Remerciements

La réalisation de ce mémoire en vue de l'obtention du diplôme de MASTER en télécommunication a été rendue possible grâce au soutien de plusieurs personnes à qui on voudrai témoigner notre reconnaissance, leurs disponibilités et leurs compétences qui nous a permis de franchir beaucoup d'obstacles, qu'ils trouvent ici le témoignage de nos gratitudes et nos remerciements les plus sincères.

On tient à remercier chaleureusement mon directeur de mémoire, monsieur Mohammed Saidi pour son grand soutien dans la réalisation de ce travail et son encadrement lors de la poursuite de notre travail ; tous ces facteurs ont grandement contribué à l'aboutissement de ce projet, qu'il soit assuré de ma reconnaissance et mon respect.

Nous remercierons l'ensemble des membres des jurys d'avoir accepté de juger et d'évaluer ce modeste travail.

Nous tenons à exprimer notre profonde gratitude à nos familles qui nous ont toujours soutenus et à tous ceux qui ont participé à la réalisation de ce mémoire. Ainsi que l'ensemble des enseignants qui ont contribué à notre formation

Dédicaces

Je dédie ce modeste travail

Au bon dieu qui m'a bénie de deux belles créations : Mon père et ma mère pour leurs
Motivations et leurs encouragements.

À mes trois neveux : Yanis, Ilyas et Farés que dieu les bénissent.

Je tiens également à exprimer mes remerciements et ma reconnaissance ;

À mes chères sœurs : Kahina et Fatima, mes frères : Tarek, Fahim et Hocine, mon beau-frère : Karim et mes belles sœurs : Fariza et Hinda auxquels je souhaite la réussite dans leurs vies

À mes trois cousines : Yassmina, Wissam et Kenza.

À ma meilleure amie : Siham, mes deux chers amis : Ibtissem, Sarah, Roumaissa

À tous mes collègues de travail au sein de mounif-procall particulièrement: Mounina, Souad, Françoise.

À mon cher binôme pour tout ce qu'elle a fait pour la réussite de ce projet.

À tous ceux qui ont participé de près ou de loin à l'élaboration de ce mémoire.

Thinhinene

Dédicaces

Je dédie ce mémoire également :

Tout d'abord, je remercie le Dieu, notre créateur de m'avoir donné de la force, la volonté et le courage afin d'accomplir ce travail modeste.

Je voudrais exprimer ma plus haute reconnaissance à mes parents pour leur soutien, leur aide et leur patience.

Mes reconnaissances envers ma famille :

Mes deux sœurs : Kahina et Sabrina, mes quatres frères : Rabah, Aissa, Farid et Samir mes neveux : Abderahmane, Abderaouf, Manel, Israe et Ritadje.

Mes remerciements à tous mes amis d'avoir été là pour moi, pour leurs présences et leurs encouragements.

J'exprime aussi ma reconnaissance à toutes les personnes qui m'ont apporté leurs soutiens, leurs amitiés tout au long de ce mémoire

Hanane

Résumé:

Malgré les énormes progrès de la communication numérique, la voix reste le principal outil avec lequel les gens échangent des idées. Cependant, la parole numérique non compressée a tendance à exiger des débits de données prohibitifs (jusqu'à 64 kbps) ce qui la rend peu pratique pour de nombreuses applications. Idéalement, on voudrait un système de codage capable de représenter le signal de parole avec un très bas débit, produisant un signal synthétisé d'une bonne qualité. La plus grande partie de ce mémoire présente les notions générales du codage de la parole comprend une enquête sur le codage paramétrique de la parole. Dans ce travail, nous avons aussi donné une présentation brève sur deux codeurs de compression de la parole ayant un débit de 2.4 Kbps en bande étroite 200-3400 Hz : MELP et LPC-10 et une évaluation de leurs implémentations en temps réel afin de comparer la synthèse de parole des deux codeurs à très bas débit.

Mots clés : codage paramétrique, codage à très bas débit, compression, LPC-10, MELP.

Abstracts:

Although the huge progress of digital communication, voice remains the main tool by which people exchange ideas. However, uncompressed digital speech tends to require prohibitive data rates (until 64 kbps) which makes it note very practical for many applications. Perfectly, we need a coding system able to represent the speech signal with a very lower flow, producing a synthesized signal of a transparent quality. The biggest part of this dissertation presents general notions of speech coding includes a survey of parametric speech coding. In this work, there is also a brief presentation on two speech compression coders having a flow of 2.4 kbps in narrow band 200-3400 Hz: MELP and LPC-10 with an evaluation of their implementation in real time in order to compare the speech synthesis of the two coders in a very low bite rate.

Key words:

Speech parametric, speech with a very low bite rate, compression, LPC-10, MELP.

ملخص :

على الرغم من التقدم الهائل في مجال الاتصالات الرقمية، يظل الصوت الأداة الرئيسية التي يتبادل بها الأفراد الأفكار. ومع ذلك، يميل الكلام الرقمي غير المضغوط إلى طلب معدلات بيانات باهظة (تصل إلى 64 كيلو بايت في الثانية)، مما يجعله غير عملي بالنسبة للعديد من التطبيقات. من الناحية المثالية، نود نظام تشفير قادر على تمثيل إشارة الكلام بمعدل بت منخفض جدًا، مما ينتج عنه إشارة مركبة ذات جودة شفافة. الجزء الأكبر من هذه الأطروحة يقدم المفاهيم العامة لترميز الكلام ويتضمن مسحًا لتشفير الكلام البارامتري. في هذا العمل، قدمنا أيضًا عرضًا موجزًا عن برنامجين للتشفير بمعدل 2.4 كيلوبت في الثانية في النطاق الضيق من 200 إلى 3400 هرتز وتقييم تطبيقاتهما في الوقت الفعلي للمقارنة بين المبرمجين عند معدل بايت منخفض للغاية هما :
LPC-10،MELP

الكلمات المفتاحية:LPC-10, MELP ضغط، ,تشفيربارامتري، تشفير منخفض جدا

Table des matières

Liste des figures

Les figures de chapitre I

Les figures de chapitre II

Les figures de chapitre III

Les figures de chapitre IV

Liste des Tableaux

Abréviations

ADPCM: la modulation de code par impulsions adaptative différentielle

AMDF: la fonction de différence d'amplitude moyenne

AR: Auto régressive

ARMA: auto regressive moving average

CR : coefficients de réflexion

DOD: département américain de la défense

DSVD: voix et données simultanées numériques.

FFT: fast fourier transform

FIR: filtre à réponse impulsionnelle finie

FS-1015: norme fédérale américaine 1015

LAR: log-Area Ratio

LP: prediction linéaire

LPC: linear predictive coding

LPF: filtre passe bas

LSF: line spectral frequencies

L'UIT: l'union internationale des télécommunications

MELP: mixed Excitation Linear Prediction

MDF: la fonction de différence de magnitude

MIC: modulation par Impulsion Codée

MIPS: millions of instructions per second

MSVQ: multi-Stage Vector Quantisation

MOS: mean opinion score

OTAN: l'organisation du traité de l'atlantique nord

PCM: pulse Code Modulation.

PESQ: perceptual evaluation of speech quality

RMS: root Mean Squared

SNR: signal to noise ratio

TFCT: la transformée de fourrier à court terme

VOIP: voice off protocole internet

Introduction générale

Les moyens de communications numériques sont actuellement en plein progression, Pour économiser les ressources des canaux de communication et un stockage efficace, la parole doit être compressée tout en gardant une très bonne qualité.

La parole est un type de signal très spécial. Elle est non stationnaire et cela la rend difficile à analyser et à modéliser. Les signaux de parole sont soumis à une compression avant la transmission.

Le codage ou la compression de la parole est un processus permettant d'obtenir une représentation des signaux de parole, en vue d'une utilisation efficace pour une transmission sur des canaux câblés ou sans fil limités en bande et également pour un stockage efficace. Le principal défi du système de codage de la parole est l'utilisation optimale de la largeur de bande des canaux qui a une incidence sur le coût de la transmission. Pour préserver la bande passante, les chercheurs ont optés à des codeurs de parole avec moins de bits ou ces codeurs sont devenus des composants essentiels pour les télécommunications et le multimédia.

L'objectif du codage de la parole est de représenter les échantillons d'un signal de parole avec un nombre minimal de bits sans réduction de la qualité de la perception c'est pour cela que c'était primordial d'obtenir un codeur de parole efficace en bande passante à faible débit de données avec une bonne qualité de parole. Sur la base de la qualité perceptuelle, afin d'obtenir le taux de compression du signal souhaité, la suppression des données redondantes reste une priorité avec une réduction correspondante du nombre de bits sans négligé le mécanisme de la parole [1].

Un codage de la parole à très bas débit, est nécessaire pour les applications de communication et de stockage vocal car un codage complet de la forme d'onde de la parole n'est pas possible. Par conséquent, les codeurs à très bas débit utilisent plutôt des modèles paramétriques pour ne représenter que les aspects les plus pertinents du point de vue perceptuel. Bien qu'il existe un certain nombre d'approches différentes pour cette modélisation.

En 1971, Atal et Hanauer ont proposé le codage vocal utilisant la prédiction linéaire(LP) qui est devenue de loin l'approche la plus populaire pour le codage à très bas débit.

Depuis les années 80, un progrès considérable a été réalisé dans le domaine du codage de la parole numérique, dans la bande téléphonique. Le développement de codeurs de parole de haute qualité, fonctionnant à débit faible, a été motivé par le marché croissant des systèmes digitaux de télécommunications et d'enregistrement, où les applications les plus importantes

sont les systèmes de radiocommunication avec les mobiles, les systèmes de communication par satellite, les systèmes de communication pour les multimédia, la téléphonie par Internet et la visiophonie. Ce progrès a été rendu possible grâce aux nouveaux processeurs rapides de traitement du signal, à la meilleure compréhension des processus de production et de perception du signal de parole et enfin au développement d'algorithmes efficaces de codage.

En 1984, la première norme en matière de communication vocale numérique a été adoptée : FS-1015 décrit un vocodeur appelé LPC10 qui fonctionne sur 2.4 kbps.

Au cours des années suivantes, le développement du vocodeur a été conduit à réaliser une modélisation paramétrique du signal sous la forme d'un signal d'excitation passant au travers d'un filtre, en exploitant d'une certaine manière les propriétés de la perception humaine, le filtre, appelé filtre de synthèse, est généralement modélisé par la prédiction linéaire (LP), le plus souvent, il s'agit d'un filtre autorégressif pur : c'est le codeur MELP qui était développé pour surmontée certaines limitations de codeur FS-1015 [2].

L'objectif de notre travail est de présenter un système capable de représenter le signal de parole avec un débit très réduit à 2.4 kbps tout en produisant un signal synthétisé d'une bonne qualité et aussi maintenir le délai du codage très court et assurer une parfaite robustesse contre les erreurs de transmission pour réduire le cout de transmission et moins d'espace de stockage.

Dans ce cadre, l'Union International des Télécommunications (UIT) a développé des normes spécifiant les procédures expérimentales à suivre pour évaluer la qualité perceptuelle du signal vocal ; parmi eux, la mesure PESQ.

La méthodologie adoptée, dans la rédaction de ce travail est présentés comme suit :

Le premier chapitre : est devisée à deux sections. La section 1 est consacrée à une description du signal de la parole. Quelques généralités de codage de la parole seront présentées dans la section 2.

Le deuxième chapitre : c'est une description générale d'un codeur a très bas débit : FS-1015

Le troisième chapitre : est consacré à l'implantation d'un codeur de parole à prédiction linéaire a mixte excitation (MELP) a 2.4 kbps.

Le quatrième chapitre : contient une évaluation des deux codeurs LPC-10 et MELP en utilisant l'évaluation perceptuelle de la qualité vocale (PESQ).

La conclusion générale et les perspectives possibles à notre travail sont présentées dans la conclusion générale.

Chapitre I : l’état de l’art de codage de la parole

I.1. Introduction

La parole est le principal moyen de communication dans toute société humaine. La parole peut être décrite comme le résultat de l'action volontaire et coordonnée d'un certain nombre de muscles. Cette action se déroule sous le contrôle du système nerveux central qui reçoit en permanence des informations par rétroaction auditive.

Il est difficile de décrire les phénomènes liés au signal de la parole, sans évoquer ses caractéristiques, ses classifications, ses paramètres qui le caractérise, et son traitement de numérisation, Il s'agit bien sûr d'une description globale.

Il est maintenant nécessaire de numériser les signaux. Le débit binaire du signal numérisé est alors égal au produit de la fréquence d'échantillonnage par le nombre d'éléments binaires nécessaire à la représentation de toutes les valeurs discrètes du signal. Pour réduire ce débit, des algorithmes vont permettre de supprimer les redondances inutiles du signal, nécessitant ainsi un système de codage, il existe des différentes méthodes de codage ainsi plusieurs techniques peuvent être utilisées.

I.2. Description et caractéristique du signal de la parole

I.2.1. Caractéristique de signal de parole

Le signal de parole est l'un des signaux les plus complexes à caractériser et analyser, ces caractéristiques sont :

I.2.1.1. La variabilité du signal de parole

Le signal vocal a contenu phonétique égal, est différent pour un même locuteur (variabilité intra locuteur) ou pour des locuteurs différents (variabilité interlocuteur) [3].

I.2.1.1.1. Variabilité intra-locuteur

Lorsqu'une personne prononce deux fois le même mot, on constate une différence dans le signal produit. Toute affection de l'appareil phonatoire peut dégrader la qualité de signal de parole. Cette dégradation peut être causée par : une simple fatigue, le stress, l'amplitude de la voix, ou l'émotion du locuteur ce qui fait que l'articulation perd de sa clarté [4].

I.2.1.1.2. Variabilité interlocuteur

La variabilité interlocuteur est d'une importance évidente car les différences au sein des classes phonétiques sont nombreuses. Sa cause principale est de nature physiologique. On trouve aussi les différences de prononciations qui existent au sein d'un milieu social et qui constituent les accents régionaux [4].

I.2.1.2. La continuité

En raison de l'anticipation du geste articulatoire, la production d'un son est influencée par le son qui le précède et par le son qui le suit. La localisation correcte d'un segment de parole isolé de son contexte est parfois impossible [3].

I.2.2. Classification des sons de la parole

Le signal de la parole ressorti deux types de sons: voisés et non voisés.

I.2.2.1. Les sons voisés

Tels que les voyelles, Les sons voisés sont produits par la vibration des cordes vocales par des impulsions périodiques de pression liées aux oscillations des cordes vocales. C'est un signal quasi périodique qui possède un spectre fréquentiel très caractéristique. La période fondamentale des différents sons voisés varie entre 2ms et 20ms [5].

I.2.2.2. Les sons non voisés

Tels que certaines consonnes, Dans ce cas les cordes vocales ne vibrent pas, donc les sons non voisés ne présentent pas de structure périodique, le signal produit est considéré comme un équivalent à un bruit blanc filtré par le filtre résultant de la partie du conduit vocal situé dans la bouche [6].

I.3. Paramètres du signal de parole

La parole est un signal continu, d'énergie finie, non stationnaire. Le signal parole est généralement caractérisé par trois paramètres: sa fréquence fondamentale, son énergie et son spectre.

I.3.1. La fréquence fondamentale

Elle représente la fréquence du cycle d'ouverture/fermeture des cordes vocales. Cette fréquence caractérise seulement les sons voisés, elle peut varier [5]:

- De 80Hz à 200Hz pour une voix masculine.
- De 150Hz à 450Hz pour une voix féminine.
- De 200Hz à 600Hz pour une voix d'enfant.

I.3.2. L'énergie

Elle correspond à l'intensité sonore qui est liée à la pression de l'air en amont du larynx. L'amplitude du signal de la parole varie au cours du temps selon le type de son, et son énergie dans une trame est donnée par [5] :

$$\mathbf{E = \sum_{n=0}^{N-1} s^2\,(n)} \qquad \textbf{(I.1)}$$

Avec N : la taille de la trame.

s(n) : signal de la parole.

I.3.3. Le spectre

La représentation fréquentielle de l'intensité de la parole définit l'enveloppe spectrale ou le spectre, en général elle est obtenue par une analyse de Fourier à court terme. La quasi stationnarité du signal de parole permet de mettre en œuvre des méthodes efficaces d'analyse et de modélisation utilisées pour le traitement à court terme du signal vocal sur des fenêtres de durée généralement comprise entre 20ms et 30msappelées trames, avec un recouvrement entre ces fenêtres qui assure la continuité temporelle des caractéristiques de l'analyse.

La transformée de Fourier à court terme (TFCT) d'un signal échantillonné est par définition la transformée du signal pondéré [5]. Dont l'expression est :

$$\widehat{\boldsymbol{S}}(\boldsymbol{k}) = \widehat{\boldsymbol{S}}\left(\boldsymbol{f} = \frac{\boldsymbol{k}}{\boldsymbol{N}}\right) = \sum_{\boldsymbol{n}=\boldsymbol{0}}^{\boldsymbol{N}-\boldsymbol{1}} \boldsymbol{s}(\boldsymbol{n}).\boldsymbol{w}(\boldsymbol{n}).\exp(-\boldsymbol{2j\pi\, nk/N}),\ \boldsymbol{0} \leq \boldsymbol{k} \leq \boldsymbol{N}-\boldsymbol{1} \qquad \textbf{(I.2)}$$

Où ; N : Le nombre de points prélevés.

S(k) : Spectre complexe.

s(n) : Segment analysé.

w(n) : Fenêtre d'analyse temporelle.

Le spectre de puissance appelé aussi densité spectrale de puissance de la transformé de Fourier est donné par :

$$|\hat{\mathbf{s}}(\mathbf{k})| \qquad \mathbf{0} \leq \mathbf{K} \leq \frac{\mathbf{N}}{\mathbf{2}} \qquad \textbf{(I.3)}$$

I.4. Automatisation de la Parole

Le traitement de signal de la parole nécessite toujours en premier lieu une conversion de signal vocal en signal électrique par le microphone, ce qui est impossible pour que la machine puisse l'interpréter ou la prédire car elle ne manipule pas des sources analogiques. Pour cela, on doit faire un traitement de numérisation sur ce signal [7].

La numérisation de signal parole est le processus qui permet de construire une représentation discrète d'un objet du monde réel. Ce procédé implique d'abord un échantillonnage suivie d'une opération de quantification puis en dernière étape le codage.

Une chaine de transmission numérique peut être représentée par différents blocs modélisant les traitements successifs apportés à l'information. Les blocs peuvent être énumères comme suit [8] :

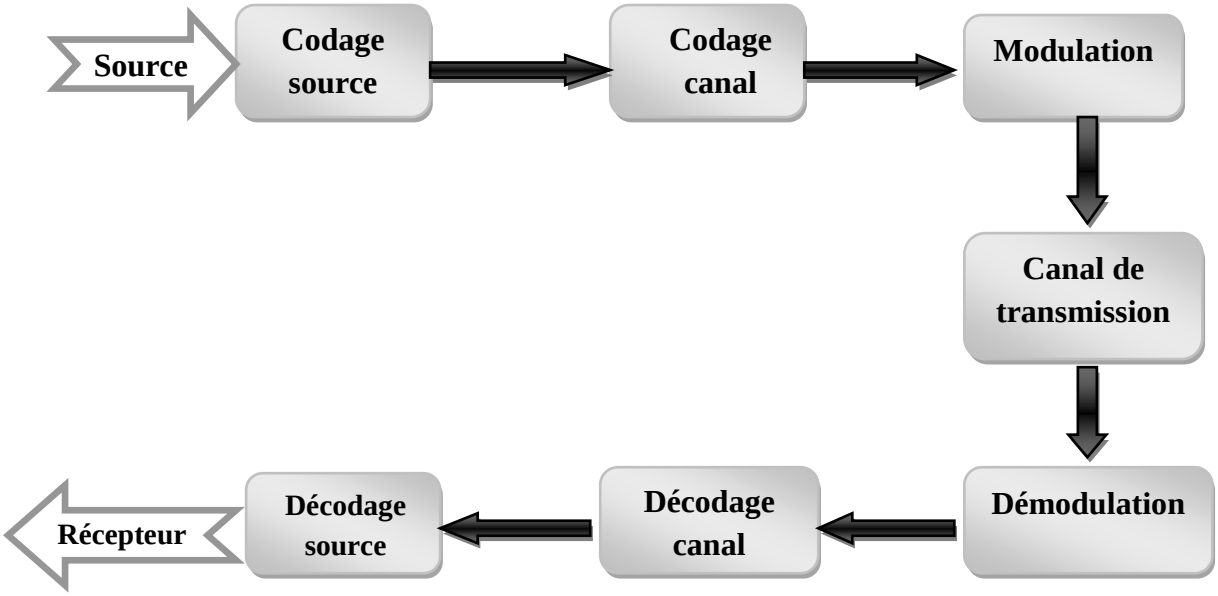


Figure. I.1: Schéma bloc d'une chaîne de transmission numérique [8]

- ✓ **La source**: génère l'information à émettre (message)
- ✓ **Le codage de source** : ou **la compression de données** :

Elle permet de réduire la taille des données à envoyer, à traité ou à stocker en supprimant les redondances contenues dans le message afin de réduire le débit et rendre la transmission plus rapide mais sans une perte d'informations initiales.

Il existe trois critères sur lesquels se base les algorithmes de compression [7]:

- Le taux de compression **:** c'est le rapport de la taille du fichier compressé sur la taille du fichier initial.il est généralement exprimé en pourcentage. La formule pour calculer ce taux est :

$$\boldsymbol{\tau = 1 - \left(b/a\right)} \qquad \textbf{(I.4)}$$

Où : $\boldsymbol{\tau}$ = le taux de compression

b= la taille du fichier compressé

a= la taille du fichier initial

- La qualité de compression **:** sans ou avec pertes (avec le pourcentage de perte).
- La vitesse de compression et de décompression.

➢ **Classification des algorithmes de compression**

Le critère de classification le plus pertinent est basé sur la perte des données. On peut distinguer deux types de compression qui seront présentés comme suit [9] :

❖ **Compression sans perte :**

Ce type de compression est très utile, car certains types de fichiers, notamment les fichiers de données, ne peuvent pas se permettre de perdre aucun bit de donnée pour ne pas perdre l'intégralité de leur sens. Les algorithmes de compression sans perte (connu aussi sous le nom de non destructible, réversible, ou conservative) sont des techniques permettant une reconstitution exacte de l'information après le cycle de compression / décompression. Le principal algorithme de compression sans perte est l'algorithme de Huffman.

- **L'algorithme de Huffman:** il consiste à diminuer au maximum le nombre de bits utilisés pour coder une suite d'information, il se base sur la fréquence d'apparition d'un caractère pour le coder : plus le caractère est fréquent, moins on utilisera de bits pour le coder. On va donc, par création de la table des fréquences, puis de l'arbre dit de Huffman, réussir à comprimer le fichier. Son but est de :

❖ Réduire le nombre de bits utilisés pour le codage des caractères fréquents.

❖ Augmenter ce nombre pour des caractères plus rares.

Il faut utiliser un autre type de compression car le taux de compression des algorithmes sans perte est en moyenne de l'ordre de 40% pour des données de type texte. Par contre, ce taux est insuffisant pour les données de type multimédia. Donc on va utiliser : la compression avec perte [10].

❖ **Compression avec perte**

L'algorithme de compression repère donc les sons "dominants" et retire toutes les données relatives aux sons "dominés" pour ne transmettre que ce qui est perceptible, il élimine l'information redondante et introduit une dégradation indiscernable a l'oreille avec un taux de compression très élevé. Donc, elle ne s'applique qu'aux données « perceptibles », Les données originales ne peuvent pas être retrouvées, donc la perte d'information est irréversible c.-à-d. non conservative.

Les données cachent les bruits de certaines fréquences et les différences de phase ne sont pas détectées par l'oreille humaine pour la plupart des fréquences. Tout cela donne la possibilité de comprimer la parole pour obtenir des résultats de bonne qualité avec un stockage minimal [9].

✓ **Le codage de canal** : insère des éléments binaires pour améliorer la qualité de la transmission, Le codage de canal protège les messages des distorsions générées par le « bruit » du canal. De ce point de vue, le processus de codage et le crypto ont des rôles similaires. Dans le cas du bruit il n'y a pas d'interception du message, mais

simplement du « brouillage » des messages. Les messages reçus peuvent être interprétés en erreur. Le but fondamental du codeur de canal, consiste à réduire la probabilité d'erreur.

- ✓ **Le modulateur**: traduit le message binaire en signal permettant son transport dans les milieux tel que l'air, l'eau, les câbles etc.
- ✓ **Le canal de transmission** : propage le signal ; lors de la propagation, le signal peut être perturbé par du bruit externe, des multi-trajets, le mouvement de l'émetteur et / ou récepteur etc.
- ✓ **Le démodulateur** : traduit le message reçu en signal binaire.
- ✓ **Le décodeur de canal** : détecte et/ou corrige les erreurs de transmission grâce aux éléments binaires ajoutés lors du codage.
- ✓ **Le décodeur de source** : régénère le message binaire [8].

I.4.1. L'échantillonnage

L'échantillonnage est la première étape de procédé de numérisation, elle consiste à transformer un signal continu en une suite de valeurs discrètes (discontinues).

Le signal analogique est découpé en échantillons. Le nombre d'échantillons par seconde représente la fréquence d'échantillonnage Fe, elle est exprimée en Hertz. Celle-ci est elle-même l'inverse de la période d'échantillonnage Te. Le choix de la fréquence d'échantillonnage n'est pas aléatoire elle doit être suffisamment grande, afin de préserver la forme du signal .il faut prélever assez de valeurs pour ne pas perdre l'information contenue. Le théorème suivant traite cette problématique : **Théorème (de Shannon)** « La fréquence d'échantillonnage assurant un non repliement du spectre doit être supérieure ou égale au double de la fréquence la plus élevée de signal analogique. ».

$$\mathbf{feh \geq 2fmax} \qquad \mathbf{(I.5)}$$

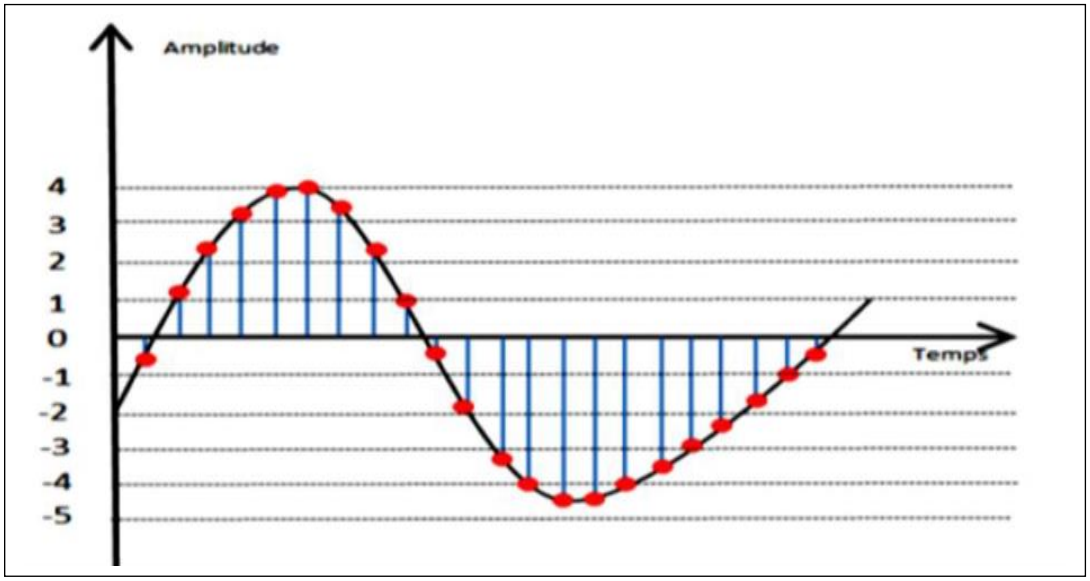


Figure. I.2: Echantillonnage d'un signal.

Pour une bonne représentation du signal de la parole, on utilise des fréquences d'échantillonnage de 16 ou 8 kHz [11].

I.4.2. Quantification

La quantification consiste à approximer les valeurs réelles à un signal instantané exact par la valeur la plus voisine selon une échelle de n niveaux appelée échelle de quantification.

Il y a donc 2^n valeurs possibles comprises entre $(-2n-1)$ et $(2n-1)$ pour les échantillons quantifiés. Le nombre de bit de quantification détermine la dynamique de la conversion, et la fréquence d'échantillonnage détermine la précision temporelle de la conversion.

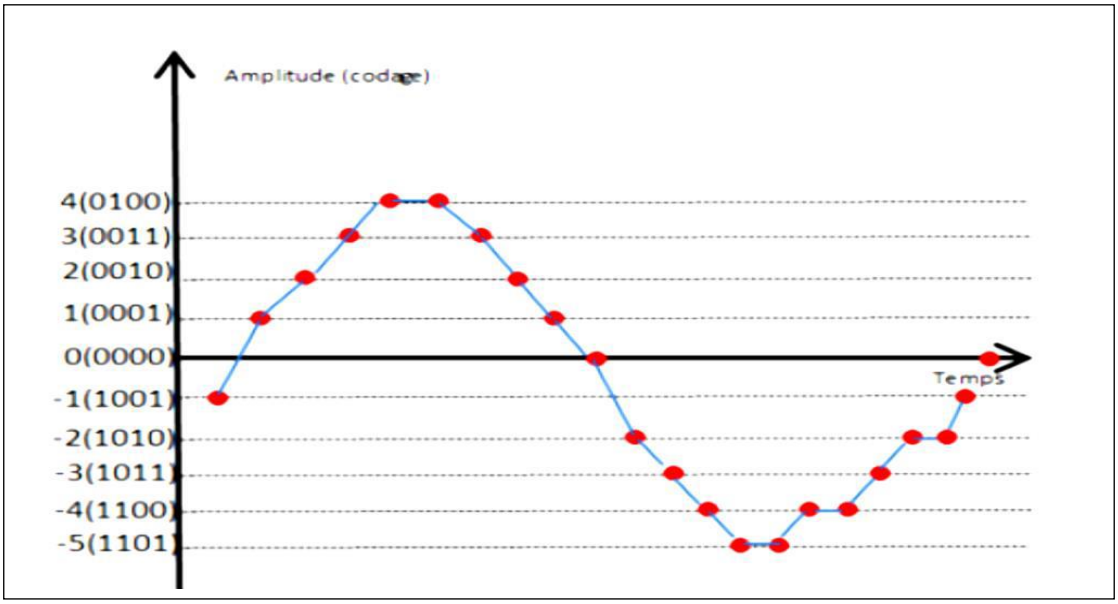


Figure. I.3: Quantification d'un signal échantillonné [11].

I.4.3. Codage des données

C'est la représentation binaire des valeurs quantifiées qui permet le traitement du signal sur machine [11].

I.5. Etat de l'art : Codage de la parole et format

I.5.1. Le processus de codage de la parole

Le codage de la parole c'est l'art qui permet de crée un minimum de représentation redondante du signal de parole pouvant être efficacement transmise ou stocké sur un support numérique, et décoder le signal avec la meilleure qualité possible [12]. Un système de codage de la parole comprend deux parties : un codeur et un décodeur. Le codeur analyse le signal pour en extraire un nombre réduit de paramètres pertinents qui sont représentés par un nombre restreint de bits pour archivage ou transmission. Le décodeur utilise ces paramètres pour

reconstruire un signal de parole synthétique [13]. Les caractéristiques principales d'un codeur sont [14] **:**

- ✓ Le débit binaire: fixe ou variable.
- ✓ La qualité de parole reconstruite.
- ✓ La complexité de calcul et mémoire requise.
- ✓ Le délai.
- ✓ La sensibilité aux erreurs dues au canal de transmission.
- ✓ La largeur de bande du signal.

La plupart des algorithmes de codage mettent à profit un modèle linéaire simple de production de la parole. Ce modèle sépare la source d'excitation, qui peut être quasi-périodique pour les sons voisés ou de type bruit pour les sons non voisés, du canal vocal qui est considéré comme un résonateur acoustique.

Quand un signal de parole numérisé est transmis, la bande passante requise est en fonction du débit binaire, de là on peut parler de compression, son but est de produire une représentation compacte des sons de parole de sorte que lorsqu'ils seront reconstruits, ils soient perçus comme très proches de l'original. Les deux mesures principales sont l'intelligibilité et la qualité. Le point de référence standard est la qualité de Toll qui est utilisé sur des lignes téléphoniques.

De là on peut dire que le rôle de tous les systèmes de codage de parole est de transmettre la parole avec une bonne qualité en employant la capacité de canal la plus petite possible. En général, il y a une corrélation positive entre l'efficacité du débit du codeur et la complexité algorithmique exigée pour le réaliser. Plus un algorithme est complexe, plus le coût du traitement et de mise en œuvre sera élevé [14].

Le codage de la parole inclut la communication cellulaire, la voix off protocole Internet (VOIP), vidéoconférence, électronique jouets, archivage, voix et données simultanées numériques (DSVD), ainsi que de nombreux jeux et logiciels sur PC et applications multimédia [12].

I.5.2. Méthode de codage de la parole

Les méthodes de codage de la parole sont généralement classées comme : sans perte et avec perte.

En codage sans perte, le signal de parole reconstruit à l'extrémité du décodeur peut avoir exactement la même forme que l'entrée du signal de la parole. Les méthodes de codage avec pertes ont le signal de parole reconstruit qui est différent du signal de parole original Les techniques de codage de la parole reposent principalement sur la technique de codage avec

perte car elle supprime les informations non pertinentes du point de vue de la qualité perceptuelle [12].

Les méthodes de codage de la parole permettent d'obtenir :

➢ Réduction du débit binaire ou équivalent de la bande passante.

➢ Réduction des besoins en mémoire, qui diminue proportionnellement par rapport au débit binaire.

➢ Réduction de la puissance de transmission requise, car le signal vocal compressé a moins de nombre de bits par seconde à transmettre.

➢ Immunité au bruit, certains des bits enregistrés par échantillon peuvent être utilisé comme des bits de contrôle d'erreur pour les paramètres de parole [12].

I.5.3. Classification des codeurs de parole

Les codeurs de parole sont classés en fonction du débit binaire auquel ils produisent une sortie avec une qualité raisonnable et aussi selon le type de codage (techniques utilisées pour coder le signal de parole).

I.5.3.1. Classification par débit binaire

Les codeurs de la parole sont classés dans quatre catégories en fonction de la plage de débits dans laquelle les codeurs de parole produisent une qualité raisonnable :

Codeurs à haut débit, codeur a débit moyen, codeurs à faible débit et codeurs à très faible débit.

Type de codeur	Codeurs à haut débit	Codeurs à débit moyen	Codeurs à faible débit	Codeurs à très bas débit
Plage de débit binaire	> 15 Kbps	5 à 15 Kbps	2.5 à 5 Kbps	<2.5 Kbps

Tableau. I.1. Classification des codeurs de parole en fonction du débit [12]

I.5.3.2. Classification par techniques de codage

Selon le type de technique de codage utilisé, les codeurs de parole tendent à maximiser le compromis entre l'efficacité, le coût et la qualité de transmission des systèmes de

communication en fonction des débits disponibles. Sont classés en trois types et sont expliqués ci-dessous :

I.5.3.2.1. Les codeurs de forme d'onde (temporel)

Les codeurs de formes d'onde numérisent le signal vocal sur un échantillon base, très simples à mettre en œuvre, Son objectif principal est de faire en sorte que la forme d'onde de sortie ressemble à la forme d'onde d'entrée et aussi à préserver l'allure temporelle du signal de parole, ce qui les rend robustes aux différents types d'entrée. Ainsi, les codeurs de forme d'onde conservent une parole de bonne qualité. Les codeurs à la forme d'onde ne sont pas complexes, ils produisent un signal de parole a des débits supérieurs à 16 Kbps environ. Quand le débit de données est abaissé au-dessous de cette valeur, la qualité de signal de parole reconstruit se dégrade. Les codeurs de forme d'onde ne sont pas spécifiques à la parole et peut être utilisé pour tout type de signaux. Les deux types des codeurs de formes d'onde sont des codeurs de domaine temporel et des codeurs de domaine fréquentiel. Les codeurs de forme d'onde du domaine temporel utilisent le schéma de numérisation basé sur les propriétés du domaine temporel du signal de la parole. Certains des exemples de techniques de codage de forme d'onde dans le domaine temporel sont la modulation par impulsions et codage PCM (pulse code modulation) ou MIC (modulation par impulsion codée) y'a aussi ADPCM (modulation de code par impulsion adaptative différentielle) [15].

I.5.3.2.2. Les codeurs paramétriques (vocodeurs)

Dans les codeurs paramétriques, le signal de parole est supposé être généré à partir d'un modèle contrôlé par certains paramètres de parole qui correspondant au mécanisme de production de la parole qui sont obtenus en analysant le signal de parole avant transmission. A la réception, le décodeur utilise ces paramètres pour reconstruire le signal de parole original. Dans ces codeurs, la sortie le signal vocal ne ressemble au signal vocal entrant. Ce type de codeur ne de conserve la forme d'origine de signal, le SNR est donc une mesure de qualité inutile. Cependant, le signal vocal de sortie émettra la même chose que le signal vocal d'entrée. Les codeurs paramétriques également connus sous le nom de vocodeurs permettre des transmissions à moyen et bas débit (entre 5 et 16 Kbits/s). La plupart des codeurs paramétriques sont basés sur le codage prédictif linéaire (LPC), Elle repose principalement sur l'hypothèse que la parole peut être modélisée par un processus linéaire, il s'agit donc de prédire le signal à un instant n à partir des p échantillons précédents. La qualité de cette classe de codeurs est limitée par la reconstruction synthétique du signal. Cependant, Aux plus bas débits, la qualité atteinte par les codeurs de forme d'onde est inférieure par rapport aux codeurs paramétriques. Exemples de codeurs de cette classe: codage par prédiction linéaire

(LPC) et prédiction linéaire à excitation mixte (MELP) [12], Un codage efficace peut actuellement être réalisé à des débits inférieurs à 2 Kbits/s avec des vocodeurs basés sur l'analyse LPC.

Dans les vocodeurs prédictifs, le conduit vocal est modélisé par un filtre tout-pôle de fonction de transfert H(z) :

$$\boldsymbol{H(z) = \frac{1}{A(z)}} \quad \textbf{(I.6)}$$

$$\mathbf{A(z) = 1 + \sum_{k=1}^{p}\left(a_k z^{-k}\right)} \quad \textbf{(I.7)}$$

Les propriétés du signal restent essentiellement constantes. Dans chaque trame, les paramètres du modèle sont estimés à partir des échantillons de parole; dans le cas présent, ces paramètres sont les suivants [15]:

- Envoi de voix: que le cadre de trame soit voisé ou non.
- Gain: principalement lié au niveau d'énergie de cadre.
- Coefficients de filtre: spécifiez la réponse du filtre de synthèse.
- Durée du pitch: dans le cas de trames vocales, durée entre deux impulsions d'excitation.

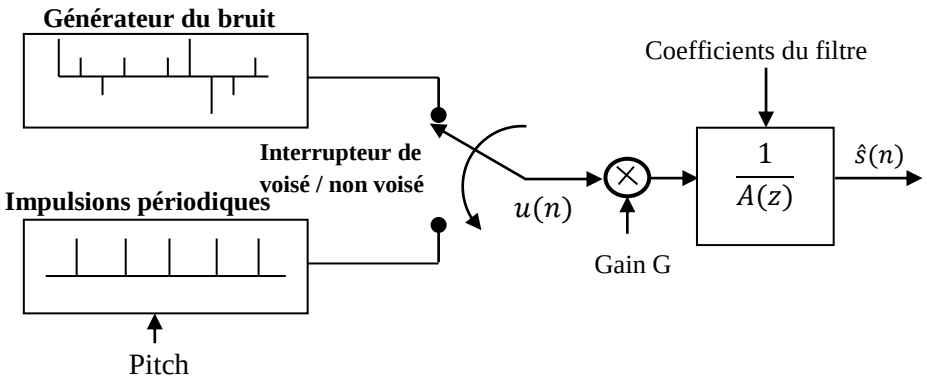


Figure. I.4: Le modèle LPC de production de la parole [15].

I.5.3.2.3. Les codeurs hybrides

Un codeur hybride est une combinaison du codeur de forme d'onde et du codeur paramétrique. L'utilisation des techniques temporelles conduit à une excellente qualité de parole mais induit un débit assez élevé. Comme codeurs paramétriques, les codeurs hybrides reposent sur la production de la parole modèle. La fréquence d'échantillonnage étant fixée, la réduction de débit des codeurs temporels fait chuter rapidement la qualité d'écoute pour des débits inférieurs à 16 Kbits/s. Une meilleure qualité pourra être observée pour des vocodeurs jusqu'à des débits de 4 Kbits/s. Mais ses applications restent réduites à cause d'une

complexité accrue. Des codeurs hybrides utilisent alors les deux méthodes temporelle et paramétrique de façon complémentaire, ce qui permet un codage de parole de bonne qualité à des débits moyens. Ces codeurs sont basés sur des techniques de codage temporel auxquelles des modèles de production de parole sont associés pour améliorer leur efficacité. Cependant, ce type de codage nécessitera des coûts de calculs plus importants. 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 temporelle modélisent respectivement le conduit vocal et le signal d'erreur résiduel.

D'un point de vue technique, la différence entre un codeur hybride et un codeur paramétrique est que le premier tente de quantifier ou de représenter le signal d'excitation au modèle de production de parole, qui est transmis en tant que partie du flux de bits codé. Ce dernier, cependant, atteint un faible débit en éliminant toutes les informations détaillées du signal d'excitation; seuls les paramètres importants sont extraits.

Un codeur hybride a tendance à se comporter comme un codeur de forme d'onde pour un débit binaire élevé et comme un codeur paramétrique à faible débit, avec une qualité passable à bonne pour un débit moyen [14].

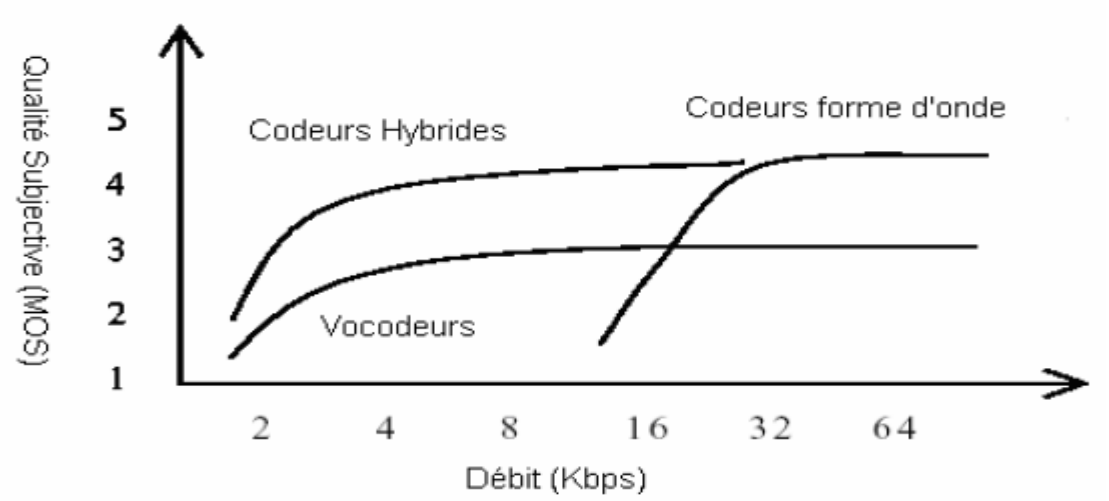


Figure. I.5: performances des codeurs temporels,paramétrique et hybrides [14].

I.6. Modélisation paramétrique

Les procédés de traitement numérique du signal appliqués aux signaux de parole se basent souvent sur une modélisation paramétrique du signal de parole. Ce modèle comporte généralement quatre paramètres [16] :

•Une information de « voisement », qui décrit le caractère plus ou moins périodique (sons voisés) ou aléatoire (sons non voisés) du signal ;

•La fréquence fondamentale ou « PITCH », pour les sons voisés ;

•L'évolution temporelle de « l'énergie » du signal de parole ;

•Son enveloppe spectrale, ou « spectre », qui est généralement obtenue par une modélisation autorégressive (filtre de prédiction linéaire ou LPC) ou par une analyse de Fourier à court terme synchrone avec le pitch. Ces quatre paramètres sont estimés périodiquement sur le signal de parole (d'une à plusieurs fois par trame selon le paramètre, pour une durée trame typiquement comprise entre10 et 30 ms).

Dans le cas général p>0 etq>0, on parle de modélisation *ARMA*. Lorsque Bz =1 (p>0, q=0), il s'agit d'une modélisation *AR* [16].

I.6.1. Modélisation AR du signal de la parole

Pour le codage prédictif (LPC) de la parole, on modélise l'onde de parole comme la sortie d'un filtre tout-pole. Cette onde est divisée en nombreux intervalles successives de 30 ms décalées de 10 ms pendant lesquels le signal de parole est considéré comme stationnaire. Pour chaque intervalle, les coefficients constants du filtre tout-pole sont estimés en prédiction linéaire par minimisation d'un critère quadratique de l'erreur de prédiction.

Si la fréquence d'échantillonnage est de 10kHz, l'estimation de chaque modèle est faite sur 300 échantillons. Un filtre AR excité par un train d'impulsion engendre un signal voisé et un filtre AR excité par un bruit blanc engendre un signal non voisé [5].

I.6.1.1. Prédiction linéaire

La Prédiction Linéaire LP (Linear Prediction) conduit à un modèle AR, c'est un des outils les plus importants de l'analyse de parole [15]. Le modèle de filtre source utilisé dans le protocole LPC comporte deux composants clés: l'analyse LPC (codage) et la synthèse LPC (décodage). Dans l'analyse LPC, le signal de parole est segmenté en blocs appelés trames. Chaque trame est examinée pour trouver:

- Si le trame est voisée ou non.
- Pitch de chaque trame.
- Paramètres nécessaires pour créer un filtre qui modélise le tract vocal.

Cette technique est prédominante dans les systèmes de codage de parole à très bas débit, elle est efficace et facile a mise en œuvre [1]. Elle est basée sur l'hypothèse que chaque échantillon de signal parole x(n) peut être approximé par une combinaison linéaire d'échantillons qui le précède.

$$\boldsymbol{x(n) = -a(1).x(n-1) - a(2).x(n-2) \ldots - a(p).x(n-p) + e(n)} \qquad \textbf{(I.8)}$$

Dans une trame de signal. En effet, l'analyse par prédiction linéaire est une procédure d'estimation permettant de trouver les paramètres AR. L'hypothèse de base est que la parole peut être modélisée comme un signal AR, ce qui en pratique s'est avéré approprié.

Dans cette expression les coefficients (i) sont les coefficients de prédiction d'ordre p, et le signal:

$$\boldsymbol{e(n) = \sum_{i=0}^{p} a(i).x(n-i)a(0) = 1} \quad \textbf{(I.9)}$$

L'énergie : En général, les sons voisés ont une énergie plus élevée que les signaux non voisés. Pour une trame (de longueur N) se terminant à l'instant m, l'énergie est donnée Par l'équation suivante [17]:

$$\mathbf{E[m] = \sum_{n=m-N+1}^{m} s^2\,[n]} \quad \textbf{(I.10)}$$

La variance : L'estimation des coefficients de prédiction est basée sur la minimisation de la variance de l'erreur de prédiction [17]:

$$\boldsymbol{\delta_e{}^2 = E[e\,(n)^2]} \quad \textbf{(I.11)}$$

$$\boldsymbol{= E[\sum_{i=0}^{p} a(i).x(n-i).\sum_{j=0}^{p} a(j).x(n-j)]} \quad \textbf{(I.12)}$$

$$\boldsymbol{E = [\sum_{i,j}^{p} a(i).a(j).x(n-i)} \quad \textbf{(I.13)}$$

Le gain de prédiction: défini comme le rapport entre l'énergie du signal et l'énergie de l'erreur de prédiction [17]:

$$\boldsymbol{PG[m] = 10\log_{10}(\frac{\sum_{n=m-N+1}^{m} s^2\,[n]}{\sum_{n=m-N+1}^{m} e^2\,[n]})} \quad \textbf{(I.14)}$$

Le taux de passage par zéro : de la trame est défini par [17] :

$$\mathbf{ZC[m] = \frac{1}{2}\sum_{n=m-N+1}^{m} |\ sgn(s[n] - sgn(s[n-1])\ |} \quad \textbf{(I.15)}$$

La prédiction linéaire à court terme considère que tout échantillon de parole peut être exprimé comme une combinaison linéaire d'échantillons précédents. Un ensemble de coefficients de prédiction est alors déterminé et utilisé pour supprimer la redondance proche entre les échantillons du signal de parole (redondance à court terme).

La prédiction à long terme prend en compte la corrélation entre des échantillons éloignés du signal de parole (présente particulièrement pour les sons voisés). L'extraction de cette périodicité est obtenue par estimation de la période du pitch [17].

I.6.2 Modélisation ARMA (auto régressive moving average) de signal de la parole

Il existe d'autres modèles linéaires que le modèle AR qui sont rencontrés moins fréquemment; et parfois sont appliqués pour des tâches spécifiques : le modèle ARMA. En utilisant un ordre élevé du modèle AR, plusieurs tentatives sont faites pour estimer les sons nasalisés et qui n'ont pas été très réussies à cause des formants parasites qui sont produites. La

contribution de la région nasale nécessite un modèle ARMA (ou pôle-zéro) [5]. Ceci peut être décrit par la fonction de transfert suivante :

$$\boldsymbol{V(z) = \frac{N(z)}{D(z)} = \frac{\sum_{k=0}^{q} b_k z^{-k}}{1+\sum_{k=1}^{p} a_k z^{-k}}} \quad \textbf{(I.16)}$$

Où :

$\boldsymbol{a_k}$: les coefficients des pôles, $\boldsymbol{b_k}$: les coefficients des zéros

p et **q** : sont les ordres des pôles et des zéros

I.7. Conclusion

Le codage de parole permet la réduction de débit de transmission du signal dans des canaux à largeur de bande limitée. La largeur de bande du canal de transmission doit être minimisée tout en préservant la qualité du signal vocal reconstruit

Tout dépend de l'utilisation du signal compressé. Lorsque qu'on veut travailler avec un format numérique tel qu'il est utilisé dans le milieu de l'audio professionnel, il est préférable d'utiliser une technique de compression sans pertes mais si on veut faire du stockage massif de donnée il est préférable d'utiliser une méthode de compression avec pertes.

Dans ce chapitre, nous avons présenté les techniques les plus utilisées pour le codage et la compression de la parole. Comme le nombre et le type de codeurs-décodeurs est élevé, nous n'avons pas pu les présenter tous.

Notre deuxième chapitre repose une étude d'un codeur à très bas débit.

Chapitre II : étude d’un codeur à très bas débit à 2.4 kbps

II.1. Introduction

Le codage prédictif linéaire (LPC) est plus efficace pour compresser les informations spectrales en un petit nombre de coefficients de filtre pour lesquels des techniques de quantification très efficaces sont facilement disponibles. La compression des signaux vocaux fait référence à la réduction de la largeur de bande nécessaire pour transmettre ou stocker un signal vocal numérisé. Idéalement, on voudrait un système capable de représenter le signal de parole avec un débit très bas, produisant un signal synthétisé d'une bonne qualité tout en maintenant une complexité de calcul raisonnable.

La plupart des méthodes de codage de la parole ont été conçues pour éliminer les redondances et les informations non pertinentes contenues dans la parole, dans le but de produire une parole de haute qualité avec des débits binaires très bas.

Dans ce chapitre on va voir le codeur FS-1015 qui facilite l'interopérabilité des systèmes de communications gouvernementales qui utilisaient un débit de 2400 bit/s pour les signaux vocaux.

II.2. Les codeurs à très bas débits

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 en 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 [13].

II.3. Codage de la parole à 2,4 kbps

Afin de produire un discours de bonne qualité à 2.4 kbps, différentes approches doivent être suivies.

Le vocodeur LPC est une méthode efficace pour encoder la parole à des débits binaires de 2,4kbits/ seconde et moins, sont connus depuis longtemps. Historiquement, la recommandation LPC-10, utilisée depuis la fin des années 1970, elle était la première norme de codage de la parole à 2,4 kbps. Le LPC-10 a été renommé plus tard Standard fédéral FS-1015. Dans ce procédé, l'enveloppe spectrale de la parole est décrite avec un filtre de synthèse LPC, tandis que la structure du spectre est produite en utilisant un train d'impulsions

périodiques en tant que filtre d'excitation pour la parole, et le bruit blanc comme excitation pour non exprimé le discours. Depuis sa création, la qualité de la parole du vocodeur LPC a considérablement augmenté, démontré parle standard FS-1015 [18].

II.3.1. Norme fédérale américaine 1015

Le codeur de parole FS-1015 est le LPC-10 à 2,4 kbps, il conforme à la norme fédérale américaine qui a été créé à partir de la fin des années 1970. Il a été normalisé par le Département américain de la défense (DoD), puis par l'Organisation du Traité de l'Atlantique Nord (OTAN), avant de devenir une norme fédérale américaine en 1984. Il a toujours été conçu pour les terminaux vocaux sécurisés Deux petites différences entre LPC-10 et la plupart des autres LPC sont que LPC-10 utilise l'extracteur de pitch de la fonction de différence d'amplitude moyenne (AMDF) et que la méthode de covariance est utilisée pour calculer les coefficients de prédicteur à court terme plutôt que la méthode d'autocorrélation [18].

Il y a deux autres différences distinctives. La première est que le signal d'excitation n'est pas une impulsion mais une excitation standardisée, comme indiqué dans le tableau (II.1). Les excitations non impulsionnelles répartissent l'énergie de manière plus uniforme au cours d'une période tonale, réduisant ainsi le rapport crête à valeur efficace et améliorant la parole synthétisée pour une précision de convertisseur N/A fixe. Une deuxième différence importante est le bloc de correction de pitch et de son, qui est utilisé pour lisser le pitch et le son de plusieurs trames. La principale limite des performances du LPC est une bonne extraction et l'extension du son. En LPC-10, un algorithme de programmation dynamique permet d'utiliser les informations de trame adjacentes pour obtenir des contours de pitch et de ton plus raisonnables. Cependant, bien que cela améliore la détermination de pitch et de la tonalité, le délai du codeur est également considérablement augmenté.

Le LPC-10 a constitué une véritable avance à son époque, au début des années 1970, mais il souffre de la limitation selon laquelle l'excitation doit tomber dans l'une des deux classes voisées ou non voisées. Comme tout discours ne peut pas être classé de cette manière, la qualité du discours synthétisé n'est souvent pas naturelle. Étant donné que l'excitation est quasi-périodique pour la parole vocale et que l'excitation est un bruit aléatoire, elle peut être trop bruité si l'excitation est trop périodique ou trop "essoufflée" si un segment partiellement voisée est non facturé. Ceci est simplement une limitation du modèle LPC. Au fil des années, de nombreuses tentatives ont été faites pour remédier à cette lacune [19].

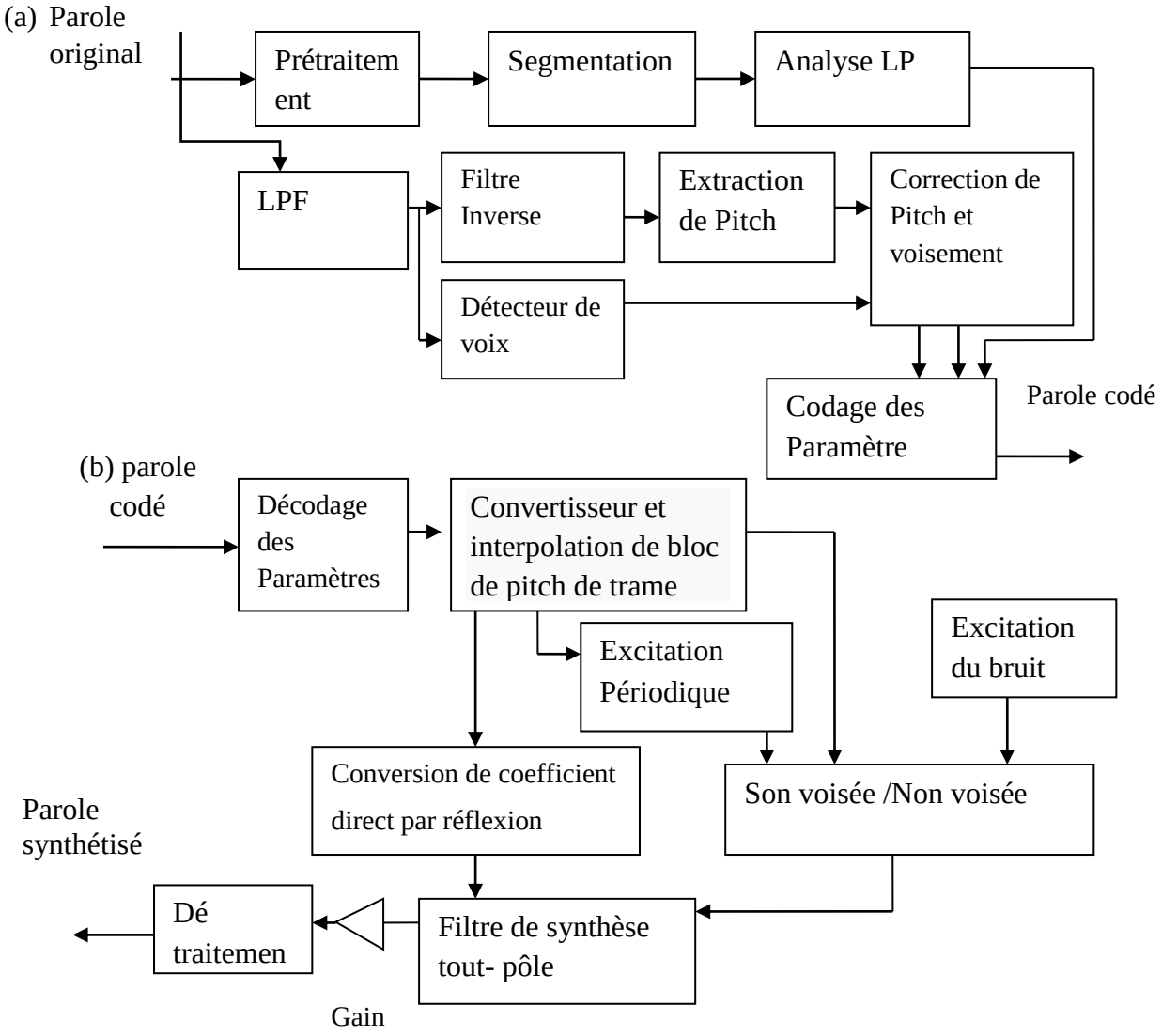


Figure. II.1: Le standard FS-1015:(a) LPC-10 codeur, (b) LPC-10 décodeur [20].

II.3.1.1. Prétraitement de la parole

La parole est pré-accentuée avec un filtre RII du premier ordre avec la fonction de transfert suivante:

$$\mathbf{H_{pre\_filter}(z) = (1 - \frac{15}{16} z^{-1})} \qquad \textbf{(II.1)}$$

Le but de ce filtre est d'améliorer la stabilité numérique de l'analyse LP. La forme d'onde de la parole présente généralement un affaiblissement à haute fréquence. La réduction de cette atténuation diminue la plage dynamique du spectre de puissance de la parole en entrée, ce qui permet une meilleure modélisation des caractéristiques dans les régions à haute fréquence du spectre de la parole [17].

II.3.1.2. Analyse LP

La norme LPC-10 spécifie qu'une méthode de covariance avec stabilisation du filtre de synthèse doit être utilisée pour déterminer le spectre LP de la parole. Cependant, la plupart

des implémentations modernes utilisent plutôt une approche d'autocorrélation en raison de son amélioration.

La stabilité numérique et l'efficacité du calcul n'affecte rien à l'interopérabilité du vocodeur. FS-1015 est favorable à une analyse LP synchrone de pitch. Cela signifie que la position de la fenêtre d'analyse LP est ajustée par rapport à la phase des impulsions de pitch. Cette conception améliore la régularité de la parole synthétisée [17].

II.3.1.3. Estimation de Pitch

Le LPC-10 autorise un pitch compris entre 50 et 400 Hz. L'estimation de pitch est obtenue comme suit.

- Filtre passe-bas du signal de parole.
- Filtre inverse de signal vocal avec une approximation du second ordre optimale.

Le prédicteur d'ordre 10 est déterminé par l'analyse LP.

- Calculez la valeur minimale de la fonction de différence de magnitude (MDF). Le

Le MDF n'est pas aussi précis que l'autocorrélation dans la détermination de pitch tonal, mais il a été choisi principalement pour des raisons liées à l'efficacité du calcul. Sur de nombreuses architectures, le coût de calcul d'une multiplication était d'un ordre de grandeur supérieur à celui d'une addition. Sur ces architectures, le MDF serait nettement plus [21].

II.3.1.4. Détection de voix

Un classificateur de modèle linéaire est utilisé pour effectuer la tâche de détection de voix. Les vecteurs de configuration sont composés des paramètres suivants:

- Énergie à faible bande.
- Rapport max / min du MDF.
- Taux de passage à zéro.
- Premier coefficient de réflexion.
- Deuxième coefficient de réflexion.
- Gains de prédiction de pitch.
- Rapport d'énergie pré-accentué

Enfin, un algorithme de contrôle est appliqué à la décision de voix. Cet algorithme est essentiellement un contrôleur médian modifié, qui prend en compte la force d'expression de chaque trame. Cela évite l'apparition de segments à une seule voix dans des segments non voisés [22].

II.3.1.5. Quantification des paramètres LP

Les deux premiers coefficients de réflexion sont convertis en rapports logarithmiques car statistiquement, seuls les deux premiers CR ont une probabilité significative d'être proches

de1. Dans LPC-10, un schéma de quantification scalaire très simple est utilisé, utilisant des codes différents pour chaque paramètre LP. Chaque code est optimisé pour le paramètre LP particulier qu'il est destiné à coder. Deux schémas de quantification différents sont utilisés, en fonction du résultat de la décision d'expression. Celles-ci sont détaillées dans le tableau (II.1).

Le choix des paramètres à quantifier, à savoir les LAR pour les deux premiers paramètres et les coefficients de réflexion par la suite est probablement motivé par les résultats déjà obtenue, à savoir que les LAR sont supérieurs pour le codage des deux premiers paramètres mais ensuite il présente aucun avantage substantiel sur les coefficients de réflexion [22].

Fréquence d'échantillonnage	8 KHZ
Ordre du LPC prédicteur	10 pour les sons voisés 4 pour les sons non voisés
Débit de données	2400bps
Longueur de trame	22.5 ms
Bits assignés /Trame	54
Pitch	AMDF méthode [51,3-400 HZ] Codage :60 valeurs
Gain	RMS valeur codage : 32 valeurs
Analyse LPC	Semipitch synchrone
Méthode d'analyse	Covariance
Synthèse	Pitchsynchrone
Excitation	Forme d'onde stockée pour une trame voisée

Tableau. II.1: Caractéristique principales de la norme fédérale FS-1015[19].

II.4. Prévision linéaire en boucle ouverte et LPC-10

Cette section décrit les algorithmes du système source qui utilisent l'analyse en boucle ouverte pour déterminer la séquence d'excitation. Les vocodeurs prédictifs linéaires en boucle ouverte sont essentiellement les vocodeurs LP de première génération. Le LPC-10 est un bon exemple d'algorithme utilisant l'analyse en boucle ouverte.il utilise un prédicteur du 10ème ordre pour estimer les paramètres du tractus vocal. La segmentation et le traitement de trame dans LPC-10 dépendent de la voix. Les informations sur le pitch tonal sont estimées à l'aide de la fonction de différence d'amplitude moyenne (AMDF). La voix est estimée à l'aide de mesures d'énergie, de mesures du passage par zéro et du rapport maximum / minimum du facteur AMDF. Le signal d'excitation pour la parole voisée dans le LPC-10 consiste en une séquence qui ressemble à une impulsion glottale échantillonnée. Cette séquence est définie dans la norme et l'apériodicité est créée par un processus de répétition d'impulsions synchrone. Le LPC-10 produit un discours synthétique acceptable. La complexité est estimée à 5 à 7 MIPS [23].

Le conditionnement variable de la taille et de trame a été ajouté. Le prototype initial d'analyse par le processus de vocodeur produit trois types de trame vocale comme défini dans le tableau (II.3).

Type de trame	Rms valeur	Longueur (bits)
Sons voisés	>min Energie	54
Sons non voisés	>min Energie	34
Silence	>min Energie	0

Tableau. II.2: LPC -10 Types de trame à taux variable [24].

II.5. Le codeur LPC10 à 2.4 kbit/s

Ce codeur c'est la base des codeurs de la parole actuelle, car il a un grand intérêt pédagogique mais plus aucun intérêt pratique. Considérons le schéma de la figure (II.2), où **x(n)**est le signal de parole original, **y(n)** le signal en sortie du filtre d'analyse, **$\hat{y}$(n)** l'entrée du filtre de synthèse et $\hat{x}(n)$le signal de parole reconstruit.

Le codeur LPC10 :

- Calcule les coefficients du filtre à partir de signal original :

$$\mathbf{A(z) = 1 + a_1 z^{-1} + apz^{-p}} \qquad \textbf{(II.2)}$$

- Détermine l'entrée du filtre de synthèse en respectant les exigences de débit 2.4 kbit/s avec une bonne qualité possible.
- En supposant que le signal est localement stationnaire dans chacune des fenêtres d'analyse qui sont d'une vingtaine de ms. Selon l'équation suivante:

$$\underline{\mathbf{x}} = [\mathbf{x(0)} \ldots . \mathbf{x(N-1)}]^{\mathbf{t}} \quad \textbf{(II.3)}$$

Pour un signal de parole échantillonné à Fe = 8 kHz : N = 160 dans chacune de ces fenêtres.

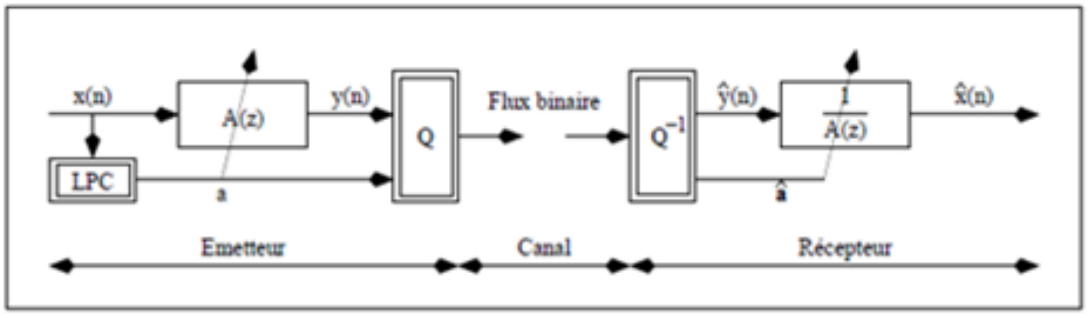


Figure. II.2: schéma très général d'un codeur de parole [25].

II.5.1. Cas des sons non voisés

La théorie de la prédiction linéaire affirme que si x(n) est considéré comme la réalisation d'un processus aléatoire AR d'ordre P0, il existe un filtre de fonction de transfert A(z) totalement blanchissant dès que son ordre P devient supérieur ou égal à P0.

la densité spectrale de puissance de l'erreur de prédiction est égale à :

$$\mathbf{S_Y(f)} = \boldsymbol{\delta}_{\mathbf{Y}}^{\mathbf{2}} \quad \textbf{(II.4)}$$

Comme on sait aussi que $\mathbf{S_Y(f)} = |\,\mathbf{A(f)}\,|^{\mathbf{2}}\mathbf{S_X(f)}$ **(II.5)**

En deduit que $\mathbf{S_X(f)} = \frac{\boldsymbol{\delta}_{\boldsymbol{Y}}^{\mathbf{2}}}{|\mathbf{A(f)}|^{\mathbf{2}}}$ **(II.6)**

Supposons que l'on choisisse comme entrée du filtre de synthèse une réalisation quelconque d'un bruit blanc de puissance

$$\boldsymbol{\delta}_{\hat{\boldsymbol{Y}}}^{\mathbf{2}} = \boldsymbol{\delta}_{\boldsymbol{Y}}^{\mathbf{2}} \quad \textbf{(II.7)}$$

On voit alors qu'on a la propriété suivante au niveau des densités spectrales de puissance.

$$\mathbf{S_{\hat{X}}(f)} = \frac{\boldsymbol{\delta}_{\boldsymbol{Y}}^{\mathbf{2}}}{|\mathbf{A(f)}|^{\mathbf{2}}} = \mathbf{S_X(f)} \quad \textbf{(II.8)}$$

En supposant que le signal de parole puisse être considéré comme la réalisation d'un processus aléatoire AR, on rétabli un signal qui a la même distribution de la puissance en

fonction de la fréquence mais les formes d'onde sont différentes. On reconstruit par cet combinaison un signal qui est perçu semblable au signal original car l'oreille est à peine sensible à des changements de phase [25]. On appelle ce type de codeur : un vocodeur, il analyse les principales composantes spectrales de la voix et fabrique un son synthétique à partir de résultat de cette analyse.

II.5.2. Cas des sons voisés

Les tracés de la figure (II.3) sont relatifs à un son voisé où l'on montre aussi bien dans le domaine temporel (à gauche) que dans le domaine fréquentiel (à droite) le signal original x(n) et l'erreur de prédiction y(n). Le filtre A(z) n'est manifestement pas totalement blanchissant [25].

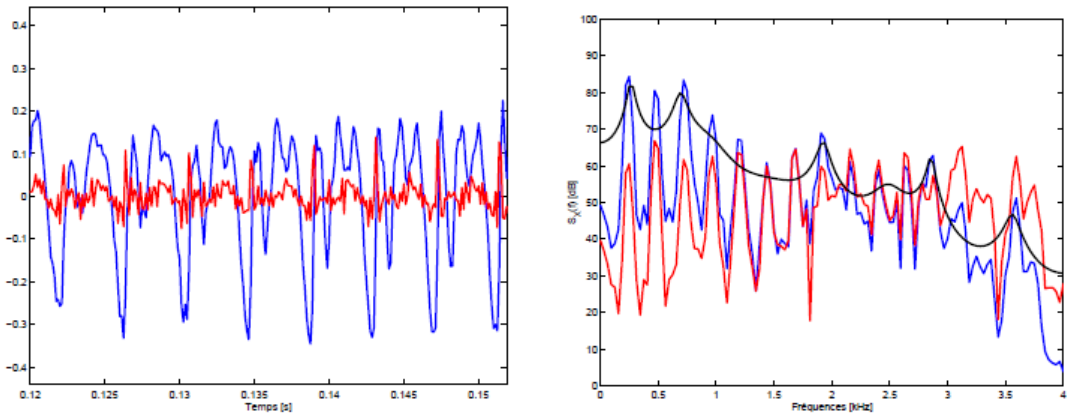


Figure. II.3: cas de sons voisés.

Il reste dans le signal ***y(n)*** une périodicité assez marquée, visible aussi bien dans le domaine temporel que dans le domaine fréquentiel. Pendant une durée de 32 ms, on a approximativement 7.5 périodes. La fréquence fondamentale est donc de l'ordre de :

$$f_0 \approx 7.5 \div 0.032 \approx 235HZ.$$

On observe bien dans le domaine fréquentiel un spectre de raies avec une fréquence fondamentale de 235 Hz (correspondant à un locuteur féminin) et les différents harmoniques.
Le problème qui se pose maintenant est de trouver un modèle ŷ(n) pour y(n) qui permette par filtrage d'obtenir :

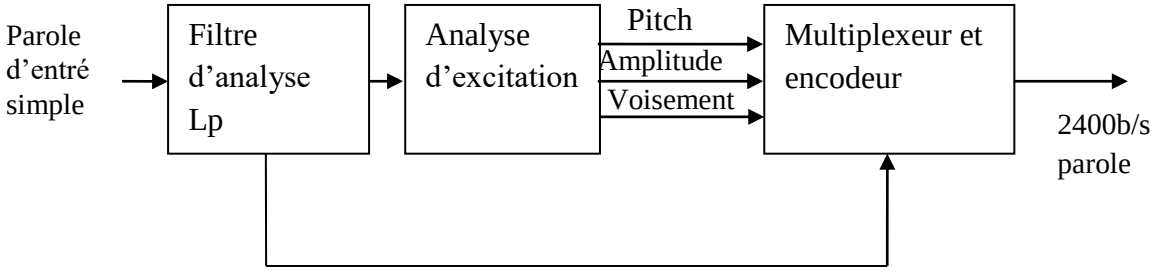
$$\boldsymbol{S_{\hat{X}}(f) \approx S_X(f)} \qquad \textbf{(II.9)}$$

Et qui soit très économe en débit. Un peigne de la forme est un bon candidat.

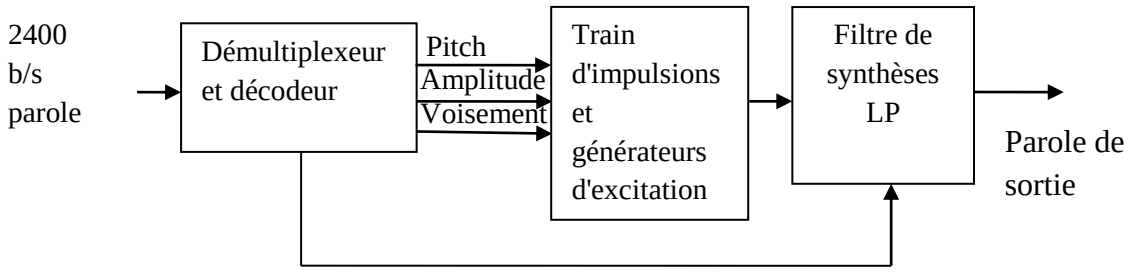
$$\boldsymbol{\hat{y}(n) = \alpha \sum_{m=-\infty}^{+\infty} \lambda(n - mT_0 + \varphi)} \qquad \textbf{(II.10)}$$

Dans cette expression, λ(n) est le symbole de Kronecker ,T0 est la période fondamentale exprimée en nombre d'échantillons et **φ** une valeur appartenant à {0…, T_0-1} traduisant notre incertitude sur la phase. Le signal ŷ(n) peut être alors interprété comme la réalisation d'un processus aléatoire ***ŷ (n)***.

Un schéma fonctionnel du vocodeur LPC-10 2.4 kbps est présenté dans la Figure (II.4) :



(a) LPC-10 encodeur



(b) LPC-10 décodeur

Figure. II.4: LPC-10 :Processus de la parole [24].

II.6. Allocation de bits

L'affectation de bits pour le codeur LPC-10 est indiquée dans le tableau (II.4). Les valeurs de période tonale sont codées sur 7 bits. La puissance est codée sur 5 bits. Il en résulte 41 bits pour les segments voisés et 20 bits pour les segments non voisés. Les bits de trame non facturés sont protégés par une protection contre les erreurs coûtant 21 bits / trame. La longueur de la trame est de 22,5 ms. Pour une parole de 54 bits / trame, cela donne un débit de 2,4 kbps pour le LPC-10.

Les bits codés sont 54 (y compris un bit de synchronisation) pour une trame de parole avec 180 échantillons. Pour 8 kHz de fréquence d'échantillonnage, 180 échantillons par trame, qui est de 22,5 ms par trame (180/8000 = 22.5 ms). Pour chaque provenant du codeur de 22,5 ms, 54 bits binaires codés sont envoyés sur le canal. Le débit binaire de codeur est 2400 bit / s ou 2,4 kb / s (54 bits / 22,5 ms = 2,4 kb / s). Le taux de compression est de 26,7 par rapport à

64 kb/s MIC (64 / 2.4). LPC-10 principalement utilisé dans les communications radio avec les transmissions vocales sécurisées. La qualité de la voix est faible dans sa nature (plus de son mécanicien), mais avec l'intelligibilité raisonnable. Certaines variantes de LPC-10 explore différentes techniques (par exemple, le sous-échantillonnage, de détection de bruit, LP variables bits codés) afin d'obtenir des débits binaires de 2400 bits /s [17] [20].

Paramètre	Sons Voisé	Non voisé
LPC	41	20
PITCH/parole voisée	7	7
Puissance	5	5
Synchronisation	1	1
Protection contre les erreurs	-	21
Totale	54	54

Tableau. II.3: allocation de bits pour le codeur FS-1015 [20].

II.7. Conclusion

Le codeur marqué FS-1015 sur les codeurs paramétriques est un élément essentiel des progrès accomplis en matière de codage à très bas débit, Il s'agit du vocodeur basé sur la prédiction linéaire, également appelé LPC-10 qui est le plus étudié dans le codage à très bas débit, fonctionnant à 2,4 Kbps avec une qualité de parole subjective jugée acceptable, il a été utilisé pour les communications vocales sécurisées. L'algorithme LPC-10 utilise un prédicteur tous pôles du 10ème ordre et s'appuie sur le modèle d'excitation de source à deux états (voisé et non voisé). Le principal inconvénient de LPC10e réside dans la décision difficile de basculer entre voix et segments non-voisés.

La qualité de la parole produite par le LPC-10 est également synthétique en ce sens qu'elle semble artificielle. Une des limitations fondamentales du codeur LPC est la classification stricte d'une trame de parole en deux classes : voisé et non voisé. Le modèle d'excitation à deux états a été amélioré ultérieurement par le modèle source d'excitation mixte.

Le prochain chapitre est réservé à une étude d'un circuit LPC à excitation mixte de 2,4 kb /s, appelé MELP, qu'est devenu le nouveau standard fédéral américain.

Chapitre III : Mise en œuvre de codeur MELP a 2.4 kbps

III.1. Introduction

Le codeur LPC utilise un modèle entièrement paramétrique pour encoder les informations importantes du signal de parole. Le codeur produit des informations parole éligible à un débit binaire voisin de 2400 bps. Cependant, un modèle au cœur du codeur LPC génère des artefacts gênants tels que comme bourdonnements, bruit de tons; ce qui est dû aux nombreuses limitations de modèle LPC.

L'algorithme de codage de la parole MELP à 2,4 kbps a été mis au point par le gouvernement américain en tant que norme suivante pour les communications vocales sécurisées à bande étroite parfaitement interopérables pour les applications stratégiques et tactiques. Le MELP est basé sur le codage prédictif linéaire du modèle de parole FS-1015, il est conçu pour surmonter certaines des limitations de LPC.

Il inclut des caractéristiques supplémentaires, telles que l'excitation mixte, les impulsions apériodiques, la dispersion des impulsions, l'amélioration spectrale adaptative et la mise à l'échelle de l'amplitude de Fourier de l'excitation voisée.

III.2. Le modèle de production de la parole de MELP

Une tentative pour améliorer le modèle FS-1015 présenté dans un schéma fonctionnel du codeur MELP à la figure (III.1). Cependant, les deux modèles partagent certaines similitudes fondamentales; comme le fait que les deux utilisent un filtre de synthèse pour traiter un signal d'excitation de manière à générer le discours synthétique.

Une des limitations fondamentales du LPC est la classification stricte d'une trame de discours en deux classes : voisé et non voisé.

Le modèle MELP utilise une excitation mixte formée de la somme d'une composante impulsionnelle et d'une composante bruit. La composante impulsionnelle est formée d'un train d'impulsions périodique ou apériodique commandé par la gigue de période (pour introduire une fluctuation de la période du pitch), qui est un nombre aléatoire uniformément distribué entre ± 25% de la période du pitch. Cette excitation est une excitation multi-bande avec une intensité de voisement (qui mesure la quantité de voisement) définie pour chaque bande de fréquence. Ceci est l'idée principale du codeur MELP qui est basée sur des observations pratiques où le signal résiduel (erreur de prédiction) est une combinaison d'un train d'impulsion avec du bruit. Ainsi, le modèle MELP est beaucoup plus réaliste que le modèle LPC-10 (FS-1015), où l'excitation est soit un train d'impulsions ou du bruit [17].

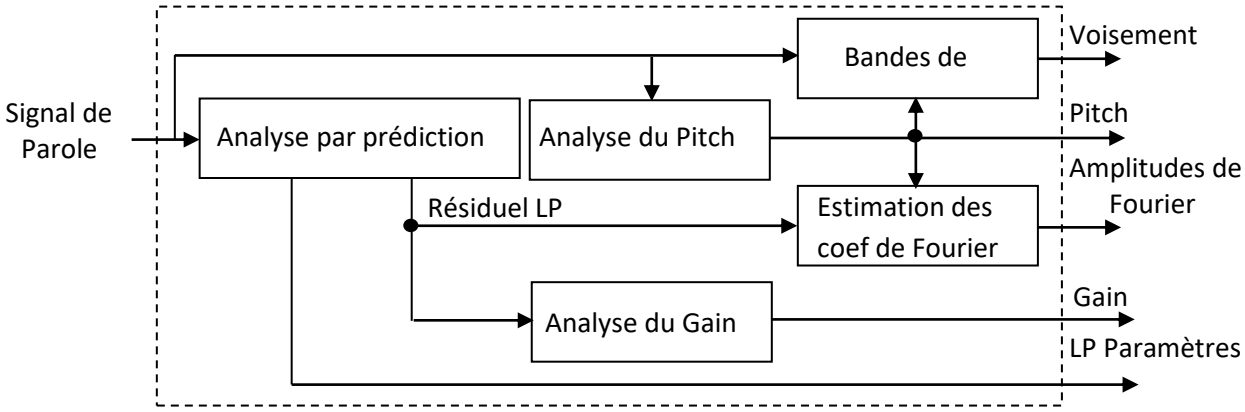


Figure.III.1. Schéma base de codeur MELP [26].

Le modèle MELP utilise une excitation mixte formée de la somme d'une composante impulsionnelle et d'une composante bruit [26].

III.3. Description du standard MELP à 2.4 Kbps

Les opérations effectuées au niveau du l'encodeur et du décodeur sont décrites brièvement ci-dessous.

III.3.1. L'encodeur

Le codeur LPC à Excitation Mixte (MELP) est basé sur un modèle paramétrique, qui inclut cinq fonctionnalités améliorées comparativement aux codeurs LPC. Celles-ci sont :

- Une excitation mixte
- Une impulsion apériodique
- Amélioration spectrale adaptative
- Un filtre de dispersion d'impulsions
- Une Modélisation par les amplitudes de Fourier

L'excitation mixte est formée de la somme d'une composante impulsionnelle et d'une composante de bruit en utilisant un mixage de filtres adaptatifs multi bandes en vue de réduire le bruit introduit par le vocodeur LPC classique. Lorsque le signal de parole est voisé, le codeur MELP synthétise ce signal en utilisant soit un train d'impulsions périodique ou des impulsions apériodiques. Ces dernières sont souvent utilisées dans les zones des transitions c'est-à-dire situées entre les segments voisés et les segments non-voisés du signal de parole ceci permet d'améliorer, lors du décodage, la reproduction des impulsions glottiques.

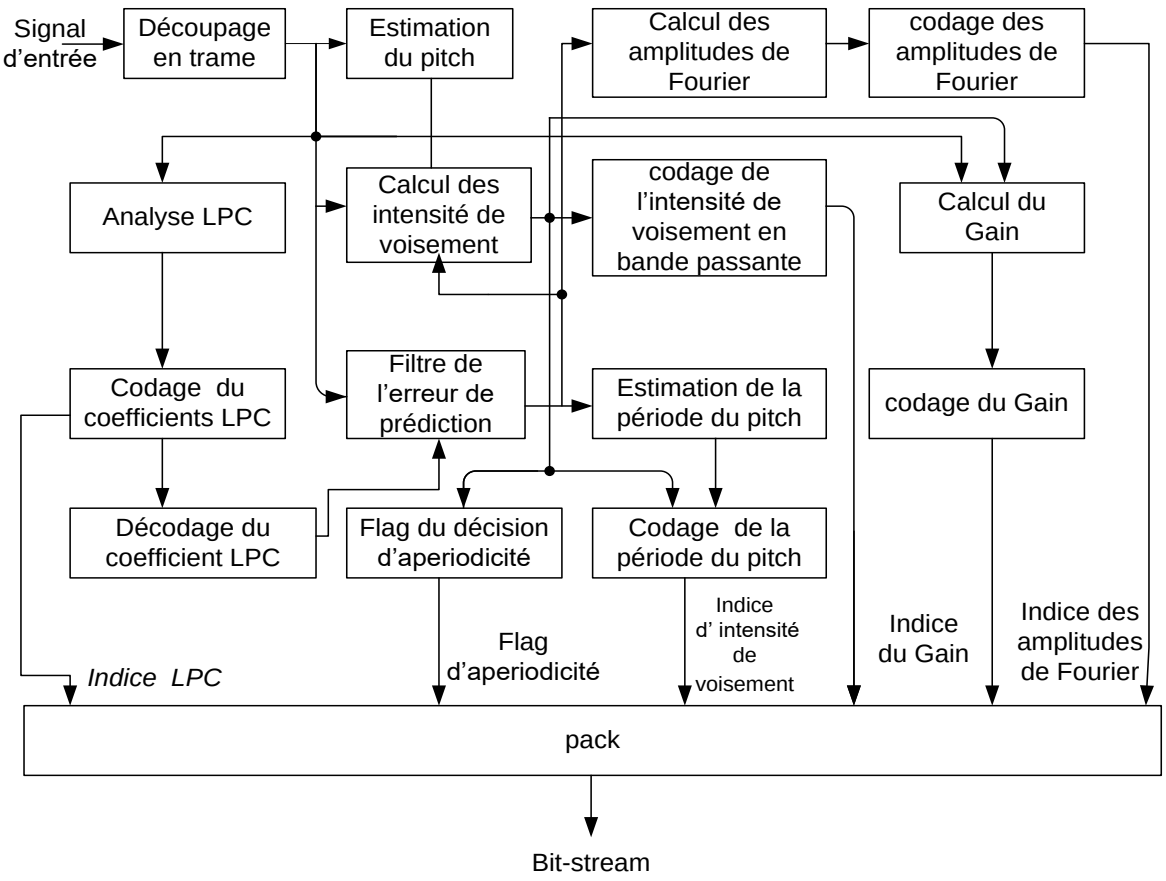


Figure.III.2: Schéma bloc du codeur MELP [17].

III.3.1.1. Prétraitement et calcul de pitch

Le prétraitement de la parole consiste uniquement en une suppression de toute énergie relative aux basses fréquences qui peut être présente dans le signal d'entrée. La parole d'entrée est filtrée avec un filtre passe-haut de Tchebychev du 4ème ordre avec une fréquence de coupure de 60Hz, le but de ce filtre est d'éliminer les composantes extrêmement basses fréquences du signal d'entrée qui sont inaudibles et qui gêneraient l'estimation des Paramètres [22].

Avant de poursuivre l'estimation de la période de pitch, les positions de quelques fenêtres de traitement des signaux lecture relatives aux trames de signal sont décrites en premier lieu. Figure (III.3) résume les principales fenêtres utilisées par le codeur MELP. Comme le FS1015Codeur LPC, chaque trame est composée de 180 échantillons. Les positions de ces fenêtres d'analyses sont sélectionnées pour faciliter l'interpolation, car les paramètres d'une parole donnée sont interpolés entre deux ensembles différents, calculés à partir des fenêtres d'analyse.

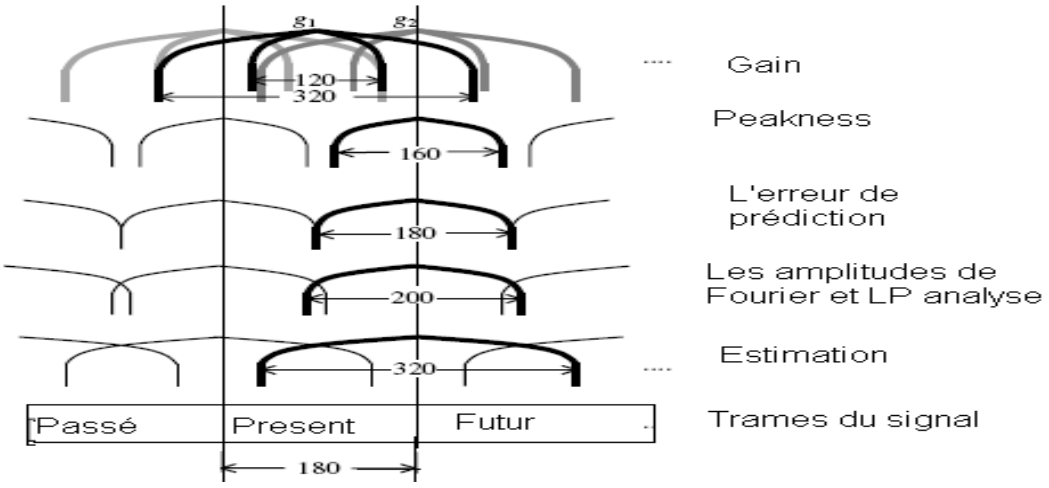


Figure.III.3: Positions des fenêtres [17].

La forme d'onde de la parole est filtrée avec un filtre passe-bas Butter Worth du 6ème ordre avec une fréquence de coupure de 1 kHz. Ce filtre supprime les composantes de fréquence supérieure qui ne sont pas nécessaires à l'estimation de la période de pitch mais qui peuvent interférer avec l'algorithme d'estimation de pitch [22].

L'estimation de la période de pitch entière est calculée en déterminant le maximum d'autocorrélation normalisée du segment de parole. La valeur entière du pitch, P_1, est la valeur de τ = 40, 41…, 160, pour laquelle la fonction d'autocorrélation normalisée, r (τ), est maximale. Cette fonction est définie par :

$$\boldsymbol{r(\tau) = \frac{c_\tau\,(0,\tau)}{\sqrt{c_\tau(0.0)c_\tau\,(\tau,\tau)}}} \qquad \textbf{(III.1)}$$

Ou : $\boldsymbol{c_\tau}$ est l'autocorrélation normalisé de segment de la parole

Avec :

$$\boldsymbol{c_\tau(m,n) = \sum_{k=-[\tau/\,2]-80}^{-[\tau/\,2]+79} s_{k+m}s_{k+n}} \qquad \textbf{(III.2)}$$

Et [τ] représente le décalage en nombre entier d'échantillons. L'échantillon S_0 dans l'équation est pris comme centre de la fenêtre d'analyse du pitch. Pour le calcul du nombre entier pitch, cette fenêtre est centrée sur le dernier échantillon dans la trame courante [17].

III.3.1.2. Filtre de mise en forme et l'affinement du pitch fractionnaire :

Chaque filtre de mise en forme se compose de cinq filtres, appelé les filtres de synthèse, puisqu'ils sont utilisés pour synthétiser le signal d'excitation mixte. Chaque filtre de synthèse commande une bande de fréquence particulière, ils sont de type Butter Worth d'ordre 6 avec les bandes passantes définies par 0-500, 500-1000, 1000-2000, 2000-3000, et 3000-4000 Hz. Les réponses de ces cinq filtres sont commandées par les cinq intensités de

voisement. En variant les intensités de voisement avec le temps, une paire de filtres variant dans le temps en résulte [17].

Le pitch réel de la forme d'onde de la parole continue, il n'est pas une valeur entière suggérée par l'autocorrélation à temps discret, mais un nombre réel. Pour cette raison, des erreurs d'échantillonnage peuvent provoquer des imprécisions dans l'analyse de pitch. Ces effets peuvent être minimisés en effectuant le suivi de la période de pitch sur la forme d'onde de la parole à un taux d'échantillonnage beaucoup plus élevé. Cependant, cela entraînerait une charge de calcul nettement plus importante et il est beaucoup plus efficace d'effectuer l'interpolation sur la fonction d'autocorrélation. La formule d'interpolation suivante est utilisée [22]:

Tout d'abord, la composante fractionnaire de la période de pitch est calculée:

$$\Delta = \frac{c_T(0,T+1)c_T(T,T) - c_T(0,T)c_T(T,T+1)}{c_T(0,T+1)[c_T(T,T) - c_T(T,T+1)] + c_T(0,T)[c_T(T+1,T+1) - c_T(T,T+1)]} \quad \textbf{(III.3)}$$

L'autocorrélation normalisée correspondant à la valeur fractionnaire de pitch est donné par :

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

Les équations (III.3) et (III.4) produisent un décalage fractionnaire et l'autocorrélation normalisée correspondante qui seraient obtenus si le signal d'entrée avait été linéairement interpolé pour obtenir des valeurs entre les temps d'échantillonnage actuels [17].

III.3.1.3. Indicateur d'apériodicité :

Une partie essentielle du modèle MELP est celle des impulsions partiellement apériodiques lors de l'excitation vocale. Le codeur signale que l'excitation est apériodique au moyen du drapeau de gigue [22]. L'indicateur d'apériodicité est mis à 1 si l'intensité de voisement de la bande la plus basse $Vbp_1 < 0.5$ et mis à 0 autrement. La valeur Vbp_1 est déterminée par l'analyse de voisement.

Pour une force de voix élevée (supérieure à 0,5), la trame est fortement périodique et le drapeau est mis à zéro. Pour une périodicité faible, le drapeau est mis à 1, signalant au décodeur MELP de générer des impulsions apériodiques sous forme d'excitation vocale (voix nerveuse). L'indicateur est transmis dans le flux de bits MELP en utilisant un bit [17].

III.3.1.4. Analyse linéaire de prédiction :

Une analyse LP au 10ème ordre est effectuée sur la trame vocale. Cela se fait de manière standard.

- Fenêtre du cadre avec une fenêtre de Hamming.
- Calculer l'autocorrélation.
- À l'aide de la récursivité Levinson-Durbin, calculez les coefficients LP.

Une analyse LP du 10 éme ordre est effectuée sur le signal vocal d'entrée en utilisant une fenêtre de Hamming à 200 échantillons (25 ms) centrée sur le dernier échantillon de la trame actuelle. La méthode d'autocorrélation est utilisée avec l'algorithme de Levinson-Durbin. Les coefficients résultants sont étendus avec une largeur de bande constante de 0,994 (15 hertz) est appliqué aux coefficients de prédiction, a_i, i = 1, 2,…, 10, où chaque coefficient est multiplié par 0.994i [17]. Les Coefficients de la prédiction linéaire sont convertis en LSF et quantifiés par MSVQ, le code-book de MSVQ se compose de quatre étapes dont les indices ont 7, 6, 6, et 6 bits, respectivement. Le vecteur LSF quantifié est la somme des vecteurs choisis par le processus de recherche, avec un vecteur choisi parmi chaque étape. Les indices de ces quatre vecteurs sont transmis. Les coefficients sont quantifiés et utilisés pour calculer le signal d'erreur de prédiction [17].

III.3.1.5. Calcul du signal résiduel et son peakiness :

Le résidu LP est calculé en filtrant la trame avec l'inverse du filtre LP. Le filtre inverse est mis en œuvre de la manière habituelle, par inversion de la fonction de transfert du filtre tous pôles LP [22]. La fenêtre résiduelle est centrée sur le dernier échantillon dans la trame courante et elle est prise suffisamment large pour être utilisée lors du calcul final de pitch. Le pic du signal d'erreur de prédiction est calculé sur une fenêtre de 160 échantillons centrée sur le dernier échantillon de la trame actuelle.

La valeur du "peakiness" est montré dans l'équation suivante :

$$\boldsymbol{peakiness} = \frac{\sqrt{\frac{1}{60}\sum_{n=1}^{160} r_n{}^2}}{\frac{1}{60}\sum_{n=1}^{160}|r_n|} \qquad \textbf{(III.5)}$$

Si le "peakiness" dépasse 1.34, alors l'intensité de voisement de la bande la plus basse, Vbp_1, est forcée à 1.0. Si le "peakiness" dépasse 1.6, alors l'intensité de voisement des trois bandes les plus basses, Vbp_i, i = 1, 2, 3, sont forcées à 1.0. C'est la seule utilisation du "peakiness"[17].

III.3.1.6. Calcul de pitch final :

Le but du calcul de pitch final est d'utiliser les informations du signal résiduel du prédicteur et du signal d'entrée pour déterminer une valeur de hauteur finale pour la trame. D'abord, le signal résiduel LP est filtré par passe-bas. Où le filtre est un Butterworth d'ordre

6, avec une fréquence de coupure de 1 kHz. L'estimation de pitch de ton fractionnaire est effectuée sur ce signal résiduel filtré passe-bas et sur le signal de parole d'origine, produisant deux candidats pitch et force vocale pour la trame. Le candidat qui a la force de voix la plus élevée est utilisée pour le reste de l'algorithme et sera appelé dorénavant le signal de pitch [22].

La procédure de contrôle du doublement du pitch est exécutée sur le signal résiduel filtré, elle permet de rechercher et corriger les valeurs du pitch qui sont des multiples du pitch réel. La vérification du doublage de pitch procède en évaluant l'autocorrélation normalisée du signal de pitch en fractions de la valeur de pitch. Si un pitch candidat (qui est une fraction de pitch actuel) est suffisamment "bonne", c'est-à-dire que le segment en question possède une autocorrélation suffisamment élevée pour cette valeur de pitch), il remplace alors le pitch actuel en tant que pitch finale de la trame [22].

III.3.1.7. Calcul du gain

Le gain du signal de la parole d'entrée est mesuré deux sous trames en utilisant une longueur de fenêtre adaptative au pitch, l'une centrée sur le centre de la trame d'analyse et référencée G1, l'autre centrée sur la fin de la trame d'analyse et nommée G2. Le calcul du gain pour G_1 et est centré 90 échantillons avant le dernier échantillon dans la trame courante. Le calcul pour G_2 et est centré sur le dernier échantillon dans la trame courante. Le gain est la valeur de RMS, mesuré en dB, du signal S_n dans la fenêtre [17] :

L'équation pour le calcul du gain est comme suit :

$$G_i = 10log(0.01 + \frac{1}{L}\sum_{n=1}^{L} S_n{}^2) \quad \textbf{(III.6)}$$

Où L : est la longueur de la fenêtre.

III.3.1.8. Magnitudes de Fourier

Les magnitudes de Fourier sont évaluées à l'aide de la magnitude FFT du résidu LP, ajouté à 512 échantillons ; La valeur de chaque amplitude de Fourier est la valeur maximale de la FFT dans la moitié d'une période de pitch provenant de chaque harmonique de pitch. Si la période de pitch est petite et la fréquence donc élevée, le pitch n'aura pas nécessairement 10 harmoniques dans la largeur de bande de Nyquist. Dans ce cas, les harmoniques de pitch qui sont supérieures à la fréquence de Nyquist sont définies sur 1. Le vecteur des amplitudes de Fourier est normalisé pour avoir une amplitude euclidienne (valeur efficace) de 1 [22].

III.3.1.9. Quantification des paramètres

D'abord, les coefficients de la prédiction linéaire, a_i i = 1, 2,…, 10, sont convertis en fréquences de raies spectrale (LSF: line spectral frequencies). Le prédicteur linéaire est quantifié en tant que fréquences du spectre linéaire. Les LSF sont obligés d'être dans l'ordre croissant et ont une séparation minimale de 50 Hz. Le vecteur LSF résultant est quantifié en utilisant un MSVQ à quatre étapes. Le livre de codes comprend une première étape de 128 niveaux et trois étapes suivantes de 64 niveaux. Le vecteur quantifie, $\hat{f}$, est la somme des vecteurs choisis par le processus de recherche, avec un vecteur choisi parmi chaque étape la métrique de quantification utilisée est une simple distance euclidienne pondérée entre les vecteurs LSF quantifiés et non quantifiés [22].

$$\boldsymbol{d^2(f,\hat{f}) = \sum_{i=1}^{10} w_i (f,\hat{f})^2} \quad \textbf{(III.7)}$$

Où : d= la distance entre les vecteurs LSF

Avec :

$$\boldsymbol{w_i = \begin{cases} p(f_i)^{0.3} & ,1 \le i \ge 8 \\ 0.64p(f_i)^{0.3} & ,i = 9 \\ 0.16p(f_i)^{0.3}, & i = 10 \end{cases}} \quad \textbf{(III.8)}$$

Ou : w_i = magnitude de fourier

III.3.1.9.1. Quantification de pitch

La période de pitch final T et l'intensité de la bande passante vs1 sont quantifiées conjointement sur 7 bits. Si vs1 est à 0.6, la trame est non facturée et le code tout à zéro est envoyé. Sinon, le log T est quantifié avec un quantificateur uniforme à 99 niveaux allant du log 20 au log 160. L'indice résultant est mappé sur le code approprié en fonction d'une table.

Le vs1 quantifié noté qvs1 est égal à 0 pour l'état non voisée et à 1 pour l'état voisée. Aucune transmission séparée n'est nécessaire car les informations peuvent être récupérées à partir de la même table [17].

III.3.1.9.2. Quantification du gain

Les deux paramètres du gain G1 et G2 sont transmis pour chaque trame. G2 est quantifiée sur 5 bits à l'aide d'un quantificateur uniforme à 32 niveaux compris entre 10 et 77 dB. La seconde valeur de gain (G1) est quantifiée sur 3 bits à l'aide d'un algorithme de quantification adaptatif. L'algorithme compare le G2 de la trame précédente et de la trame

actuelle. Si les deux niveaux de gain diffèrent de moins de 5 dB et que la valeur mesurée de G¡ diffère de la moyenne des deux G, les valeurs de moins de 3 dB, un indice spécial est alors envoyé pour indiquer que G est approximativement la moyenne des 2 G2. Sinon, G1 est quantifié à l'aide d'un quantificateur uniforme à sept niveaux [22].

III.3.1.9.3. Quantification des amplitudes de Fourier

Les magnitudes de Fourier sont quantifiées à l'aide d'un quantificateur vectoriel standard à une étape, avec une métrique de distance euclidienne pondérée utilisant des poids fixes pour mettre en valeur les basses fréquences. Les valeurs des poids sont données par:

$$\boldsymbol{w_i = \left[\frac{117}{25+75(1+1.4(\frac{f_i}{1000})^2)^{0.69}}\right]^2 \quad , i = 1, 2, 3 \ldots \ldots 10} \qquad \textbf{(III.9)}$$

où $f_i = 8000i / 60$ est la fréquence en Hertz correspondant au $i^{ième}$ harmonique pour une période du pitch par défaut de 60 échantillons. Les poids sont appliqués à la différence carrée entre les amplitudes de Fourier de l'entrée et les valeurs du code-book [22].

III.3.1.10. Allocation des bits

Le tableau récapitule le schéma d'allocation de bits du codeur MELP. Les LPC sont quantifiés en tant que LSF à l'aide de MSVQ. La synchronisation est un modèle alternant : un / zéro. Une protection contre les erreurs est fournie pour les trames non voisées, utilisant 13 bits. Un total de 54 bits est transmis par trame, avec une longueur de trame de 22,5 ms. Un débit binaire de 2400 bps [17].

Paramètres	**MELP a 2.4 kbps**	
Fréquence d'échantillonnage Taille de la trame Débit en trame	8khz 180 échantillons (22.5 ms) 44.44 trame/seconde	
Mode de voisement	V	N/V
10 LSF	25	25
Pitch	7	7
10 Amplitudes de Fourier	8	-
5 Bandes de voisement	4	-
2 Gains	8	8
Flag	1	-
Protection	-	13
Synchronisation	1	1
Total de bits par trame	**54 bits**	
Débit total	**54*44.44=2400 bps**	

Tableau.III.1: Allocation des bites des codeurs MELP de 2.4 kbps [27].

III.3.2. Le décodeur

L'entrée du décodeur est un train de bits et la sortie est un signal de parole synthétisé. La figure 3.4 montre le schéma fonctionnel du décodeur MELP, dans lequel le train de bits est décompressé avec les index destinés au décodeur correspondant. En comparant avec la figure 3.1nous pouvons voir que le modèle de production de parole est intégré à la structure du décodeur. Deux filtres supplémentaires sont ajoutés tout au long du traitement de chemin: le filtre de rehaussement spectral prend en entrée l'excitation mixte et le filtre de dispersion des

impulsions à la fin de la chaîne de traitement. Ces filtres améliorent la qualité de la perception du discours synthétique.

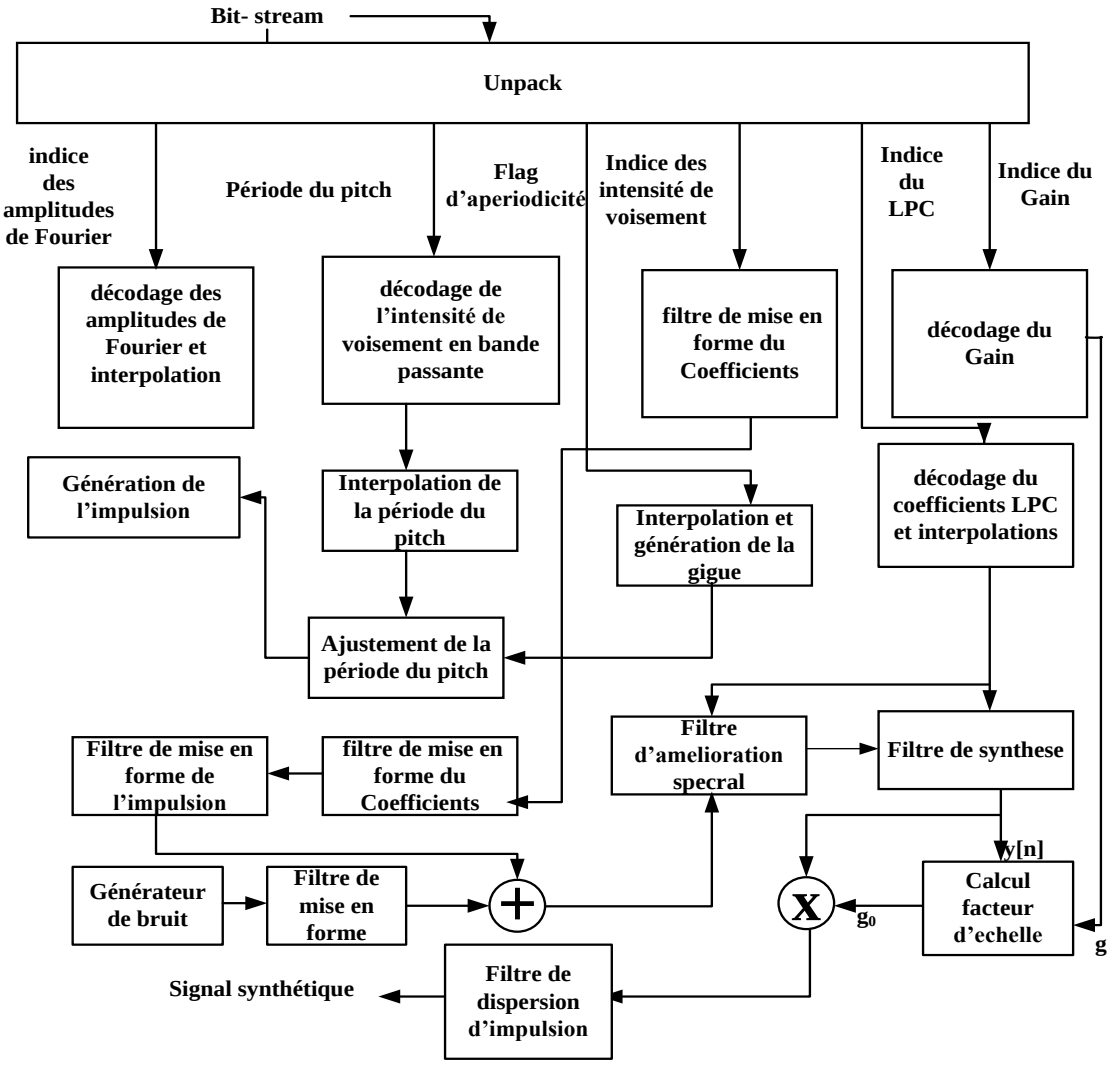


Figure.III.4: schéma bloc du décodeur MELP [17].

III.3.2.1. Décodage des paramètres et interpolation

Dans le décodage MELP, les paramètres du flux binaire sont décompactés et décodés conformément au schéma approprié. Ces paramètres sont :

- LPC (LSF).
- Période de pitch
- La force de la parole en bande basse (Vbpi)
- La force de la bande passante de la parole (Vs)
- Gains (g1 et g2)
- Période de flag

- Les magnitudes de Fourier.

Après décodage des paramètres reçus, le décodeur interpole les différents paramètres de manière synchrone au pitch. Pour les trames non voisées (détectables à partir du code de la période du pitch/intensité de voisement de la bande la plus basse), des valeurs par défaut pour certains paramètres sont utilisés : 50 pour la période du pitch, 0.25 pour la gigue, les amplitudes de Fourier sont initialisées à 1, et les intensités de voisement sont mises à 0. Les valeurs par défaut sont nécessaires pour les trames non voisées puisque l'interpolation linéaire est effectuée de manière synchrone au pitch. Le signal d'excitation est synthétisé une fois par période de pitch. Chaque impulsion (de l'excitation impulsionnelle) est obtenue sur une période de pitch et possède une certaine forme différente d'une impulsion idéale. La forme de l'impulsion est capturée par les amplitudes de Fourier. Les amplitudes de Fourier sont utilisées pour générer la réponse impulsionnelle du filtre de génération d'impulsion, responsable de la synthèse de l'excitation impulsionnelle. Un nouveau paramètre est introduit : La gigue. Elle est utilisée uniquement dans le décodeur pour contrôler la quantité de caractère aléatoire pendant la génération d'excitation apériodique. Lorsque l'indicateur d'périodicités (flag) est positionné, une gigue est appliquée à la valeur de la période fondamentale. La valeur de la gigue est assignée comme suite : gigue est à 0.25 si la période de flag est égale à 1 ; autrement, gigue sera à 0. Cette possibilité d'excitation impulsionnelle non périodique est particulièrement intéressante pour les zones de transitions entre sons [17].

III.3.2.2. Génération d'excitation mixte

L'excitation mixte est produite comme la somme de l'impulsion filtrée et des excitations de bruit. L'excitation impulsionnelle, $e_p(n)$, n = 0, 1,…, T-1, est calculée par la transformée de Fourier discrète inverse d'une période du pitch:

$$\boldsymbol{e_p(n) = \frac{1}{t}\sum_{k=0}^{T-1} M(k) e^{j2\pi nkT}} \qquad \textbf{(III.10)}$$

Ou : $e_p(n)$= l'excitation impulsionnelle

La séquence d'impulsions de l'échantillon T générée à partir des magnitudes de Fourier a une valeur efficace unitaire (Rms). Cette séquence est filtrée par le filtre de mise en forme d'impulsion et ajoutée à la séquence de bruit filtrée pour former l'excitation mixte. Le bruit est généré par un nombre aléatoire uniforme moyen nul ayant une valeur efficace unitaire. Les coefficients des filtres sont interpolés de manière synchrone [17].

III.3.2.3. Filtre d'amélioration spectrale

Le filtre de l'amélioration spectrale adaptative est appliqué pour le signal d'excitation mixte. Ses coefficients sont produits par l'expansion de largeur de bande de la fonction de transfert linéaire du filtre de la prédiction linéaire, $A(z^{-1})$, correspondant aux LSF interpolés. La fonction de système est donnée par :

$$\boldsymbol{H(z) = (1 - \mu z^{-1})} \ \frac{\mathbf{1+\sum_{i=1}^{10} a_i \ B^i Z^i}}{\mathbf{1+ \sum_{i=1}^{10} a_i \ \propto^i Z^i}} \qquad \textbf{(III.11)}$$

Où ai : les coefficients de prédiction linéaire.

Les paramètres $\boldsymbol{\mu}$, $\boldsymbol{\alpha}$ et $\boldsymbol{\beta}$ sont rendus adaptatifs en fonction des conditions du signal. Le filtre de rehaussement spectral, comme son nom l'indique, a pour seul objectif : d'améliorer la qualité de la perception du discours synthétique en ressortant les caractéristiques spectrales originales [17].

III.3.2.4. Filtre de synthèse LP

Il s'agit d'un filtre de synthèse de formants sous forme directe, avec les coefficients correspondant aux LSF interpolés.

III.3.2.5. Calcul du facteur d'échelle

La puissance de sortie du filtre de synthèse doit être égale au gain interpolé de la période en cours. L'excitation étant générée à un niveau arbitraire, un facteur d'échelle 'g0' est calculé afin d'échelonner la sortie du filtre de synthèse pour produire le niveau approprié. Ceci est donné par l'équation suivante [17] :

$$\boldsymbol{g_0 = \frac{10^{g/20}}{\sqrt{\frac{1}{T}\sum_n y^2\,[n]}}} \qquad \textbf{(III.12)}$$

En multipliant 'g0' par y [n], la séquence d'échantillon T résultante aura une puissance de $\mathbf{10^{g/20}}$, ou **g** est en dB [17].

III.3.2.6. Filtre à dispersion d'impulsion

C'est le dernier bloc de la chaîne de décodage. Le filtre est un filtre FIR à 65 prises dérivé d'une impulsion triangulaire à aplatissement spectral. Il s'agit presque d'un filtre passe-tout où les changements de réponse en amplitude sont relativement faibles.

Ce filtre améliore la correspondance entre les formes d'onde de voix synthétiques et naturelles filtrées par passe-bande dans les régions qui ne contient pas de résonance de formants. On peut penser à la fonction du filtre de "remuer" légèrement le spectre de la sortie du filtre de synthèse afin d'améliorer le naturel; cela est bénéfique car une excitation mixte est

formée en combinant le bruit et l'impulsion à travers un groupe de filtres à largeur de bande fixe [17].

III.4. Conclusion

Dans ce chapitre, nous avons présenté en détails le principe du codeur MELP à 2.4 kbps et ces différents blocs. On a confirmé sa classification au tant que codeur paramétrique puisqu'il ne préserve pas la forme d'onde. Le codeur MELP est basé sur le modèle LPC classique. L'ajout de certaines étages (étage excitation, étage impulsion apériodique, étage amélioration spectrale adaptatif, filtre de dispersion d'impulsion, modélisation par des amplitudes de fourriers) a donné naissance avec des performances plus robustes.

Chapitre IV : Résultat et discussion de simulation

IV.1. Introduction

La manière la plus performante pour évaluer la qualité de la parole transmise par un système de télécommunication est de recourir sur des tests d'écoutes, des tests durant lesquels les participants, qu'on soumet à écouter uniquement des échantillons vocaux donnent leurs jugements sur une échelle de qualité bien définie. Cependant, cette méthode s'avère coûteuse en temps et en moyens financiers.

Des méthodes de mesures dites « objectives », fondées soit sur une analyse du signal de parole soit sur des informations issues du réseau, sont développées. Actuellement des méthodes de plus en plus « sophistiquées » sont apparues, connues sous le nom de modèles perceptifs. Les modèles les plus performants sont normalisés au sein de l'Union Internationale des Télécommunications (UIT) tel que le modèle PESQ qui permet d'évaluer la qualité d'écoute dans de nombreuses conditions de dégradation.

Notre objectif est de synthétiser la parole avec une bonne perception afin de comparer la synthèse de la parole à très bas débit. Pour cela, nous avons utilisé les deux codeurs qui fonctionnent à 2.4 kbps LPC-10 et le MELP.

Ce chapitre est divisé en deux parties essentielles. Dans la première partie on a simulé le codeur MELP à 2.4 kbps en utilisant le langage C (Builder C++ 6.0) pour la partie programmation et Matlab pour les représentations. La deuxième partie est consacrée pour la phase d'évaluation ; notre choix s'est porté sur les mesures PESQ comme mesures perceptuelles en raison de leur bonne corrélation avec les tests subjectifs.

Une comparaison de performance entre le codeur LPC-10 et MELP est effectuée avec la mesure objective PESQ en utilisant des échantillons extraits des bases de données de deux langues différentes (arabe, anglais). Ces bases de données vocales varient pour des voix masculines féminine.

IV.2. Présentation du logiciel d'évaluation PESQ

En 2001, UIT-T a rendu public son système d'évaluation perceptuelle de la qualité vocale (Perceptual Evaluation of Speech Quality, PESQ) [ITU-862]. PESQ est un outil d'évaluation de la qualité de la parole transmise pas un système de télécommunication. Ce système a un caractère perceptuel, il est connu pour sa facilité d'utilisation et la fiabilité de ses résultats tout en réduisant les coûts de mesure.

Sur la base des données numériques des appréciations, une opinion moyenne de la qualité d'écoute MOS est ensuite calculée. Elle permet de déterminer le degré de satisfaction concernant une certaine ligne téléphonique. La valeur MOS résultante est la valeur moyenne

de tous les écouteurs pour chacun des discours testés. Elle s'applique sur toute forme de dégradation qui peuvent trouvée dans la parole testée, par exemple la limitation de la bande passante, le bruit additif, l'écho, la distorsion non linéaire, etc [28].

La valeur du MOS est calculée par :

$\mathrm{MOS} = \frac{\sum_i N_i i}{N}$ i= {1,…,5}

N : nombre d'auditeurs ayant participé au test.

Ni : nombre d'auditeurs qui ont choisi la catégorie i .

i : 1 jusqu'à 5 sachant que i est un nombre naturel.

Pour donner un aperçu sur la position des limites entre la qualité « bonne » et « basse » et entre celle « basse » et celle « inacceptable » la recommandation P. 862 donne les intervalles de la valeur de PESQ suivants sur une échelle allant de -0,5 (dégradation très gênante) à 4,5 (dégradation imperceptible). [29]:

- PESQ donne une valeur numérique entre 0(aucune similitude) et 4.5 (signaux identiques), qui simule la perception humaine de la qualité de la parole.
- Des scores PESQ entre 3 et 4.5 désignent une qualité très acceptable (avec un 3.8 comme seuil de la qualité dans les systèmes téléphoniques traditionnels), niveau qu'on va appeler qualité bonne.
- Des valeurs entre 2.5 et 3 indiquent une qualité acceptable entre 2 et 2.5, niveau de qualité est dit bas. Un effort est alors nécessaire pour la compréhension. Dans ce, cas on va référer à une qualité dite basse.
- Des valeurs inférieures à 2 signifient que la dégradation a rendu la communication très difficile ou impossible. En d'autre terme, l'intelligibilité est perdue ; par conséquent, la qualité est inacceptable.

IV.2.1. Algorithme PESQ :

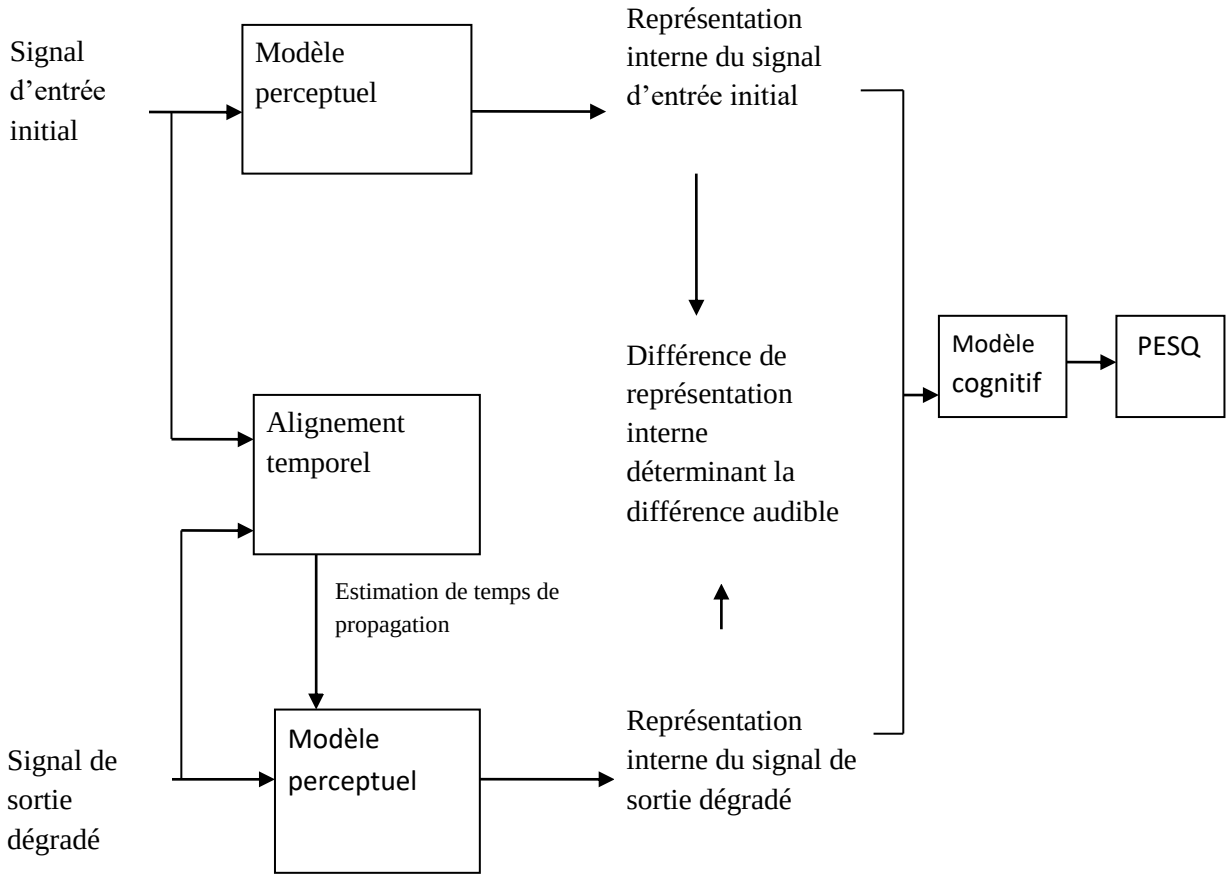


Figure. IV.1: principe de fonctionnement de modèle PESQ d'apres UIT-T [30].

L'algorithme PESQ réalise un alignement temporel dynamique entre le signal d'origine et le signal dégradé à l'aide d'un modèle informatique, comme illustré sur la figure 4.1. Cette étape s'appelle l'algorithme d'alignement temporel Il consiste à effectuer d'abord une estimation du retard de l'ensemble du signal, puis en effectuant un alignement sur des sections du signal défini par un silence, appelées énonciations [18].

Le modèle informatique remplace le sujet (humain). Ce modèle est constitué de deux modèles. Le premier est le modèle de perception responsable de l'extraction des paramètres de la parole.et le second est le « modèle cognitif » qui permet de prendre en compte le fait qu'une dégradation n'a pas le même impact selon qu'elle est additive, soustractive, ou selon son contexte (segment de parole ou non) et sa distribution (localisée ou non) ce que rend le jugement réel. Le modèle PESQ transforme le signal dégradé et le signal d'origine correspondant en représentations internes, puis utilise leur différence pour calculer une note de qualité d'écoute [27].

IV.3. Description de corpus de parole utilisé

Les signaux de test ont été pris à partir de plusieurs bases de données dont la fréquence d'échantillonnage est de 8 KHz, des échantillons extraits des bases de données de test pour les langues anglais (TIMIT) et arabe. Les phrases sont prononcées par des locuteurs féminins et masculins.

La stratégie de test peut être résumée comme suit : Afin de tester ces codeurs nous avons utilisé un corpus de deux langues : arabe, anglais qui est composé de phrases phonétiquement équilibrées. Les 20 premières phrases [enregistrés par plusieurs personnes] pour la langue arabe et la langue anglaise sont prononcées par 10 locuteurs féminins et 10 locuteurs masculins.

IV.4. Résultats de simulation

Nous commençons de premier principe par le codeur LPC-10 à 2.4 kbps suivi du codeur MELP a 2.4 kbps. Pour ce faire nous avons simulé le corpus cité précédemment, pour illustrer ce travail nous allons présenter la simulation de deux phrases prononcées par un locuteur et une locutrice pour les deux langues (Arabe et Anglais).

Les figures en dessous montrent le déroulement de la simulation pour les codecs MELP et LPC-10réalisé avec le langage C, suivi des figures qui représentent les signaux (original, synthétique et résiduel) sous Matlab.

En premier lieu, on a introduit un fichier son de type (.**wav**) enregistré dans la base de données cité précédemment. Pour avoir en sortie un flux binaire qui sera ensuite pris comme entrée du décodeur afin de restituer notre signal synthétique.

IV.4.1. Codec LPC-10 à 2.4 Kbps

Voilà l'exécution d'une simulation du codeur FS1015 à 2.4 Kbps sous C++.

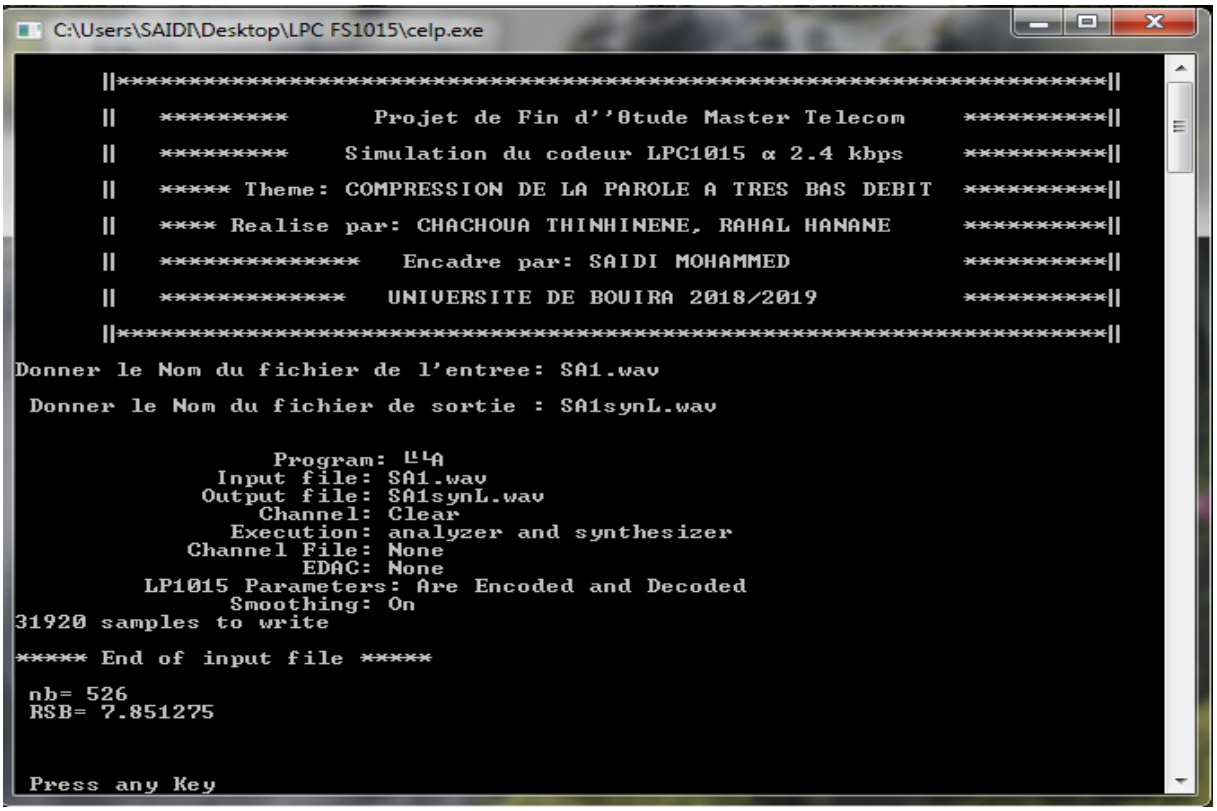


Figure. IV.2 : l'exécution de la simulation du codeur FS-1015 sous le C++.

1)- Langue Arabe :

Les deux figures en dessous représentent le résultat de la simulation sous MATLAB pour une phrase arabe prononcée par locuteur et une locutrice.

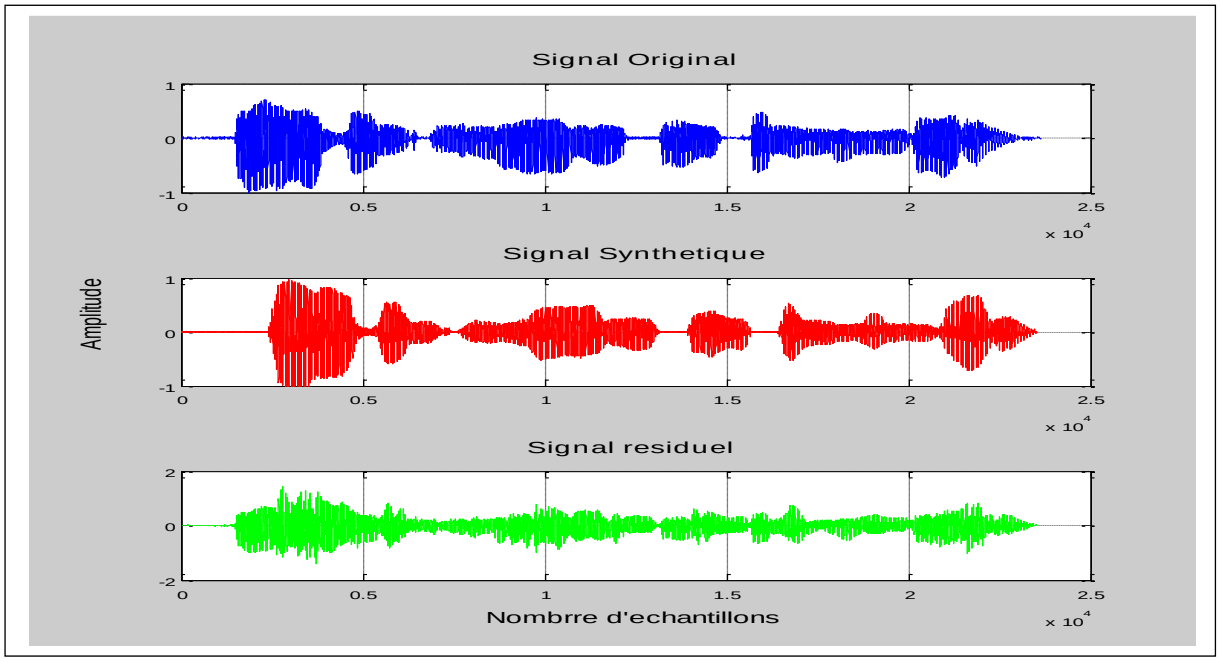


Figure. IV.3 : Phrase prononcée par un locuteur « صعد الإمام فوق المنبر ».

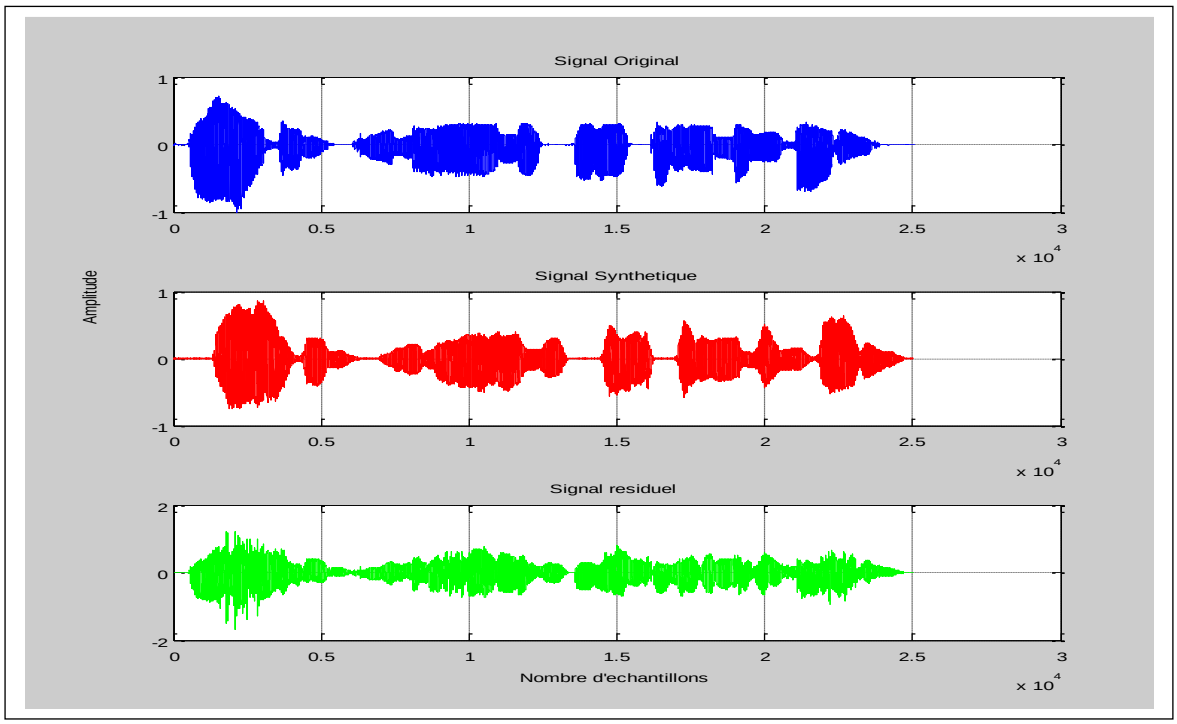


Figure. IV.4: Phrase prononcée par une locutrice « صعد الإمام فوق المنبر ».

2)- Langue Anglaise :

Les deux figures en dessous représentent le résultat de la simulation sous MATLAB pour une phrase en anglais prononcée par locuteur et une locutrice.

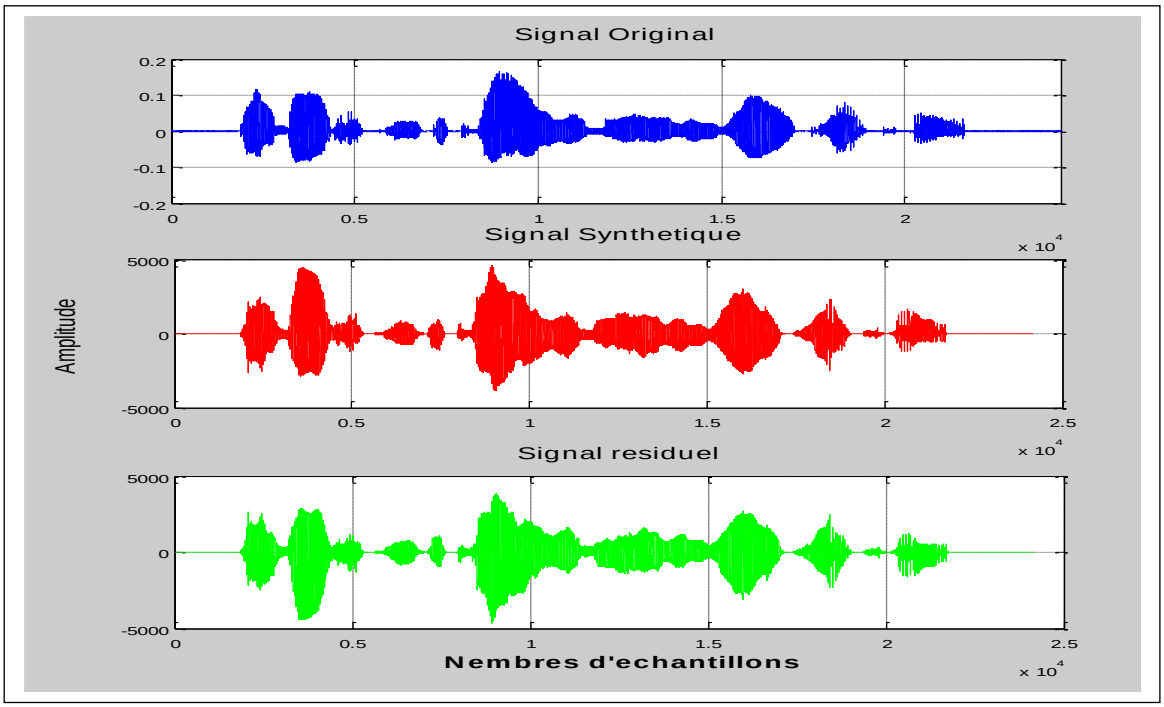


Figure. IV.5: Phrase prononcée par un locuteur ‘‘don’t ask me to carry an oily rag like that”.

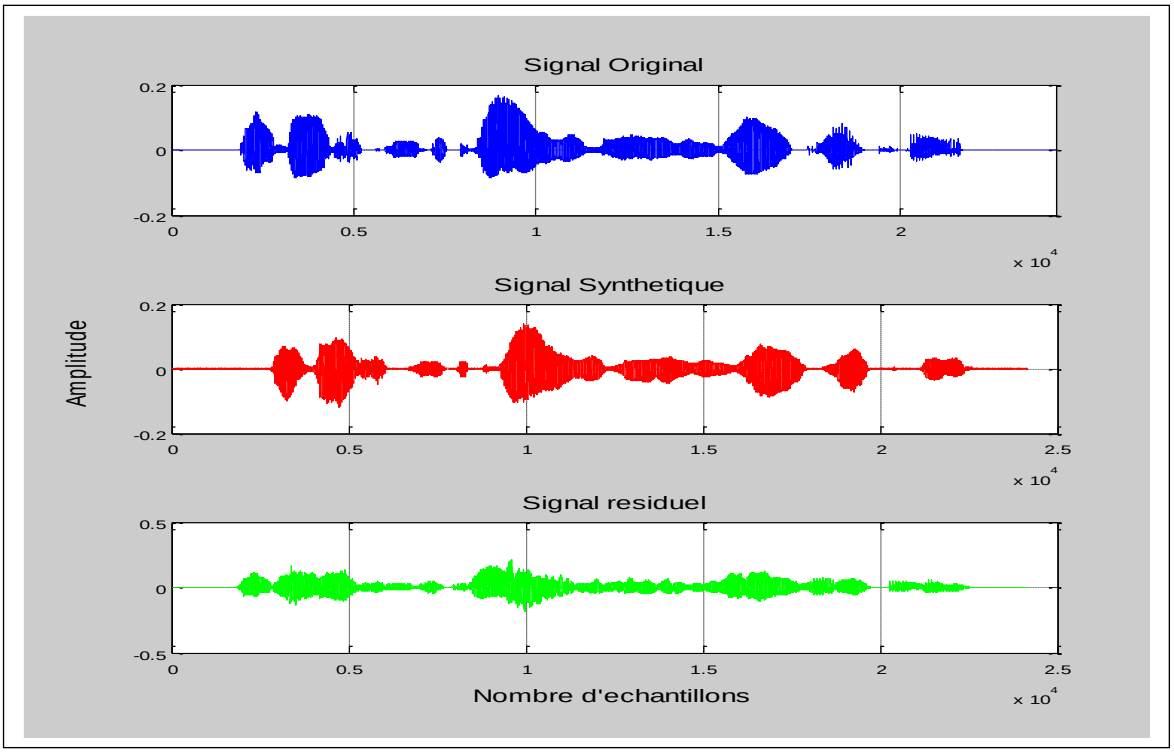


Figure. IV.6: Phrase prononcée par une locutrice "don't ask me to carry an oily rag like that.

IV.4.2. Codec MELP à 2.4 Kbps

Voilà l'exécution d'une simulation du codeur MELP à 2.4 Kbps sous C++.

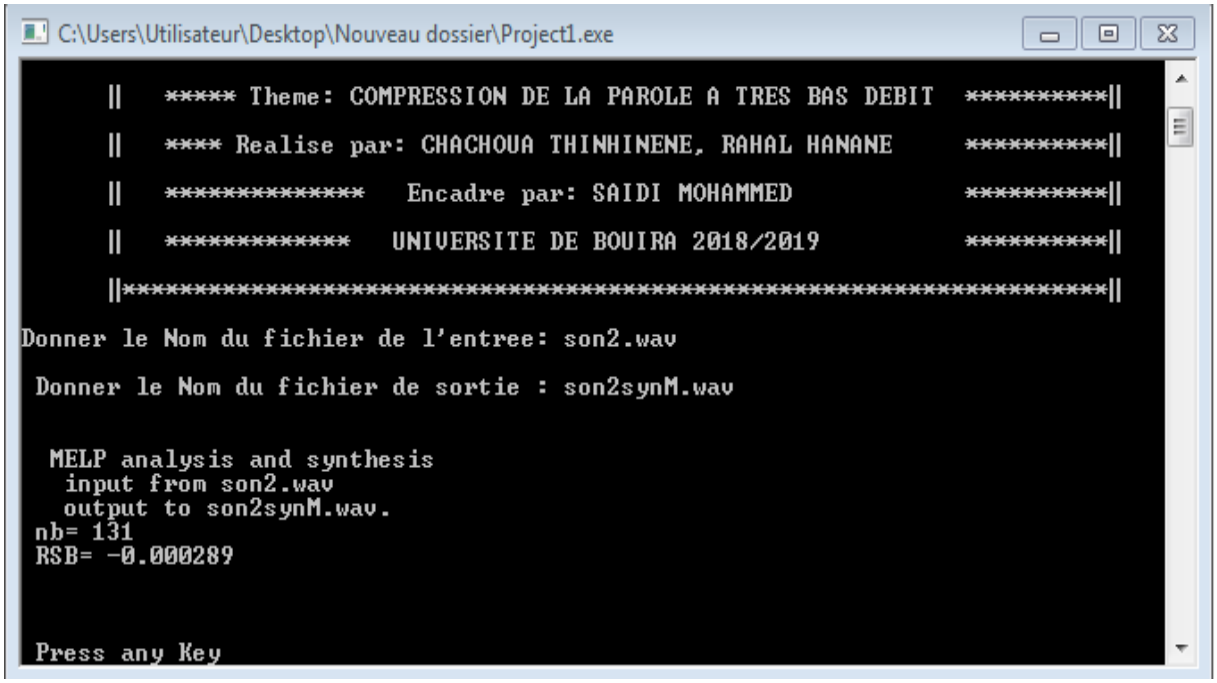


Figure. IV.7 : l'exécution de la simulation du codeur MELP sous le C++.

1)- Langue Arabe :

Les deux figures en dessous représentent le résultat de la simulation sous MATLAB pour une phrase en arabe prononcée par locuteur et une locutrice.

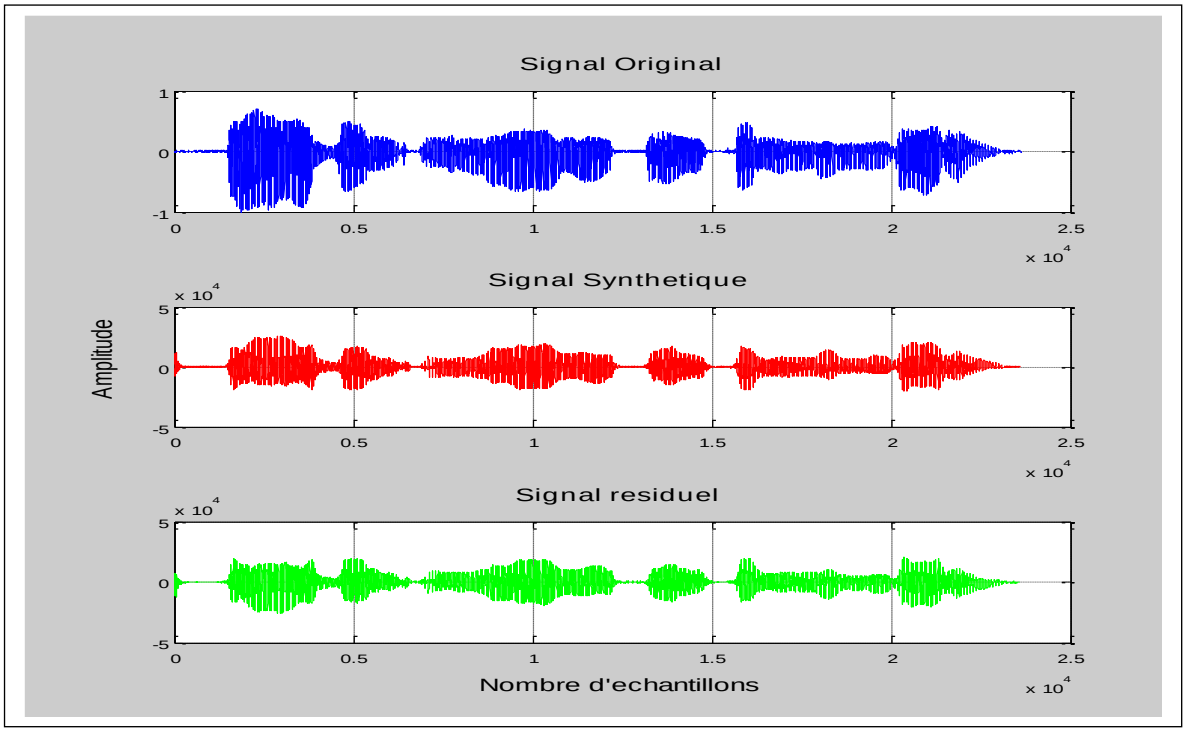


Figure. IV.8 : Phrase prononcée par un locuteur « صعد الإمام فوق المنبر ».

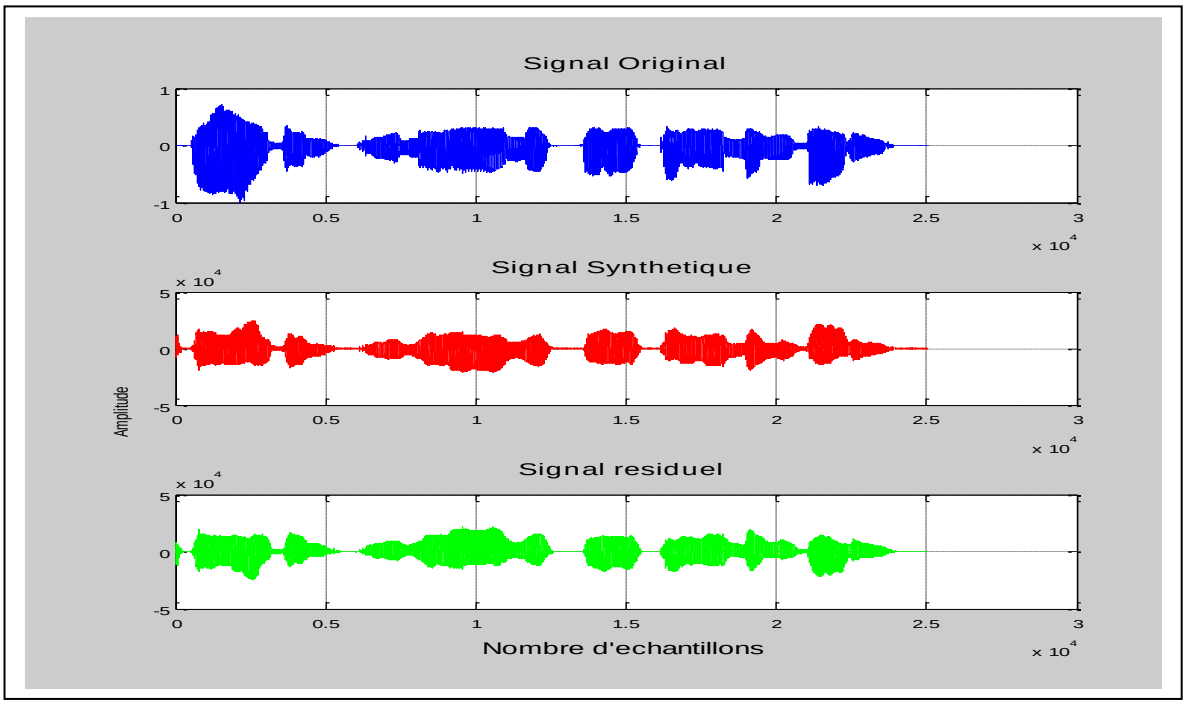


Figure. IV.9 : Phrase prononcée par une locutrice « صعد الإمام فوق المنبر ».

2) - Langue Anglaise:

Les deux figures en dessous représentent le résultat de la simulation sous MATLAB pour une phrase en anglais prononcée par locuteur et une locutrice.

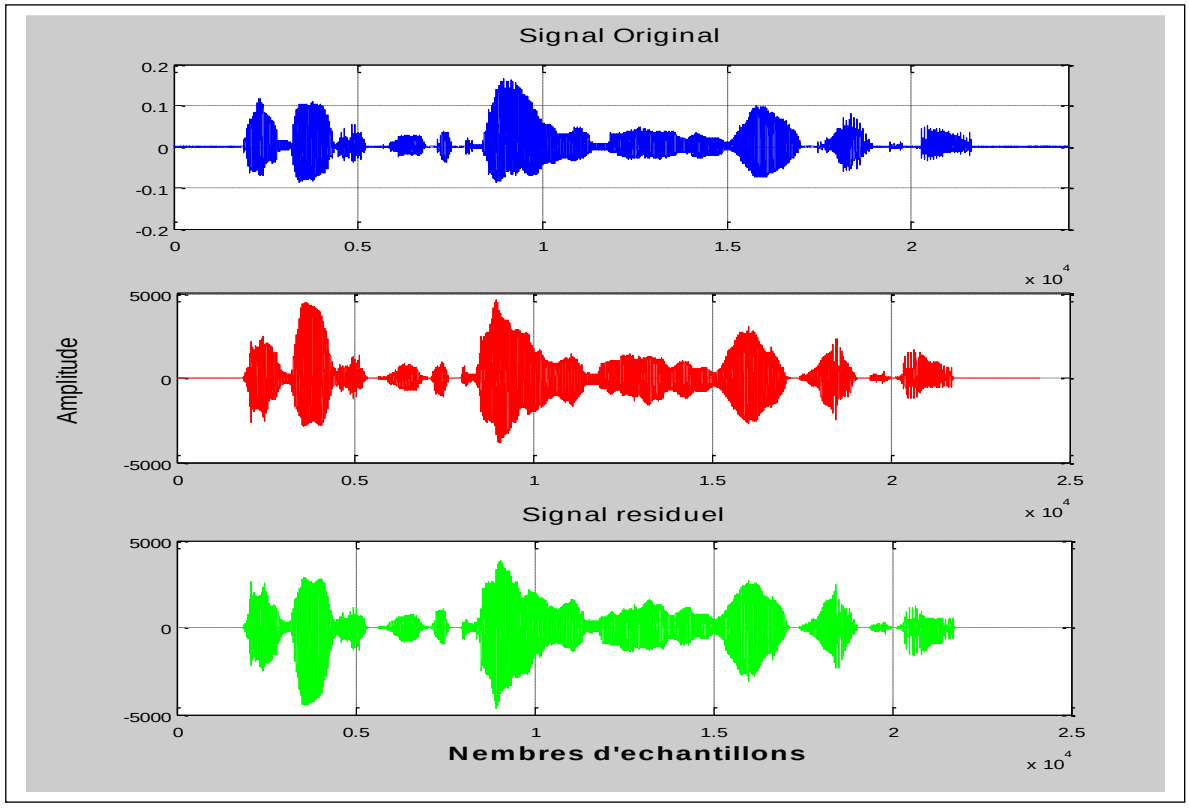


Figure. IV.10: Phrase prononcée par un locuteur" don't ask me to carry an oily rag like that."

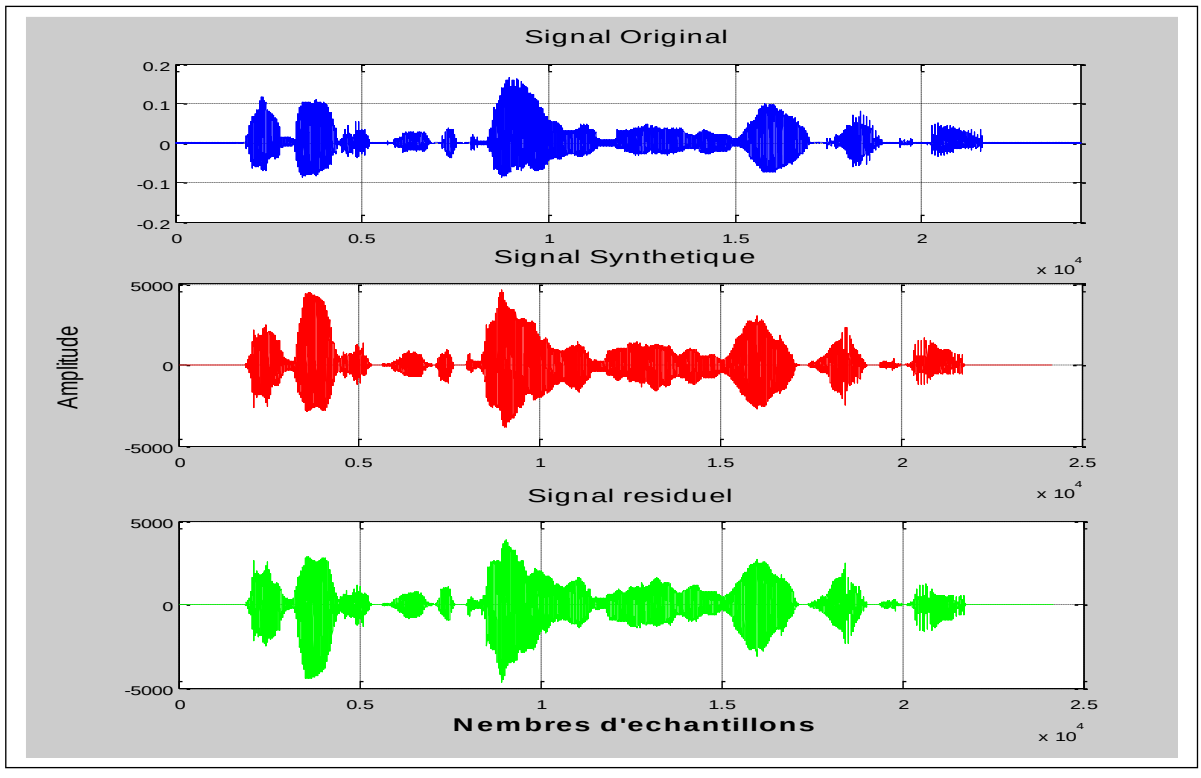


Figure. IV.11: Phrase prononcée par une locutrice "don't ask me to carry an oily rag like that"

IV.5. Evaluation objective des Résultats :

Plusieurs simulations, ont été réalisées pour évaluer les performances de ces deux codeurs fonctionnant à2.4 kbps. Le but est de chiffrer la qualité perceptuelle de nos codeurs et d'évaluer leurs performances.

La figure ci-dessous présente l'exécution sous le logiciel PESQ. On a inséré un signal d'entrée original de type (.wav) et le signal synthétisé avec l'un des deux codecs à 2.4 kbps pour les comparer et avoir la note de qualité qui détermine la différence entre ces deux signaux. Sachant que la fréquence d'échantillonnage introduite est de 8KHz et la bande étroite.

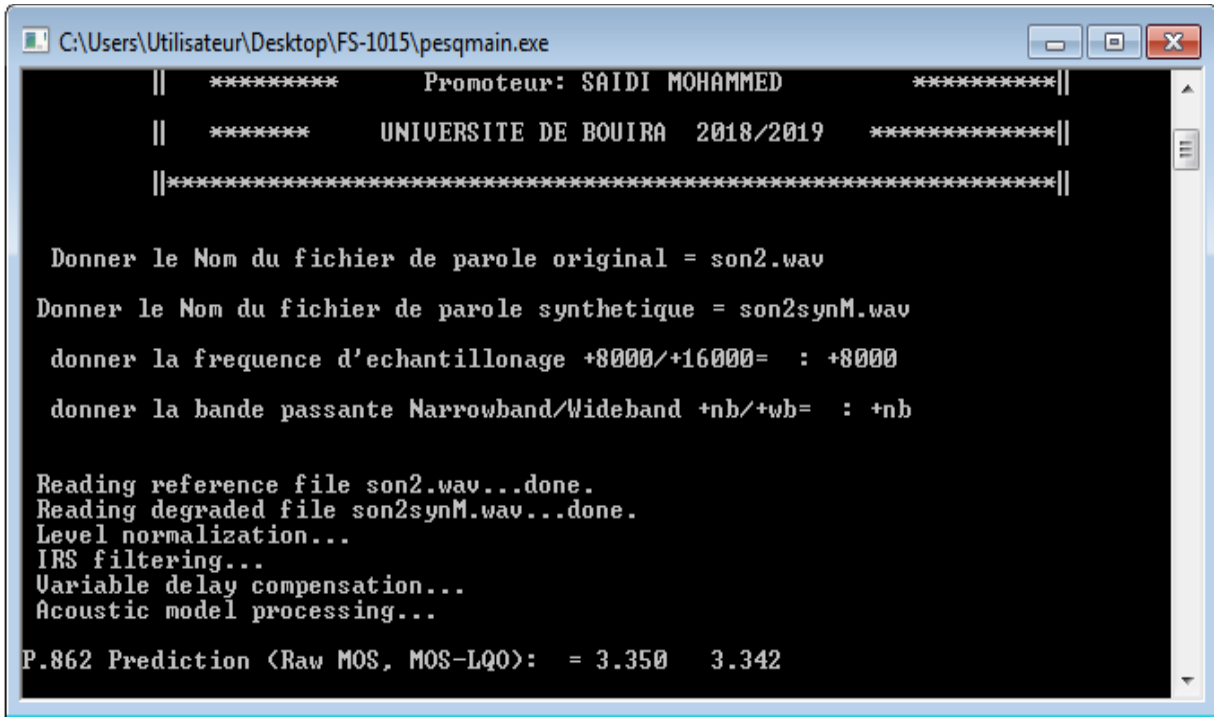


Figure.IV.12 : l'exécution sous le logiciel PESQ.

Les deux tableaux ci-dessous résument la moyenne du score PESQ pour les bases de données utilisées. Les mesures de ces notes ont été faites suivant cette exécution.

Tableau. IV.1: Scores PESQ pour la langue arabe.

PESQ LPC-10 à 2.4 Kbps		**PESQ MELP à 2.4 Kbps**	
Masculin	2,67235	Masculin	3,16665
Féminin	2,6094	Féminin	2,99585

Tableau. IV.2: Scores PESQ pour la langue anglais.

PESQ LPC-10 à 2.4 Kbps		**PESQ MELP à 2.4 Kbps**	
Masculin	2,737	Masculin	2,929
Féminin	2,716	Féminin	2,878

D’après les résultats obtenus, nous constatons que :

- Le score PESQ obtenu est meilleur pour les locuteurs masculins que les locutrices féminines pour les deux codeurs MELP et FS-1015.
- Le score PESQ du codeur MELP est meilleur que celui du FS-1015.
- Pour les deux codeurs, l’intelligibilité synthétique est assurée à un niveau assez bon.

IV.5.1. Comparaison entre les deux codeurs :

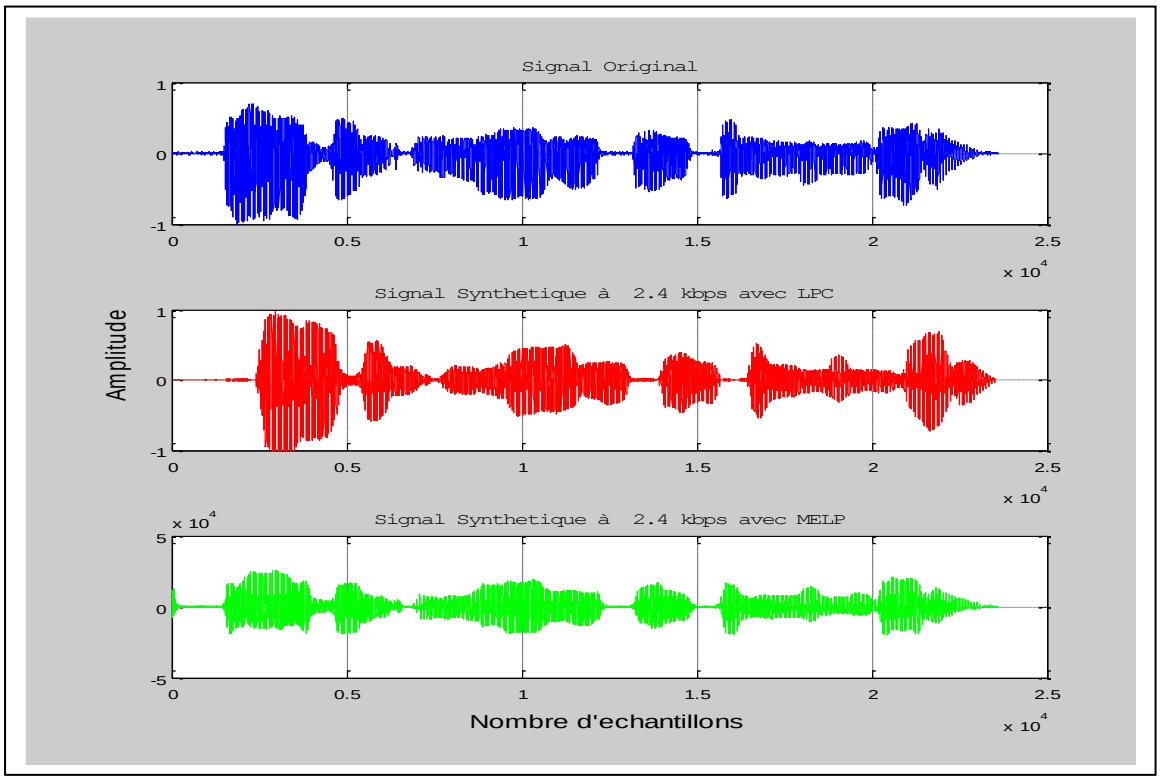


Figure. IV.13: Phrase prononcée par un locuteur «صعد الإمام فوق المنبر »

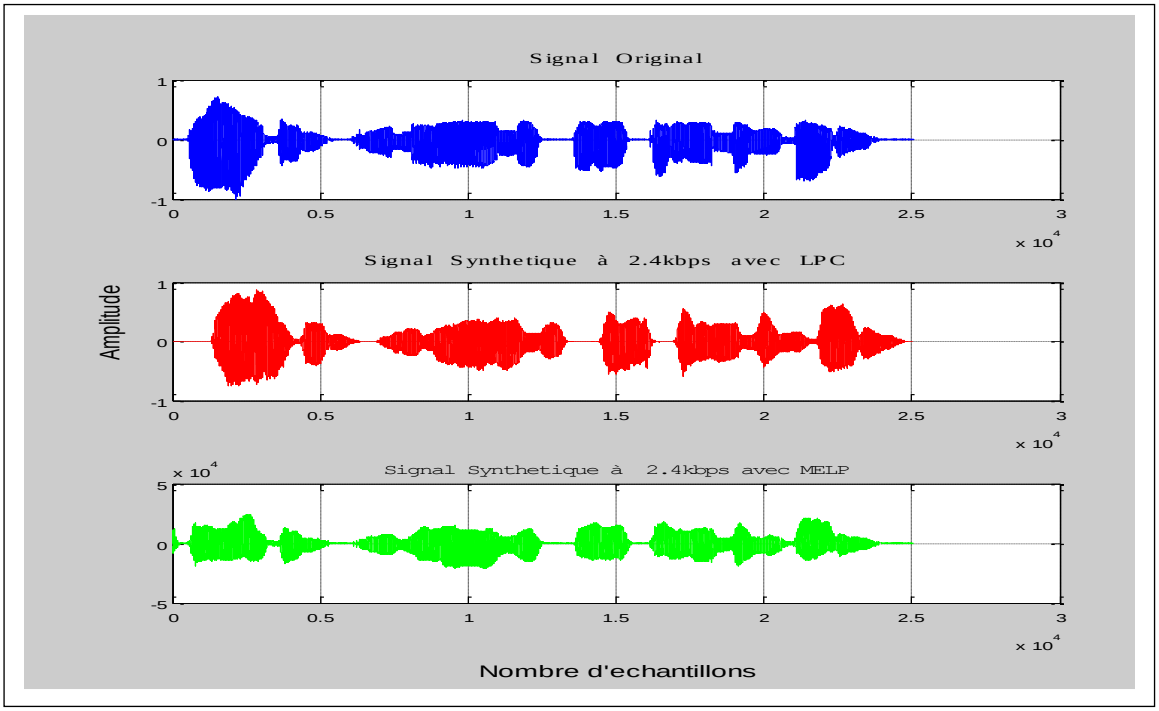


Figure. IV.14: Phrase prononcée par une locutrice« صعد الإمام فوق المنبر»

- **Comparaison des résultats obtenus :**

D'après les graphes obtenus après les opérations de codage et décodage avec les standards FS-1015 et MELP a 2.4 Kbps on peut dire que :

- Le signal original et signal synthétique n'ont pas la même forme, donc le codage a très bas débit ne garde pas la même forme.
- Les signaux résiduel et synthétique ont une grande ressemblance en terme de forme mais pas en terme de qualité.
- L'analyse des tableaux des notes obtenues avec l'évaluation PESQ et la comparaison entre les signaux synthétiques à 2.4 Kbps pour les deux codeurs permet de dire que le MELP donne une meilleure qualité que LPC-10 avec tout la complexité de calcul.

IV.6. Conclusion

Dans ce chapitre nous avons présenté une évaluation de deux codeurs LPC-10 et MELP. Les codeurs sont des systèmes de codage mixtes qui utilisent des représentations paramétriques du signal de parole. Notre objectif était focalisé envers l'étude des codeurs à 2.4 Kbps, et afin de tester et comparer leurs performances pour un ensemble de langues, nous avons effectué une implémentation software que nous avons testé avec une base de données phonétiquement équilibré.

Etant donné que les deux codeurs appartiennent à la famille de codeurs paramétriques qui ne préservent pas la forme d'onde, on se référence sur un fort critère d'évaluation qui est le PESQ. Les différents résultats d'évaluation objective avec le PESQ ont montré que les deux codeurs appartiennent à la bonne catégorie de qualité perceptuelle mais le MELP à 2.4 Kbps offre de meilleurs résultats que le FS-1015.

Conclusion Générale

Le traitement de la parole s'est développé pendant les dernières décennies dans le domaine des télécommunications. Des avances importantes ont été réussies sur le plan du traitement analogique et fondamentalement sur le traitement numérique grâce à des techniques comme les algorithmes de compression. L'objectif de la compression/codage des signaux numériques est la réduction des couts de transmission et de stockage. Cependant il ne faut perdre de vue qu'un système de compression est toujours le résultat d'un compromis entre quatre critères principaux et qui sont: La qualité du signal restitué, débit, complexité et le retard. Le codage de la parole à très bas débit sera la question de recherche la plus importante.

Dans ce travail nous avons mis en œuvre deux standards de codage de la parole à très bas débit, le premier codeur est le LPC-10 à 2.4 Kbps et le second c'est le MELP à 2.4 Kbps dans le but de comparer la synthèse de la parole obtenue par ces deux types codeurs. Nous avons simulé le codeur MELP avec le langage C (Builder C++ 6.0) pour la partie programmation et Matlab pour les représentations. Pour tester la qualité du discours obtenu par les deux codeurs, nous avons utilisé la méthode dite objective : PESQ (évaluation perceptuelle de la qualité de la parole) afin de juger les performances de chacun d'eux.

Les résultats obtenus montrent que les deux codeurs appartiennent à la bonne catégorie, avec des scores PESQ désignant une qualité perçue assez bonne, mais la comparaison entre les signaux synthétiques nous permet de dire que la qualité du signal de sortie de codeur MELP est meilleure que celle de LPC-10 malgré toute sa complexité de calcul.

Avant de terminer cette conclusion, on peut dire que ce travail nous a apporté des intérêts considérables tant sur le plan théorique qu'expérimental. En effet, de plus qu'il nous a permis de creuser dans la simulation et la programmation, nous avons été amenés, par ce travail, à nous instruire dans le domaine de codage et le traitement de la parole.

Nous avons, aussi pu nous apprivoiser avec un domaine d'actualité c'est le codage de parole à très bas débit, nous avons vu les différentes classes des codeurs de parole et en détails le codeur LPC-10 et MELP, les mesures d'évaluation des performances dans le codage de la parole. Notamment, on a mis en œuvre les codecs LPC-10 et MELP opérant à des débits de 2.4 Kbps. Enfin, en implémentant ces codeurs on a pu les évaluer avec la mesure PESQ en utilisant un corpus de parole de langues différentes.

Perspectives

Comme perspectives, l'ensemble des codeurs présentés a été programmé en langage haut niveau C/C++, nous souhaitons que cette contribution serve de base pour de développement dans le domaine de traitement numérique de la parole. Particulièrement, pour ceux qui auront objectif d'intégrer des fonctionnalités de codage en temps réel sur les applications DSP en vue d'une évaluation de toutes leurs performances.

Références bibliographiques

[1] S.K.J agtap, M.S.Mulye, M.D.Uplane, « speech coding techniques»,2015.

[2] Jacob Benesty, M. Mohan Sondhi, Yiteng Huang, « springer handbook of speech processing », pp.331-350, 2008.

[3**]** A.Hacine gharbi, « sélection de paramètres acoustiques pertinents pour la reconnaissance de la parole », thèse de doctorat en sciences, spécialité: électronique, université Ferhat Abbas-Sétif, Algérie, 2012.

[4] L. Buniet, « traitement automatique de la parole au milieu bruité : étude de modèles connexionniste statiques et dynamiques », thèse de doctorat, spécialité informatique, université Henri Poincaré-Nancy I, 1997.

[5] Y. Aziza, « modélisation AR et ARMA de la parole pour une vérification robuste du locuteur dans un milieu bruite en mode dépendant de texte », thèse magister, Option: communication, université Ferhat Abbas-Setif1-UFAS(Algérie), 2013.

[6] C. Plapous, « Traitements pour la réduction de bruit. Application à la communication parlée », thèse de doctorat, Université de RENNES 1,2005.

[7] S. Bouasli, A. Noumeri, « compression et codage de la parole par la transformée KLT automatisation », mémoire de master, option : systèmes de télécommunications, université Djilali Bounaama Khemis Miliana, 2016.

[8] M. Mehassouel, « application de la technologie MIMO à la 4G du mobile » thèse de magister, option: communication, université Ferhat Abbas-Setif1, 2014.

[9] J. Raverdino, J M. Guatteri, Techniques multimédia pour le son.

[10] J. Hernandez, « algorithmes d'acquisition, compression et restitution de la parole à vitesse variable. Etude et mise en place », école nationale supérieure de l'électronique et de ses applications cergy-pontoise, Paris**,** 1995.

[11] R. Benammar, « traitement automatique de la parole arabe par les HMMs : calculatrice vocale », université Abou Bekr Belkaid, Tlemcen, 2012.

[12] M. Hasegawa-Johnson, "speech coding: fundamentals and application", university of Illinois at Urbana–Champaign Urbana, Illinois ABEER ALWAN, university of California at Los Angeles Los Angeles, California, 2003.

[13] G. Baudoin, J. Cernocký, P. Gournay, G. Chollet, « Codage de la parole à bas et très bas débit », n° 9-10, pp.2-11, Tchèque, 2000.

[14] N. Lachachi, « Codage paramétrique de la parole en vue de transmission sur Internet », Mémoire de Magistère, Spécialité : Informatique, Université D'Oran-es-senia- faculté des Sciences, Octobre 2006.

[15] M. Djamah, « Codage échelonnable à granularité fine de la parole utilisant la quantification vectorielle arborescente », thèse doctorat en télécommunications, université du Québec INRS (centre énergie matériaux télécommunications).

[16] P.Gournay, G.Quagliaro, A.Goye, G.Guilmin, F.Chartier,« procédé de correction auditive utilisant un modèle paramétrique du signal de parole »,France, 1999.

[17] W.C. Chu, «speech coding algorithms: foundation and evolution of standardized coders», John Wiley & Sons, 2003.

[18] L.Hanzo, F.Clare Somerville, J.Woodard, "voice and audio compression for wireless communication", 2 éme édition,pp.543-544, 2007.

[19]D.Gibson, T. Berger, T. Lookabaugh, Dave. Lindbergh, Richard L Baker, "digital compression for multimedia: principles and standards", pp.190-193, 1998.

[20] T. Ogunfunni, M. Narasimha, "Principale of speech coding", pp.173-175, 2010.

[21] R. Goldberg, L. Riek, "a practical handbook of speech coders" CRC press, 2000.

[22] C .Kritzinger, "low bite rate speech coding", mémoire de master, université Stellenbosch, 2006.

[23] A. Ray, T. Acharya, "information technology: principles and application", PP.265, New Delhi, India, 2004.

[24] J. Macker, R. Adamson**,** "a variable rate voice coder using LPC-10E" 1994.

[25] N. Moreau, « outils pour la compression application à la compression des signaux audio », Novembre 2009.

[26] H.Kheddar, B.Boudraa, " implementation of intervaling methods on MELP 2.4 kbps coder to reduce packet loss in the voice over IP (VOIP) transmission'' journal of engineering research and application, VOL.5(issu N 3),march 2015.

[27] M. Saidi1, L. Falek, B. Boudraa, ''codage par prédiction linéaire (LPC) à bas et très bas débit '', Alger.

[28] A.Cheradid, « Étude des approches adaptatives liées à la QoE dans le cadre des applications de téléphonie sur IP », mémoire de magistère en informatique, option : technologies de l'information et de la communication, université Kasdi Merbah, Ouargla, 2014.

[29] R. Beuran, ‘‘ mesure de la qualité dans les réseaux informatiques’’. Thèse de doctorat, université de Jean Monnet St Etienne, 2004.

[30] M.Guéguin, «évaluation objective de la qualité vocale en contexte de conversation», thèse de doctorat en traitement de signal et télécommunications, université de Rennes 1 ,France,2007.