de ce mémoire, peut être une aide pour tous ceux qui veulent approfondir leurs idées dans la quantification des composantes du codeur ; soit améliorer la méthode proposée, ou même présenter d'autre méthodes qui quantifient la composante lente SEW, ainsi on aura un schéma complet de la quantification des forme d'ondes CW, donc on pourra avoir une évaluation réelle de la qualité de la parole reconstruite. On croit aussi que, employer la WI dans les signaux à large bande (wide band), peut conduire à des résultats performants.

Références Bibliographiques

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

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

[3] 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.

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

[5] Guillaume Madre, Application de la Transformée en Nombres Entiers à l'Etude et au Développement d'un Codeur de Parole, Université de Bretagne Occidentale, Octobre 2004.

[6] G. Baudoin, Codage de la Parole à Bas et Très Bas Débit, Mémoire d'Habilitation à Diriger des Recherches, Université de Marne la Vallée, Novembre 2000.

[7] M.R. Schroeder, B.S. Atal, Code-Excited Linear Prediction (CELP): High Quality Speech at Very Low Bit Rates, Proceedings IEEE ICASSP-85, pp. 937-940, Tamp, 1985.

[8] 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.

[9] K.K.Paliwal and B.S.Atal, Efficient Vector Quantization of LPC Parameters at 24 Bits/ Frame. IEEE Transactions on Speech and Audio Processing, Vol 1. No 1, January 1993.

[10] 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.

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

[12] I. Tasmanna, Interpolation of Linear Prediction Coefficients for Speech Coding Department of Electrical Engineering McGill University Montreal, Canada April, 2000.

[13] B. Razvan, Mesure de la Qualité dans les Réseaux Informatiques Université Jean Monnet St. Etienne Thèse pour obtenir le grade de Docteur, juillet 2004.

[14] P. Noll Speech, and Audio Coding for Multimedia Communications, Proceeding International Cost 254 Workshops on Intelligent Communication Technologies and Applications University, Berlin, 1999.

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

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

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

[18] 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.

[19] 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.

[20] T. Ericsson and W.Bastiaan Kleijn, on Waveform-Interpolation Coding with Asymptotically Perfect Reconstruction, Department of Speech, Music and Hearing, (Royal Institute of Technology) 100 44 Stockholm Sweeden, pp 93- 95 IEEE, 1999.

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

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

[23] Selectable Mode Vocoder (SMV) Service Option for Wideband Spread Spectrum Communication Systems, January 2004.

[24] 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.

[25] 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.

[26] G. Blanchet, M. Charbit, Traitement Numérique du Signal, simulation sous Matlab.

[27] M. A. Khan, Coding of Excitation Signals in a Waveform Interpolation Speech Coder; Department of Electrical & Computer Engineering McGill University Montreal, Canada July 2001.

[28] 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.

[29] H. Khalil K. Rose, MSVQ Design for Packet Networks with Application to LSF Quantisation, signal compression laboratory, University of California Santa Barbara CA 93106, USA.

[30] O. Gottesman and A. Gersho, Enhanced Waveform Interpolative Coding at 4 Kbps Signal Compression Laboratory, Department of Electrical and Computer Engineering, University of California, Santa Barbara, California 93106, USA.

[31] J.Max, Quantization for Minimum Distortion, IRE Transactions on Information Theory IT-6, 7–12, 1960.

[32] Y.Linde, A.Buzo, and R.M.Gray, an Algorithm for Vector Quantizer Design, IEEE Transactions on Communications 28, no.1, pp 84-95, 1980.

[33] A. Das, A. V. Rao, and A. Gersho, Variable-Dimension Vector Quantisation, IEEE Signal processing letters, vol 3, pp 200-202, July 1996.

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

[35] J. Thyssen, W. Bastiaan Kleijin, R. Haagen, Using a Perception-Based Frequency Scale in Waveform Interpolation information principles research laboratory, AT & T Laboratories, Murray Hill. NJ 07974, USA, IEEE, 1997.

[36] 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.

[37] 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.

[38] DFT, DCT, MDCT, DST and Signal Fourier Spectrum Analysis ; L. Yaroslavsky, Department of Interdisciplinary Studies, Faculty of Engineering, Israel ; and Ye Wang, Nokia Research Center, Tampere, Finland.

[39] 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, Tampere, Finland.

[40] N. Ahmed, T. Natarajan, and K.R. Rao, Discrete Cosine Transform, IEEE Transactions on Computer, vol. C-23, pp 90-94, January 1974.

[41] X. Huo, Sparse Image Representation Via Combined Transforms, Stanford University, August 1999.

[42] Numerical Methods Lecture 5 - Curve Fitting Techniques CGN 3421 - Computers Methods pp 89-102.

Annexe A :

Représentation Bidimensionnelle des CW

Comme 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 recherchées doit permettre d'avoir un signal bidimensionnel.

Tout d'abord, 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 C W de longueur P (période du pitch) :

$$s(m) \in R \qquad m = 0,1,\cdots,P-1 \qquad \textbf{(A.1)}$$

On connaît que toute fonction à bande limitée[1] peut être décomposée en série de Fourier à temps discret DTFS, c'est `a dire sous forme de série de sinus et cosinus ; on écrit alors, pour toute fonction $s(m)$:

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

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 :

[1] Dans notre cas on traite un signal parole réel et à bande limitée B, tel que B= [300, 3400 Hz]

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

$$\left.\begin{aligned} A_k &= \frac{1}{P}\sum_{m=0}^{P-1}\left[s(m)\cos\left(\frac{2\pi km}{P}\right)\right] \\ B_k &= \frac{1}{P}\sum_{m=0}^{P-1}\left[s(m)\sin\left(\frac{2\pi km}{P}\right)\right] \end{aligned}\right\} \quad \textit{pour} \quad k = 0 \quad \textit{et} \quad P/2 \qquad \textbf{(A.3)}$$

Quand *P* est impair :

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

$$\left.\begin{aligned} A_k &= \frac{1}{P}\sum_{m=0}^{P-1}\left[s(m)\cos\left(\frac{2\pi km}{P}\right)\right] \\ B_k &= \frac{1}{P}\sum_{m=0}^{P-1}\left[s(m)\sin\left(\frac{2\pi km}{P}\right)\right] \end{aligned}\right\} \quad \textit{pour} \quad k = 0 \qquad \textbf{(A.4)}$$

La forme d'une CW peut maintenant, être décrite par un ensemble de coefficients DTFS {*Ak*, B_k}. Notons que l'indice *m* dans (A.2) n'est pas nécessairement entier ; il peut prendre n'importe quelle valeur réelle dans l'intervalle $0 \leq m < P$. En d'autres termes, les valeurs situées entre deux instants discrets peuvent être calculées aisément par (A.2).

Après avoir obtenu la représentation pour une CW, nous sommes maintenant prêts à construire une représentation bidimensionnelle pour une séquence de CW. En fait, cette représentation est simplement obtenue en ajoutant une modification à (A.2). Ainsi, on attache un indice de temps discret *n* à tous les paramètres dans (A.2) qui varient dans le temps, ces paramètres sont {A_k}, {B_k} et *P*. L'équation (A.2) peut donc être écrite comme suit :

$$s(n,m)=\sum_{k=0}^{[P(n)/2]}\left[A_k(n)\cos\left(\frac{2\pi km}{P(n)}\right)+B_k(n)\sin\left(\frac{2\pi km}{P(n)}\right)\right] \qquad 0\le m<P(n) \quad \textbf{(A.5)}$$

Où les coefficients $\{A_k\}$ et $\{B_k\}$ sont maintenant variants dans le temps, de même que la valeur du pitch $P(n)$. Il faut noter que le coefficient A_0 représente la moyenne ou la composante continue du signal et le coefficient B_0 est toujours redondant a valeur nulle (sin(0) = 0).

L'équation (A.5) 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 l'ensemble des CW évolue à travers le temps discret le long de l'axe n.

Cependant, la longueur de la CW dans (A.5) dépend de la valeur 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 [18]. Cette normalisation peut être accomplie en substituant

$$\phi=\phi(m)=\frac{2\pi.m}{P(n)} \qquad \textbf{(A.6)}$$

dans (A.5), ainsi on obtient :

$$s(n,\phi)=\sum_{k=0}^{[P(n)/2]}\left[A_k(n)\cos(k\phi)+B_k(n)\sin(k\phi)\right] \qquad 0\le\phi(\cdot)<2\pi \qquad \textbf{(A.7)}$$

Annexe B :

Transformée de Cosinus Discrète

La DCT (Discrete Cosine Transform) est une transformée de Fourier dont on a remplacé les sinus par des cosinus [40]. On peut interpréter la DCT, comme une transformation prenant en entrée un signal de base a une dimension $x(n)$ de longueur N (pour une DCT unidimensionnelle) ; la DCT décompose $x(n)$ en un signal $y(k)$ de longueur K . C'est une transformée très utilisée en traitement de signal, particulièrement en traitement d'image. Elle est définit par :

$$y(k)=w(k)\sum_{n=0}^{N-1}x(n).\cos\left(\frac{\pi.k(2n+1)}{2N}\right) \qquad \textbf{(B.1)}$$

Où

$$w(k)=\begin{cases}\dfrac{1}{\sqrt{N}} & si \quad k=0\\ \sqrt{\dfrac{2}{N}} & si\ non\end{cases} \qquad \textbf{(B.2)}$$

Sa transformée inverse définit par :

$$x(n)=\sum_{k=0}^{N-1}w(k).y(k).\cos\left(\frac{\pi.k(2n+1)}{2N}\right) \qquad \textbf{(B.3)}$$

Avantages de la DCT sur la FFT

En effet la formule de la DFT (Discrete Fourier Transform) est assez proche de celle de la DCT [41] :

$$y(k)=\sum_{n=0}^{N-1} x(n).\exp(-j2\pi kn/N) \qquad \textbf{(B.4)}$$

Ainsi en appliquant l'algorithme rapide de la DFT ou la FFT (Fast Fourier Transform), on obtient deux vecteurs, un vecteur de la partie réelle et un vecteur de la partie imaginaire. On peut penser que la FFT nous a doublé la taille des données. En fait la taille des données n'est pas doublée car il y a des redondances prévisibles. Ainsi cela ne représente qu'un faible avantage. Bien qu'étant plus longue à calculer que la FFT, la DCT comporte certains avantages. En effet, la DCT n'utilise pas de coefficients complexes, ceci rend la programmation légèrement plus simple.

Mais il y a un avantage bien plus grand qui est le fait que la DCT repartit l'énergie du signal sur une bande de fréquences plus basses que la FFT et donc il y a moins de coefficients importants que pour la transformée de Fourier. De ce fait, si on néglige les petits coefficients dans la FFT, on aura une perte très grande. C'est pour cela que l'on utilise la DCT.

Cet avantage de la DCT est due au fait qu'avec la DCT, on observe le signal sur une période 2**N* contre *N* pour la FFT. (Voir B.1 et B.4) ; Ainsi comme le calcul se fait sur une période 2**N*, alors l'énergie du signal se trouve dans des fréquences plus basses, et donc on a besoin de moins de coefficients.

D'autre part la DCT présente pour les effets de bord de meilleures performances que la DFT.

Annexe C :

Approximation Polynomiale

Une quantité mesurée est souvent reliée à une variable, par exemple $y = f(x)$. Cette fonction peut être de n'importe qu'elle forme, linéaire, quadratique, harmonique, ou arbitraire. Les courbes polynomiales ajustées (Polynomial Curve Fitting), essayent de trouver des modèles mathématiques d'un ensemble de données. En effet c'est une procédure dans laquelle le problème de base, est de traverser une courbe ou un ensemble de points, qui représentent des données expérimentales. Le plus intéressant est l'ajustement des données par une droite, où ces dernières sont supposées relier linéairement par la formule suivante :

$$y = ax + b \qquad \textbf{(C.1)}$$

Une des méthodes les plus généralement utilisées dans l'ajustement, est celle des moindres carrés [42]. Dans cette méthode les paramètres sont choisis d'une façon à minimiser la somme des carrés des déviations des variables dépendantes (y_i) (l'incertitude en x_i est supposée négligeable) à partir de la courbe $f(x_i)$.

$$S = \sum_{i=1}^{n} [y_i - f(x_i)]^2 \qquad \textbf{(C.2)}$$

Calcule des cœfficients du polynôme par l'approche des moindres carrés

Considérons la forme générale d'un polynôme d'ordre j :

$$f(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots + a_j x^j = a_0 + \sum_{k=1}^{j} a_k x^k \qquad \textbf{(C.3)}$$

D'après (C.2), l'expression générale de l'erreur est donnée par :

$$\boldsymbol{erreur} = \sum (\boldsymbol{d}_i)^2 = (\boldsymbol{y}_1 - \boldsymbol{f}(\boldsymbol{x}_1))^2 + (\boldsymbol{y}_2 - \boldsymbol{f}(\boldsymbol{x}_2))^2 + (\boldsymbol{y}_3 - \boldsymbol{f}(\boldsymbol{x}_3))^2 + \cdots + (\boldsymbol{y}_i - \boldsymbol{f}(\boldsymbol{x}_i))^2 \quad \textbf{(C.4)}$$

Substituons (C.3) dans (C.4) :

$$\boldsymbol{erreur} = \sum_{i=1}^{n} \left(\boldsymbol{y}_i - \left(\boldsymbol{a}_0 + \boldsymbol{a}_1 \boldsymbol{x}_i + \boldsymbol{a}_2 \boldsymbol{x}_i^{\,2} + \boldsymbol{a}_3 \boldsymbol{x}_i^{\,3} + \cdots + \boldsymbol{a}_j \boldsymbol{x}_i^{\,j} \right) \right)^2 \quad \textbf{(C.5)}$$

$$\boldsymbol{erreur} = \sum_{i=1}^{n} \left(\boldsymbol{y}_i - \left(\boldsymbol{a}_0 + \sum_{k=1}^{j} \boldsymbol{a}_k \boldsymbol{x}^k \right) \right)^2 \quad \textbf{(C.6)}$$

Où *n* représente le nombre de points ou de données, et *j* est l'ordre du polynôme. Maintenant l'estimation de l'ensemble des coefficients a_0 ..... a_k nous mène à minimiser l'équation de l'erreur, cette minimisation consiste à annuler la dérivée partielle pour chaque variable a_0 ..... a_k , k=1,..., j de l'équation (C.6).

$$\begin{aligned}
\frac{\partial \boldsymbol{erreur}}{\partial \boldsymbol{a}_0} &= -2 \sum_{i=1}^{n} \left(\boldsymbol{y}_i - \left(\boldsymbol{a}_0 + \sum_{k=1}^{j} \boldsymbol{a}_k \boldsymbol{x}^k \right) \right) = \boldsymbol{0} \\
\frac{\partial \boldsymbol{erreur}}{\partial \boldsymbol{a}_1} &= -2 \sum_{i=1}^{n} \left(\boldsymbol{y}_i - \left(\boldsymbol{a}_0 + \sum_{k=1}^{j} \boldsymbol{a}_k \boldsymbol{x}^k \right) \right) \boldsymbol{x} = \boldsymbol{0} \\
\frac{\partial \boldsymbol{erreur}}{\partial \boldsymbol{a}_2} &= -2 \sum_{i=1}^{n} \left(\boldsymbol{y}_i - \left(\boldsymbol{a}_0 + \sum_{k=1}^{j} \boldsymbol{a}_k \boldsymbol{x}^k \right) \right) \boldsymbol{x}^2 = \boldsymbol{0} \\
&\vdots \\
&\vdots \\
\frac{\partial \boldsymbol{erreur}}{\partial \boldsymbol{a}_j} &= -2 \sum_{i=1}^{n} \left(\boldsymbol{y}_i - \left(\boldsymbol{a}_0 + \sum_{k=1}^{j} \boldsymbol{a}_k \boldsymbol{x}^k \right) \right) \boldsymbol{x}^j = \boldsymbol{0}
\end{aligned} \quad \textbf{(C.7)}$$

L'équation (C.7) est mise sous la forme matricielle d'ordre j+1 :

$$\begin{bmatrix} n & \sum x_i & \sum x_i^2 & \cdots & \sum x_i^j \\ \sum x_i & \sum x_i^2 & \sum x_i^3 & \cdots & \sum x_i^{j+1} \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 & \cdots & \sum x_i^{j+2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \sum x_i^j & \sum x_i^{j+1} & \sum x_i^{j+2} & \cdots & \sum x_i^{j+j} \end{bmatrix} \times \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_j \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum (x_i y_i) \\ \sum (x_i^2 y_i) \\ \vdots \\ \sum (x_i^j y_i) \end{bmatrix} \qquad \textbf{(C.8)}$$

Tel que :

$$A = \begin{bmatrix} n & \sum x_i & \sum x_i^2 & \cdots & \sum x_i^j \\ \sum x_i & \sum x_i^2 & \sum x_i^3 & \cdots & \sum x_i^{j+1} \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 & \cdots & \sum x_i^{j+2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \sum x_i^j & \sum x_i^{j+1} & \sum x_i^{j+2} & \cdots & \sum x_i^{j+j} \end{bmatrix}, \quad X = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_j \end{bmatrix}, \quad B = \begin{bmatrix} \sum y_i \\ \sum (x_i y_i) \\ \sum (x_i^2 y_i) \\ \vdots \\ \sum (x_i^j y_i) \end{bmatrix} \qquad \textbf{(C.9)}$$

Ainsi la résolution de ce système linéaire consiste à multiplier l'inverse de la matrice *A* par le vecteur *B.*

$$X = A^{-1} * B \qquad \textbf{(C.10)}$$