RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE

MINISTÈRE DE L´ENSEIGNEMENT SUPÉRIEURE
ET DE LA RECHERCHE SCIENTIFIQUE

ÉCOLE NATIONALE POLYTECHNIQUE

DÉPARTEMENT D´ÉLECTRONIQUE

PROJET DE FIN D'ÉTUDES EN VUE DE L'OBTENTION DU DIPLÔME D'INGÉNIEUR D'ÉTAT EN ÉLECTRONIQUE

THÈME :

AMÉLIORATION DU CODEC G.729 PAR ENTRELACEMENT DES TRAMES

Proposé et dirigé par :
Melle: F.MERAZKA

Réalisé par:
BOUAFIA Abdeldjalil
BELGROUNE Nadia

PROMOTION 2006

Laboratoire Signal et Communications
E.N.P. 10, Avenue Hassen-Badi, El Harrach, ALGER

REMERCIEMENTS

Ce travail a été effectué au sein du laboratoire de signal et communications du département d'électronique de l'Ecole Nationale Polytechnique, sous la direction du Dr F.MERAZKA

Nous tenons à lui exprimer nos plus sincères remerciements pour ses conseils, son aide et sa patience tout au long de ce travail.

Nous exprimons notre plus sincère gratitude au Professeur D.BERKANI, pour son aide et sa disponibilité et qui a rendu possible l'entreprise de ce travail.

Nous tenons à exprimer notre très grande gratitude, et notre profonde affection à nos chers parents pour leur encouragement, leur patience et leur grand soutien, durant toutes ces années d'études.

Nous tenons à remercier NADIR, pour son aide sans laquelle ce travail n'aurait pu aboutir.

Nous tenons également à remercier tous nos amis et camarades, pour leur sincère amitié et leur précieux soutien.

ملخّص

لاحظنا في الرّامزة النموذجية G.729, المقدّمة من طرف الاتحاد الدّولي للاتصالات اللاسلكية, أنّه بعد ضياع قطع من الكلام عبر شبكة الإنترنت, تصحّح الرّامزة النموذجية G.729 هذه القطع الضائعة, لكن بكل غير فعّال.
لمعالجة هذا الإشكال, أدخلنا طريقة تقوم على تشابك القطع, تمكّن من تخفيض أثر القطع الضائعة, و ذلك بتقسيمها إلى وحدات صغيرة الشّيء الذي يزيد من فعاليّة الرّامزة النموذجيّة G.729 ضدّ ضياع قطع.

الكلمات المفتاحيّة :
ترميز الكلام, الصوت عبر شبكة الإنترنت, ضياع قطع, إخفاء القطع, تشابك القطع.

ABSTRACT

In the codec G.729 of ITU (*International Telecommunication Union*), we have observed that after a frame erasure the standard concealment of the CoDec does not manage to cover this loss perfectly, because the error remains considerable.

To cure this problem, we implemented a method based on interleaving. It makes it possible to decrease the effect of the packet loss, by division of the size of a frame into units. Which improves the performances of G.729 against the losses.

Key words:
Speech coding, Voice over IP, Packet loss, Loss concealment, Frame interleaving.

RESUME

Dans le CoDec G.729 de l'ITU (*International Telecommmunication Union*), nous avons observé qu'après une perte de trame, à travers une liaison IP, la dissimulation standard du CoDec ne parvient pas à couvrir parfaitement cette perte, car l'erreur reste considérable.

Pour remédier à ce problème, nous avons implémenté une méthode basée sur l'entrelacement. Elle permet de diminuer l'effet de la perte de trames, par division de la taille d'une trame en unités. Ce qui améliore les performances du G.729 vis-à-vis des pertes.

Mots clefs :
Codage de la parole, Voix sur IP, Perte des trames, Masquage des pertes, Entrelacement.

Table des matières

Liste des figures ……..……...**i**
Liste des tableaux ……..…**iii**
Introduction ……**iv**
Chapitre I ……..**1**
I.1 Caractéristiques d'un signal vocal ……...……...…2
I.1.1 Production de la voix ……...…2
I.1.2 Sons voisés ou non voisés ……3
I.1.3 Perception de la parole et redondance ……...……...…...5
I.1.4 Modélisation de la parole ……...……6
I.1.5 Codage prédictif et analyse par synthèse ……...7
I.1.5.1 Prédiction linéaire ……..……..8
I.1.5.2 Estimation des coefficients de prédiction linéaire ……..……12
I.1.5.3 Considérations pratiques ……...……...17
I.1.5.4 Pondération perceptive ……18
I.2 Principe de quantification ……...……19
I.2.1 Quantification scalaire ……..……..19
I.2.2 Quantification vectorielle ……..20
I.3 Techniques de codages de la parole ……..…..20
I.3.1 Le codage de forme d'onde ……..21
I.3.2 Le codage paramétrique ……..……22
I.3.3 Le codage hybride ……...……22
I.4 Qualité des codeurs ……..…...23
I.4.1 Mesure de distorsion subjective ……...……..……23
I.4.2 Mesure de distorsion objective ……24
I.4.2.1 Domaine temporel ……...……25
I.4.2.2 Domaine fréquentiel ……...……25
I.4.3 Mesure de distance euclidienne LSP pondérée ……26
Chapitre II ……**28**
II.1 Généralités ……...……..29
II.2 Principe de fonctionnement ……..……30
II.2.1 Architecture des réseaux ……..……31
II.2.1.1 Modèle OSI ……31
II.2.1.2 Modèle TCP/IP ……..……33
II.2.1.3 Modèle UIT-T ……...……36
II.2.2 Acheminement de l'information dans les réseaux ……...……..36
II.3 Qualité de service ……..……36

II.3.1 Les niveaux de qualité ……..……36
II.3.2 Les facteurs affectant la Qualité de Service ……..…..…37
II.3.2.1 Le Retard ……...….37
II.3.2.2 La Gigue ……..……..….38
II.3.2.3 La bande passante ……...……...…39
II.3.2.4 Les Pertes de Paquets ……..…….....39
II.3.3 Les Techniques de Masquage des Paquets perdus ……...….40
II.3.3.1 Masquage Basé sur l'Emetteur ……..……40
II.3.3.2 Masquage Basé sur le Récepteur ……..…….42
II.4 Perspectives et enjeux économiques ……44
Chapitre III ……..**46**
III.1 Description général du standard G.729 ……47
III.1.1 Codeur ……..……47
III.1.2 Décodeur ……..…51
III.2 Procédure de masquage des trames effacées du G729 ……..52
III.2.1 Répétition des paramètres du filtre de synthèse ……53
III.2.2 Affaiblissement des gains du dictionnaire adaptatif et fixe ……...53
III.2.3 Affaiblissement de l'énergie mémorisée par le prédicteur de gain …..54
III.2.4 Production de l'excitation de remplacement ……54
Chapitre IV ……..**55**
IV.1 Présentation de la méthode d'entrelacement ……56
IV.1.1 Entrelacement de quatre trames ……...57
IV.1.2 Entrelacement de huit trames ……57
IV.1.3 Entrelacement de seize trames ……..57
IV.2 Implémentation, simulation et résultats ……...58
IV.2.1 Implémentation ……58
IV.2.2 Simulation ……..……58
IV.2.2.1 Base de données vocales ……...……58
IV.2.2.2 Modèle du réseau IP ……...59
IV.2.3 Résultats ……..……60
IV.2.3.1 Distorsion spectrale ……..……..60
IV.2.3.2 PESQ ……..…60
IV.2.3.3 EMBSD ……..……60
IV.3 Présentation des résultats ……...……..61
IV.3.1 Distorsion spectrale ……...61
IV.3.2 PESQ ……..……65
IV.3.3 EMBSD ……..……69
IV.4 Analyse et interprétations ……..…..73
Conclusion ……...……**74**
Bibliographie ……..……..**76**

Liste des Figures

Figure I-1 : Système phonatoire …….....3
Figure I-2 : Spectre d'un signal vocal voisé ……....4
Figure I-3 : Un signal vocal non voisé et son spectre …….....5
Figure I-4 : Modélisation de la production de la parole …….....6
Figure I-5 : Synthèse de parole à deux états d'excitation ....…….....7
Figure I-6 : (a) signal de parole voisée et non voisée …….....7
(b) signal d'excitation correspondant …..…….....7
Figure I-7 : Diagramme du codeur LPAS ……..8
Figure I-8 Modèle simplifié de production de la parole ……....10
Figure I-9 : Analyse et synthèse LP ……...11
Figure I-10 : Modélisation LPC d'une voix d'homme ……...11
Figure I-11 Localisation possible des racines pour $P(z)$ et $Q(z)$ d'ordre pair ……...16
Figure I-12 : Réponse fréquentielle du filtre de pondération perceptive ……19
Figure I-13 Quantification scalaire ……20
Figure I-14 : Comparaison de la qualité de codage de parole ……21
Figure II-1 : Principe de la transmission de la voix par paquets ……30
Figure II-2 : Modèle de référence OSI ……...31
Figure II-3 : Modèle TCP/IP ……..33
Figure II-4 : Mise en paquet de l'information ……35
Figure II-5 : Qualité de service en fonction des pertes de paquets ……37
Figure II-6 : Schéma illustrant la GIGUE ……..39
Figure II-7 : Classification des techniques de réparations basées à l'émetteur ……...40
Figure II-8 : Exemple de FEC ……40
Figure II-9 : Exemple *d'Interleaving* ……41
Figure II-10 : Classification des techniques de masquage d'erreur ……42
Figure II-11 : Evolution du trafic téléphonique international sur IP ……..45
Figure III-1 : Réponse du filtre de prétraitement (a) impulsionnelle (b) fréquentiel …….48
Figure III-2 : Procédure de fenêtrage en analyse (LP)……...48
Figure III-3 : Principales procédures du codeur G.729 ……..49
Figure III-4 : Principe du codeur CS-ACELP G.729 ……50
Figure III-5 : Principe du décodeur CS-ACELP G.729 ……52

Figure III-6 : Masquage des paquets IP perdus …….…….53
Figure IV-1 : Principe de la méthode d'Entrelacement ……..……56
Figure IV-2 : Distribution de l'erreur par la méthode d'entrelacement ……..57
Figure IV-3 : Pertes de paquets modélisées par un processus aléatoire de Markov ……...59
Figure IV-4 : Distorsion spectrale de l'Entrelacement de 4 trames ……61
Figure IV-5 : Distorsion spectrale de l'Entrelacement de 8 trames ……62
Figure IV-6 : Distorsion spectrale de la méthode d'Entrelacement de 16 trames …….…….63
Figure IV-7 : Distorsion spectrale de la méthode d'Entrelacement de 4, 8 et 16 trames ..….64
Figure IV-8 : PESQ de l'entrelacement de 4 trames pour une voix masculine ……..65
Figure IV-9 : PESQ de l'entrelacement de 8 trames pour une voix masculine ……..65
Figure IV-10 : PESQ de l'entrelacement de 16 trames pour une voix masculine ……..66
Figure IV-11 : PESQ de l'entrelacement de 4, 8 et 16 trames pour une voix masculine …...66
Figure IV-12 : PESQ de l'entrelacement de 4 trames pour une voix féminine ……..67
Figure IV-13 : PESQ de l'entrelacement de 8 trames pour une voix féminine ……..67
Figure IV-14 : PESQ de l'entrelacement de 16 trames pour une voix féminine ……68
Figure IV-15 : PESQ de l'entrelacement de 4, 8 et 16 trames pour une voix féminine …….68
Figure IV-16 : EMBSD de l'entrelacement de 4 trames pour une voix masculine ……69
Figure IV-17 : EMBSD de l'entrelacement de 8 trames pour une voix masculine ……69
Figure IV-18 : EMBSD de l'entrelacement de 16 trames pour une voix masculine ……..70
Figure IV-19 : EMBSD de l'entrelacement de 4, 8 et 16 trames pour une voix masculine ...70
Figure IV-20 : EMBSD de l'entrelacement de 4 trames pour une voix féminine ……..71
Figure IV-21 : EMBSD de l'entrelacement de 8 trames pour une voix féminine ……..71
Figure IV-22 : EMBSD de l'entrelacement de 16 trames pour une voix féminine ……72
Figure IV-23 : EMBSD de l'entrelacement de 4, 8 et 16 trames pour une voix féminine ….72
Figure IV-24 : Segment de 10 Micro-secondes de voix féminine et masculine ……..73

Liste des Tableaux

Tableau I-1 : Qualité avec la mesure MOS ……24

Tableau III-1 : Affectation des bits dans l'algorithme de codage CS-ACELP à 8 kbit/s …...47

Tableau IV-1 : Les taux de pertes simulés ……59

Tableau IV-2 : Distorsion spectrale de l'Entrelacement de 4 trames pour une voix masculine……..61

Tableau IV-3 : Distorsion spectrale de l'Entrelacement de 4 trames pour une voix féminine……61

Tableau IV-4 : Distorsion spectrale de l'Entrelacement de 8 trames pour une voix masculine.....62

Tableau IV-5 : Distorsion spectrale de l'Entrelacement de 8 trames pour une voix féminine……62

Tableau IV-6 : Distorsion spectrale de l'Entrelacement de 16 trames pour une voix masculine……..63

Tableau IV-7 : Distorsion spectrale de l'Entrelacement de 16 trames pour une voix féminine……63

Tableau IV-8 : Valeurs du PESQ de la méthode d'Entrelacement pour une voix masculine……..65

Tableau IV-9 : Valeurs du PESQ de la méthode d'Entrelacement pour une voix féminine……67

Tableau IV-10 : Valeurs de l'EMBSD de la méthode d'Entrelacement pour une voix masculine……..69

Tableau IV-11 : Valeurs de l'EMBSD de la méthode d'Entrelacement pour une voix féminine……71

Introduction

L'objet de notre étude se situe dans le domaine -particulièrement important pour les systèmes de télécommunications modernes- de la compression de signaux de parole.

Pour transmettre des signaux de parole, des chaînes de communication analogique ont été utilisées, presque exclusivement, jusqu'à une époque récente : en raison des perturbations et des bruits apparaissant inévitablement dans le canal de transmission, le signal reconstruit au récepteur ne pouvait être une réplique exacte du signal émis.

Des systèmes numériques, dont le principal avantage est de transmettre des informations codées par numérisation avec une meilleure fiabilité, ont été élaborés.

Depuis les années 80 l'idée de transmettre la voix sur les réseau Internet et Intranet s'impose de plus en plus.

La voix sur IP (*Voice over IP*) est une technologie de communication vocale en pleine émergence. En effet, la convergence du triple *play* (voix, données et vidéo) fait partie des enjeux principaux retenus par les acteurs de la télécommunication aujourd'hui.

Comme toute innovation technologique, la VoIP doit non seulement simplifier le travail mais aussi faire économiser de l'argent : En particulier, plus les interlocuteurs sont éloignés et plus la différence de prix est intéressante. De plus, la téléphonie sur *IP* utilise jusqu'à 10 fois moins de bande passante que la téléphonie traditionnelle.

Les caractéristiques de la téléphonie IP, à l'exemple de la possibilité d'une plus grande compression de l'information par paquets, en font une application ambitieuse. Cependant, le réseau IP étant basé sur le principe dit "*best effort*", il ne garantie pas la réception des paquets envoyés.

La téléphonie IP ne pourra s'imposer que dans la mesure où la qualité de la parole reçue sera suffisante. C'est pourquoi, des algorithmes performants ont été développés pour réduire l'effet des pertes de paquet sur le signal vocal.

Nos travaux se sont concentrés sur le codeur G.729 , retenu parmi les recommandations de l'IUT (*International Union of Télécommunications*) , qui standardise un codage de la parole à 8 kbits/s basé sur des techniques de Prédiction Linéaire de type CELP. Le choix de ce dernier a été validé car il présentait déjà des caractéristiques particulièrement adaptées aux applications DSVD (*Digital Simultaneous Voice and Data*). De plus, la faible complexité d'implantation du G.729 en fait un choix intéressant pour la téléphonie sur Internet.

En effet, le G729 a de faibles exigences en ce qui concerne le débit. Cependant, une éventuelle perte de trames nuis constatablement à son efficacité. Un algorithme de masquage des pertes, basé sur la méthode d'interpolation, lui est incorporé, mais il n'introduit pas une grande amélioration.

Notre étude a donc été menée pour améliorer ses performances en implémentant une méthode de masquage des trames effacées plus performante que celle adoptée par ce dernier pour alléger les effets des pertes de paquets.

L'objectif principal de notre travail consiste à implémenter une nouvelle méthode de dissimulation des trames perdues basée sur l'entrelacement afin d'améliorer les performances du CoDec G.729.

Ce rapport comprend quatre chapitres organisés comme suit :

Le premier chapitre donne des généralités sur le codage de la parole, les caractéristiques et la modélisation du système phonatoire humain et les principes des différentes techniques connues aujourd'hui pour compresser un signal de parole.

Le deuxième chapitre est consacré à la transmission de la voix sur IP, la description du réseau IP et de ses caractéristiques liées aux différentes méthodes de recouvrement des pertes.

Le troisième chapitre décrit le fonctionnement du CoDec G.729, les principes de codage et ceux du décodage.

Le quatrième chapitre décrit la méthode implémentée, les simulations et affichage des résultas obtenus.

Enfin une conclusion générale résume et clos le travail.

I – Codage de la parole

La parole est incontestablement le moyen de communication le plus important pour l'humanité. Cela fait d'elle un élément central de la communication numérique et donc un vecteur essentiel dans l'évolution des technologies des télécommunications.

Ce chapitre regroupe des généralités sur les notions fondamentales de la production du signal parole, ses propriétés ainsi que sa perception dans le but de présenter les principes de base du codage de parole, dont l'intérêt récent a été motivé par l'introduction des services de télécommunications entièrement numériques. Le codage de parole permet la réduction du débit de transmission du signal et des communications dans des canaux à largeur de bande limitée. La largeur de bande d'une transmission devra être minimiser pour réduire la mémoire nécessaire dans le système de stockage de la parole tout en préservant la qualité du signal vocal reconstruit [1] et en répondant aux autres exigences liées à l'application.

I.1 Caractéristiques d'un signal vocal

En raison des caractéristiques du conduit vocal humain, le signal de la parole est fortement redondant. Ces redondances permettent aux algorithmes de codage de compresser le signal en enlevant l'information non pertinente contenue dans le signal. La connaissance du système vocal et des propriétés du signal de parole est essentielle pour concevoir des codeurs efficaces. Les propriétés du système auditif humain peuvent également être exploitées pour améliorer la qualité perceptuelle du signal codé.

Avant d'aborder le problème plus précis du codage de la parole, nous présenterons quelques caractéristiques du signal qui permettront de mieux appréhender les différentes techniques de codage présentées par la suite. Dans cette section, une partie simple de la théorie acoustique est représentée et les notions de phonème, de formant, de son voisé et non voisé et de pitch [2] [3] sont définies.

I.1.1 Production de la voix

Le signal vocal est engendré par l'appareil phonatoire avant d'être émis puis détecté par un auditeur. A la perception l'oreille analyse ce signal et transmet au cerveau les informations nécessaires à son interprétation. La figure I-1 présente le système vocal qui se compose d'une soufflerie (poumon et conduit trachéo-bronchique), du larynx et du conduit vocal formé par le pharynx ainsi que les cavités buccales et nasales. La parole est généralement produite par expiration de l'air à travers la glotte et le conduit vocal.

L'air venant des poumons est modulé par vibrations des cordes vocales par déformation (élargissement ou resserrement) du conduit. Ce système phonatoire permet le processus de production de la parole qui comporte trois étapes principales :

- La génération d'une énergie ventilatoire qui mettra en mouvement oscillatoire les cordes vocales.
- Les vibrations des cordes vocales qui donneront naissance à des sons voisés.
- L'articulation qui sera réalisée dans les cavités supra-glottiques (conduit vocal et fosses nasales).

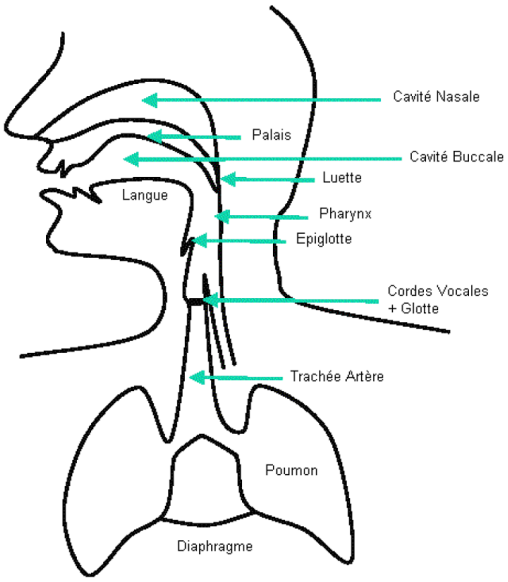


Figure I-1 : Système phonatoire

Le signal de parole étant un signal réel, continu, à énergie finie et non stationnaire, les caractéristiques du signal de parole et du conduit vocal évoluent dans le temps. Les positions du système phonatoire agissent comme une opération de filtrage, en augmentant certaines fréquences tout en atténuant d'autres. Ces composantes (cordes vocales, langue, lèvres, dents, mâchoire, etc.) sont déplacées pour former différents bruits [1]. Notons que la position du larynx varie avec le sexe et l'âge du locuteur. Elle s'abaisse progressivement jusqu'à l'âge adulte mais reste plus élevée chez la femme. Cette différence aura des conséquences sur la position des formants et la hauteur de voix.

I.1.2 Sons voisés ou non voisés

La parole est composée de vagues de pression acoustique créées par le flux d'air, qui provient des poumons, à travers le conduit vocal. Les cordes vocales dans le larynx se ferment quasi-périodiquement pour interrompre ce flux. Malgré sa structure complexe, un signal de parole non stationnaire, tantôt périodique, tantôt aléatoire, peut être répertorié, par simplification, par sa périodicité et son énergie. Une décomposition simplifiée d'un signal de parole fait ressortir deux types de sons, voisés (les voyelles) ou non voisés (les consonnes) :

• Les sons voisés, tels que les voyelles, sont produits par le passage de l'air des poumons à travers la trachée qui met en vibration les cordes vocales. Ce mode, qui représente 80% du temps de phonation, est caractérisé en général par une quasi-périodicité et une énergie élevée.

• Les sons non voisés, comme des consonnes, sont obtenus par resserrement du conduit vocal, et sont habituellement d'énergie inférieur aux sons voisés. Les cordes vocales sont écartées et n'entrent pas en vibration. Les consonnes sont un exemple de son non voisé, apériodique. Ces sons sont considérés comme ayant les mêmes caractéristiques que le bruit.

La figure I-2 représente une trame de parole voisée on peu donc remarquer une certaine périodicité dans le signal due à la vibration des cordes vocales sous l'effet du passage de l'air à travers la glotte. Leur fréquence de vibration est appelée fréquence fondamentale F_0 ou ***pitch***. Cette dernière varie en fonction de chaque individu [4], en général la variation s'étend :

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

Le flux laryngé étant modulé par un résonateur pharingo-buccal, le conduit modifie la distribution de l'énergie du spectre de la source vocale et présente plusieurs fréquences de résonance qui forment une enveloppe. Comme on le remarque dans figure I-2, la représentation d'un segment de parole périodique dans le domaine fréquentiel forme des harmoniques dont l'amplitude est modulée par l'effet de filtrage du conduit vocal. Il en résulte plusieurs zones de fréquences renforcées appelées ***formants*** [5] et qui correspondent aux fréquences propres F_i du conduit vocal (structure formantique). Les trois premiers formants sont essentiels pour caractériser le spectre vocal, les formants d'ordre supérieur ont une influence moindre.

La figure I-3 représente une trame de parole non voisée, on remarque qu'elle ne représente pas de structure périodique. Il peut être considéré comme un bruit blanc filtré par les lèvres, son spectre ne présente donc pas de structure de *pitch*.

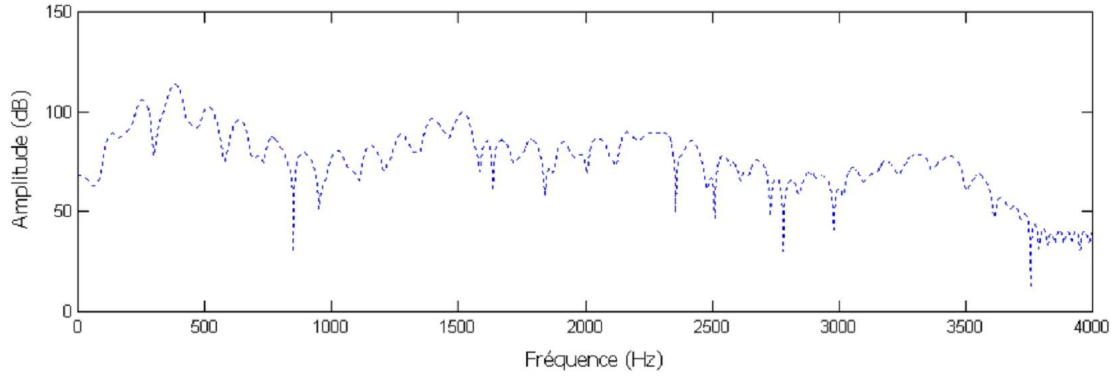


Figure I-2 : Spectre d'un signal vocal voisé

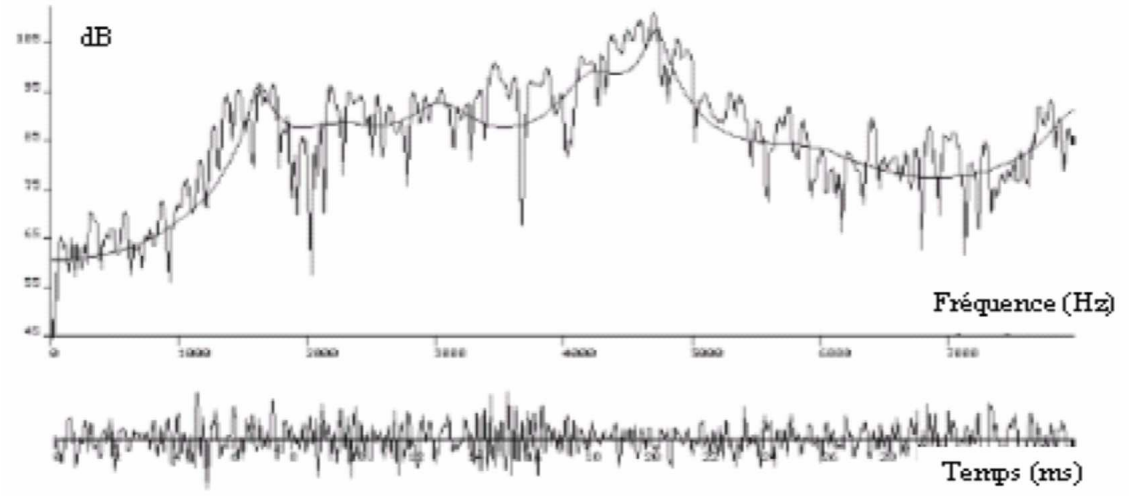


Figure I-3 : Un signal vocal non voisé et son spectre

I.1.3 Perception de la parole et redondance

Le système auditif fonctionne comme un filtre passe bande, la perception auditive est plus sensible dans la gamme de fréquence appartenant à l'intervalle [200-5600] Hz [1].

Les caractéristiques de la perception de la parole par le système auditif doivent être prises en compte lors de l'analyse temporelle ou spectrale de la parole afin d'améliorer l'efficacité des algorithmes de codage de la parole. Les aspects simples exploitables dans les codeurs de parole sont :

• La sensibilité de phase : les composantes de phase d'un signal de parole jouent un rôle négligeable dans la perception du langage [6]. L'oreille humaine perçoit principalement la parole grâce aux informations du spectre d'amplitude, ce qui justifie l'utilisation fréquente de systèmes à minimum de phase pour représenter un système vocal qui ne l'est pas.

• La perception de la forme spectrale : sachant que les pics du spectre sont plus importants pour la perception que les vallées du spectre [7], les méthodes d'évaluation devront modéliser, au mieux et en priorité, les formants du signal.

• Masquage de fréquences : tout bruit inférieur à un seuil pourra être masqué par le signal utile et devenir inaudible. Pour cela, un masque fréquentiel dont la forme est semblable à l'enveloppe spectrale du signal est utilisé. Des techniques efficaces de compression codent le bruit en fonction de ce seuil, ou son approximation, et réduisent au minimum la distorsion perceptive audible.

La perception ne dépend pas uniquement de la détection des informations audibles par l'oreille mais aussi du contexte, l'attention de l'auditeur, la familiarité de l'auditeur avec l'orateur, le sujet de conversation et la présence de bruit.

Le caractère redondant du signal parole aide à la perception dans un environnement bruité ou dans une conversation simple avec un interlocuteur familier. Le codage prédictif pourra donc exploiter la redondance du signal parole pour réduire le débit.

I.1.4 Modélisation de la parole

Pour réduire le débit en conservant une qualité suffisante ou pour améliorer la qualité pour un débit imposé, il faut chercher les paramètres pertinents qui constituent le signal parole grâce à l'analyse de la parole. Pour cela un modèle simplifié du système phonatoire ne retenant que les paramètres les plus significatifs du signal est recherché. Dans une grande majorité de codeurs, la production de parole est modélisée par une opération de filtrage [1], où une source excite un filtre censé représenter le conduit vocal. Une modélisation précise du conduit vocal et de la source d'excitation est nécessaire pour produire respectivement un signal intelligible et naturel. Le modèle électrique linéaire à été proposé par *Fant* [8] en 1960.

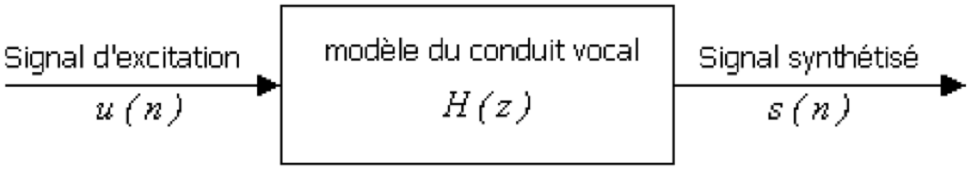


Figure I-4 : Modélisation de la production de la parole

Un signal d'excitation et la modélisation du conduit vocal permettent donc de caractériser la génération de la voix (figure I-4). Le modèle du conduit vocal $H(z)$ est excité par un signal glottal discret $U(z)$ pour produire un signal de parole $S(z)$:

$$S(z) = H(z)\,U(z) \qquad \text{(I.1)}$$

Le signal est donc modélisé par la sortie d'un filtre de corrélation court terme LPC (*Linear Predictive Coding*) dont la fonction est de modéliser l'enveloppe spectrale du signal, excité par un train d'impulsions périodiques pour les sons voisés et un bruit blanc pour les sons non voisés (figure I-5).

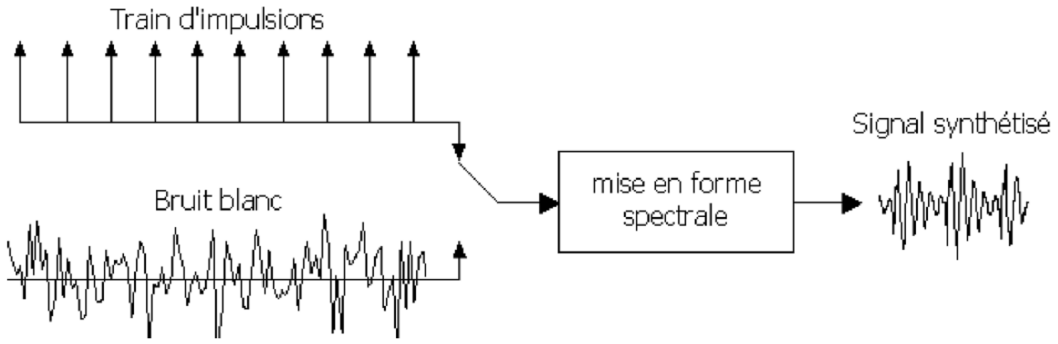


Figure I-5 : Synthèse de parole à deux états d'excitation

La figure I-6(a) donne le signal d'excitation optimal à l'entrée du filtre LPC pour obtenir le signal synthétisé figure I-6(b).

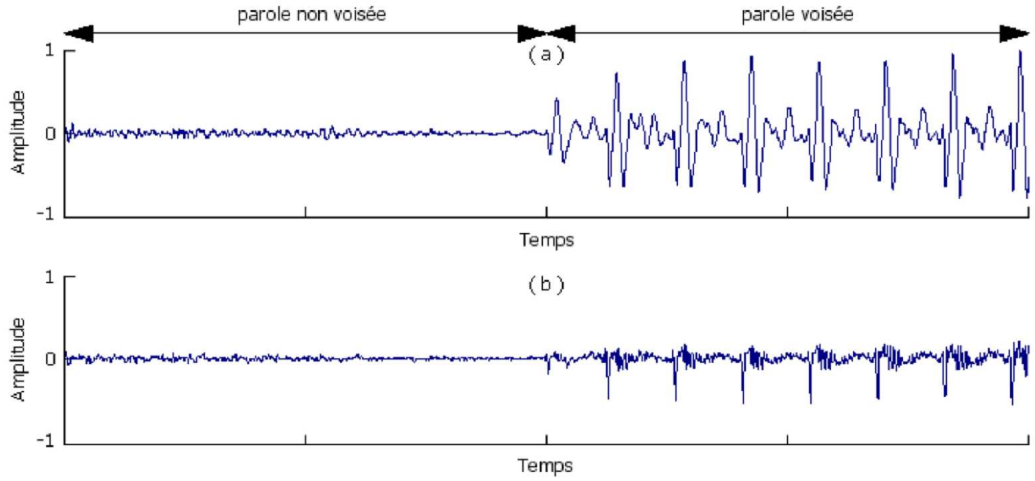


Figure I-6 :(a) signal de parole voisée et non voisée
(b) signal d'excitation correspondant

Notons que le signal d'excitation ressemble bien à un bruit blanc pour le segment de signal non voisé et à un train d'impulsions pour la partie parole voisée.

I.1.5 Codage prédictif et analyse par synthèse

La plus part des méthodes actuelles permettant des taux de compression significatifs cherchent à exploiter la redondance propre du signal. Il suffira alors de ne transmettre au décodeur que l'information non prédictible. Le modèle de synthèse, défini précédemment, va permettre de donner la structure d'un codeur de parole basé sur le codage prédictif et l'analyse par synthèse (LPAS *: Linear Prediction Analysis by Synthesis*) [9] [10]. Dans le codage LPAS (figure I-7), le signal d'entrée est analysé et un signal d'excitation est déterminé. La fonction

du codage prédictif consiste à définir les coefficients du filtre de prédiction tandis que le signal d'excitation est modélisé par l'analyse par synthèse. L'erreur entre le signal d'entrée et celui mis en forme par le filtre de synthèse, reproduisant les résonances (formants) du conduit vocal, est alors minimisée par le critère des moindre carrés (*MMSE* : *Minimum Mean Square Error*) pou choisir le meilleur signal d'excitation.

De par leur construction, les codeurs prédictifs ne sont plus des codeurs temporels. En effet, les données à transmettre ne sont plus des valeurs du signal échantillonné mais des paramètres résultant de la prédiction linéaire et du codage du signal d'excitation. Ces opérations effectuées pour le calcul des différents paramètres sont répétés à toutes les trames, voire plusieurs fois par trame. Les paramètres correspondants sont alors transmis au décodeur qui reconstruira le signal de parole grâce à la même structure de synthèse.

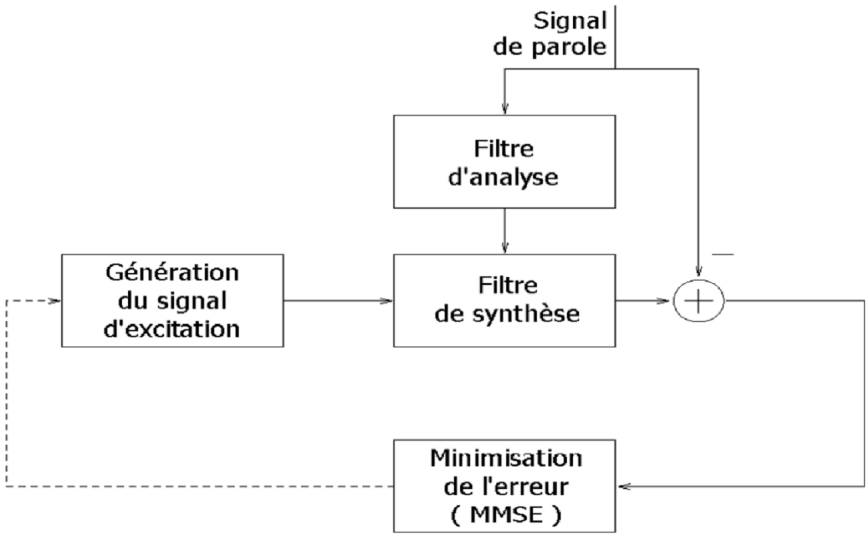


Figure I-7 : Diagramme du codeur LPAS

Une des raisons du succès du codage LPAS est la possibilité d'incorporer dans sa structure une fonction qui prend en compte la perception de l'appareil auditif humain. Ce principe a pour but de minimiser un critère d'erreur plus subjectif entre le signal de parole réel et sa modélisation paramétrique. Ceci est réalisé par pondération des fréquences perceptuelles sur le signal d'erreur pendant la sélection du signal d'excitation.

I.1.5.1 Prédiction linéaire

Le codage de parole à bas débit exige des représentations compactes et précises du filtre du conduit vocal et de la source d'excitation. C'est pourquoi la prédiction linéaire (LP : *Linear Prediction*), qui exploite les redondances du signal de parole en le modélisant par un

nombre restreint de paramètres sous la forme d'un filtre linéaire [11],est un des outils les plus important de l'analyse de parole [12]. L'idée de base du codage prédictif linéaire LPC (*Linear Predictive Coding*) est de considérer que tout échantillon de parole peut être exprimé comme une combinaison linéaire d'échantillons antérieurs. Un ensemble unique de coefficients prédicteurs peut alors être déterminer et utilisé pour supprimer les redondances à court terme du signal.

Lors d'une analyse à court terme, la redondance proche entre les échantillons du signal de parole est supprimée par le filtre d'analyse LP, représentant le conduit vocal. Ce filtre permet d'extraire la structure des *formants* du signal d'entrée et d'obtenir un signal de sortie de faible énergie, correspondant à l'erreur de prédiction appelée signal résiduel ou excitation.

Le filtre inverse du filtre d'analyse est le filtre de synthèse LP dont la fonction de transfert décrit l'enveloppe spectrale du signal parole. Chaque trame de parole est donc modélisée en sortie du système linéaire LP par un signal d'excitation.

Le système LP (figure I-4) modélise le conduit vocal qui caractérise la production de la voix. Le modèle $H(z)$ le plus utilisé pour représenter le conduit vocal est un modèle pôle-zéro ou AutoRégressif à Moyenne Ajustée ARMA *(Auto Regressive Movieng Average*). Le signal $s(n)$ est alors une combinaison linéaire de ses échantillons antérieurs et du signal d'excitation du système $u(n)$:

$$s(n) = G\sum_{l=0}^{q} b_l u(n-l) - \sum_{k=1}^{p} a_k s(n-k) \tag{I.2}$$

Où le gain G et les coefficients $\{a_k\}$et $\{b_k\}$du filtre LP sont les paramètres du système avec $b_0 = 1.$ Les p échantillons passés étant considérés, p est l'ordre de prédiction linéaire.

La fonction de transfert du système $H_{ARMA}(z)$, dont les pôles du filtre représentent les formants du spectre est donnée par :

$$H_{ARMA}(z) = \frac{S(z)}{U(z)} = G\frac{1+\sum_{l=1}^{q} b_l z^{-l}}{1+\sum_{k=1}^{p} a_k z^{-k}} \tag{I.3}$$

Avec $S(z) = \sum_{n=-\infty}^{n=+\infty} s(n)z^{-n}$ la transformée en Z de $s(n)$ et $U(n)$ la transformée en Z de $u(n)$.

Deux cas particuliers sont alors possibles pour ce filtre :

• Si $a_k = 0$ pour k = 1,…, p, $H_{ARMA}(z)$ devient un modèle tout-zéro ou à Moyenne Ajustée (*MA* : *Moving Average*). Ce filtre $H_{MA}(z)$ permet de modéliser parfaitement les segments du

signal dont le spectre tend vers zéro. Ce pendant, les zéros qui transcrivent les sons non voisés sont difficiles à déterminer car un ensemble d'équations non linéaires doit être résolu.

• Si $b_l = 0$ pour l = 1,…, q, $H_{ARMA}(z)$ est réduit à un modèle tout-pôle appelé modèle AutoRégressif (*AR* : *Auto Regressive*). Le modèle *AR*, qui représente parfaitement les résonances spectrales des sons voisés, est souvent utilisé pour sa simplicité et son efficacité dans des systèmes temps réel. La fonction de transfert du conduit vocal est alors approximée par le modèle tout-pôle $H_{AR}(z)$:

$$H_{AR}(z) = \frac{G}{1 + \sum_{k=1}^{p} a_k z^{-k}} \qquad \text{I.4}$$

Le facteur de gain G est généralement égal à 1, pour obtenir le filtre $H_{AR}(z) = \frac{1}{A(z)}$ avec le polynôme $A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k}$ dont les paramètres $\{a_k\}$ sont appelés coefficient LP (figure I-9). Ils sont déterminés, à l'aide de différentes techniques telles que les méthodes d'autocorrélation et de covariance [1], à partir d'un système à *p* inconnues de telle sorte que le signal reconstitué soit le plus proche possible du signal original.

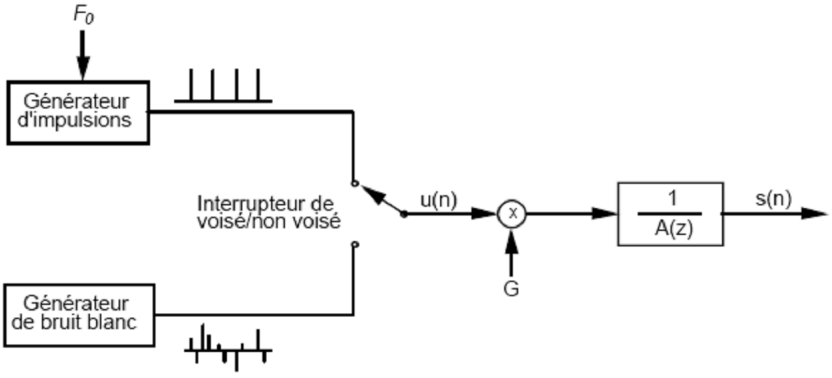


Figure I-8 Modèle simplifié de production de la parole [8][13]

La modélisation complète peut maintenant être décomposée en deux parties (figure I-9) :

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

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

$$e(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k) \qquad \text{I.5}$$

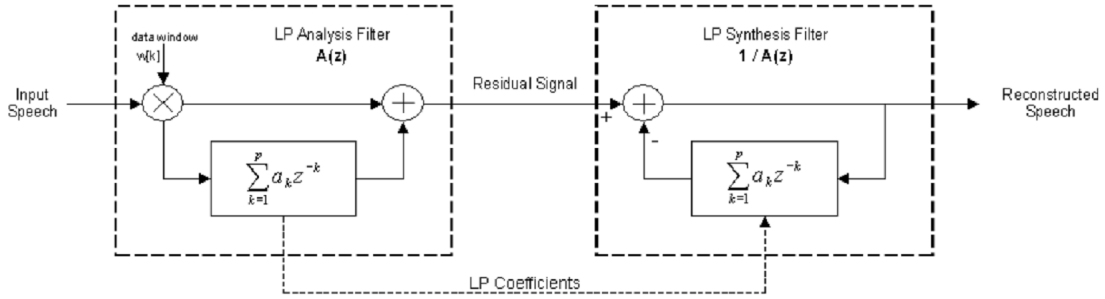


Figure I-9 : Analyse et synthèse LP

Le signal résiduel e est l'excitation idéale du modèle LPC du conduit vocal $H_{AR}(z)$. Une modélisation précise de ce signal permet d'obtenir un signal reconstruit naturel. Or l'estimation des paramètres LP du modèle vocal entraîne une approximation du signal d'excitation. A mesure que l'ordre du modèle LPC augmente un meilleur ajustement au spectre de la voix est obtenu. Cependant, plus l'ordre *p* sera élevé, plus le nombre de paramètres à transmettre augmentera. L'ordre doit ainsi résulter d'un compromis entre une bonne représentation de la structure formantique du signal de parole, la complexité de calcul et le débit de transmission. En général, deux pôles sont nécessaires pour représenter chaque formant et jusqu'à quatre autres sont employés pour approximer les vallées du spectre.

L'amélioration de la prédiction devient minime lorsque l'ordre *p* est supérieur à 10. Des modèles LPC les plus courants utilisent un nombre de coefficients de l'ordre de 8 à 16.

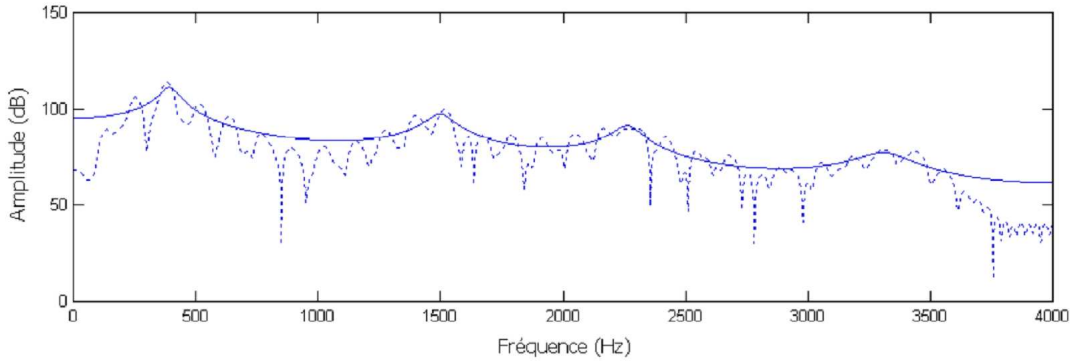


Figure I-10 : Modélisation LPC d'une voix d'homme

La figure I-10 montre un exemple de modélisation LPC appliqué à une section de voix masculine [19]. La réponse fréquentielle du filtre $H_{AR}(z)$ à l'ordre p = 10 (trait plein) suit parfaitement l'enveloppe spectrale du signal (trait pointillé). La modélisation accentue quatre pics sur le spectre, situés à environ 400, 1500, 2300 et 3300 Hz, qui correspondent aux fréquences de résonances. De plus, notons que le modèle de LPC s'adapte mieux aux crêtes (formants) qu'aux vallées (anti-formants). Ceci est dû à la plus grande contribution des formants dans le critère de minimisation d'erreur quadratique moyenne (MMSE) [1].

I.1.5.2 Estimation des coefficients de prédiction linéaire

Dans la synthèse de la parole, le signal parole est produit par un modèle LP. Pour se faire chaque paramètre à besoin d'être estimé. Cette estimation se compose de deux parties :

• Estimation des paramètres du filtre LP.

• Estimation des paramètres du signal d'excitation du filtre LP (spécialement le pitch).

I.1.5.2.1 Fenêtrage

Les caractéristiques du filtre du conduit vocal et de la parole évoluent avec le temps, un signal vocal est non stationnaire. Or les caractéristiques spectrales du signal de parole doivent évoluer très peu pendant l'analyse (stationnarité locale). Le signal étant généralement considéré comme quasi-stationnaire sur de faibles périodes de l'ordre de quelques dizaines de millisecondes (entre 5 et 20 ms) [1], une segmentation est effectuée en multipliant le signal de parole numérique $s(n)$ par une fenêtre d'analyse $w(n)$.

Soit $w(n)$ la fenêtre d'analyse de longueur finie correspondant à N échantillons, on obtient le segment du signal de parole fenêtré $s_w(n)$ [1] suivant :

$$s_w(n) = w(n)s(n) \qquad \text{I.6}$$

La fenêtre d'analyse la plus utilisée dans le codage de la parole est la fenêtre de *Hamming* :

$$w(n) = \begin{cases} 0.54 - 0.46\cos\left(\dfrac{2\pi n}{399}\right) & n = 0,\cdots,199 \\ \cos\left(\dfrac{2\pi\,(n-200)}{159}\right) & n = 200,\cdots,239 \end{cases} \qquad \text{I.7}$$

I.1.5.2.2 Méthode d'autocorrélation

Après avoir multiplié le signal $s(n)$ par la fenêtre $w(n)$ est obtenu le signal fenêtré $s_w(n)$:

$$s_w(n) = \begin{cases} w(n).s(n) & pour \quad 0 \le n \le N-1 \\ 0 & ailleurs \end{cases} \qquad \text{I.8}$$

Nous allons minimiser l'énergie dans le signal résiduel *E* définie comme suit :

$$E = \sum_{n=-\infty}^{\infty} e^2(n) = \sum_{n=-\infty}^{\infty} \left(s_w(n) + \sum_{k=1}^{p} a_k s_w(n-k) \right)^2 \qquad \text{I.9}$$

Les valeurs des $\{a_k\}$ qui minimisent *E* sont obtenues en annulant les dérivées partielles de *E* par rapport à aux coefficients du filtre $\{a_k\}$. La méthode d'autocorrélation revient mettre $\frac{\partial E}{\partial a_k} = 0$, pour k=1, …, *p*, on aura *p* équations avec *p* inconnues variables $\{a_k\}$comme suit :

$$\sum_{k=1}^{p} a_k \sum_{n=-\infty}^{\infty} s_w(n-i)s_w(n-k) + \sum_{n=-\infty}^{\infty} s_w(n-i)s_w(n) = 0 \qquad 1 \le i \le p \qquad \text{I.10}$$

Dans l'équation (I.10), le signal fenêtré $s_w(n)$=0 hors de la fenêtre $w(n)$. Les équations linéaires peuvent être exprimées en termes de fonctions d'autocorrélation. Car la fonction d'autocorrélation du segment fenêtré $s_w(n)$ est définie comme :

$$R(i) = \sum_{n=i}^{N-1} s_w(n)s_w(n-i) \qquad 0 \le i \le p \qquad \text{I.11}$$

Où *N* est la longueur de la fenêtre. La fonction d'autocorrélation est une fonction paire : $R(i)=R(-i)$. Des équations I.10 et I.11 on obtient les équations linéaires suivantes :

$$\sum_{k=1}^{p} a_k R(|i-k|) = -R(i) \qquad 0 \le i \le p \qquad \text{I.12}$$

Ce système peut se mettre sous forme matricielle par ***Ra = v*** comme suit :

$$\begin{bmatrix} R(0) & R(1) & \cdots & R(p-1) \\ R(1) & P(0) & \cdots & R(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ R(p-1) & R(p-2) & \cdots & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = - \begin{bmatrix} R(1) \\ R(2) \\ \vdots \\ R(p) \end{bmatrix} \qquad \text{I.13}$$

La matrice d'autocorrélation *R* de dimension $(p \times p)$ dont tous les éléments de la diagonale sont égaux est symétrique, elle présente la propriété d'être une matrice de *Toeplitz* symétrique. Le vecteur des coefficients de prédiction $[a_1 \cdots a_p]^t$ peut alors être obtenu par inversion de la matrice *R* en utilisant la méthode de *Levinson-Durbin* [14] (annexe A) qui tire profit de la structure symétrique de la matrice d'autocorrélation.

Notons que la résolution de ce système d'équations permet de garantir le minimum de phase au filtre de prédiction linéaire $A(z)$ (tous les zéros du filtre sont à l'intérieur du cercle unité), ce qui entraîne la stabilité du filtre de synthèse LP, essentielle pour obtenir un signal de parole reconstruit de bonne qualité[15].

I.1.5.2.3 Méthode de covariance

La méthode de covariance est très similaire à celle d'autocorrélation. La majeure différence réside dans l'emplacement de la fenêtre d'analyse. Dans la méthode de covariance le fenêtrage se fait au niveau du signal résiduel au lieu du signal parole original. L'énergie E signal résiduel fenêtré est :

$$E = \sum_{n=-\infty}^{\infty} e_w^2(n) = \sum_{n=-\infty}^{\infty} e^2(n) w^2(n) \qquad \text{I.14}$$

En mettant les dérivées partielles de l'énergie par rapport aux coefficients $\{a_k\}$ à zéro ($\frac{\partial E}{\partial a_k} = 0$) on obtiendra les p équations linéaires suivantes :

$$\sum_{k=1}^{p} \Phi(i,k) a_k = -\Phi(i,0) \qquad 1 \le i \le p \qquad \text{I.15}$$

Où $\Phi(i,k)$ est la fonction de covariance de $s(n)$:

$$\Phi(i,k) = \sum_{n=-\infty}^{\infty} w(n) s(n-i) s(n-k) \qquad \text{I.16}$$

Les p équations peuvent être exprimées sous forme matricielle : $\Phi \cdot a = \Psi$

$$\begin{bmatrix} \Phi(1,1) & \Phi(1,2) & \cdots & \Phi(1,p) \\ \Phi(2,1) & \Phi(2,2) & \cdots & \Phi(2,p) \\ \vdots & \vdots & \ddots & \vdots \\ \Phi(p,1) & \Phi(p,2) & \cdots & \Phi(p,p) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = - \begin{bmatrix} \Psi(1) \\ \Psi(2) \\ \vdots \\ \Psi(p) \end{bmatrix} \qquad \text{I.17}$$

Tel que : $\Psi(i) = \Phi(i,0) \qquad 1 \le i \le p$

La matrice Φ n'est pas une matrice de *Toeplitz*, elle est symétrique et définie positive mais ne garantie pas la stabilité du filtre LPC. Ces p équations ne peuvent pas être résolues par l'algorithme de *Levinson-Durbin*. La matrice de covariance peut être décomposée en deux matrices, l'une triangulaire inférieure L et l'autre triangulaire supérieure U.

$$\Phi = L \cdot U \qquad \text{I.18}$$

La décomposition de *Cholesky* peut être utilisée pour convertir la matrice de covariance sous la forme :

$$\Phi = C \cdot C^T \qquad \text{I.19}$$

Tel que : $C = L$ et $C^T = U$; On commence d'abord par résoudre l'équation suivante :

$$L \cdot y = \Psi \qquad \text{I.20}$$

Apres avoir obtenu les valeurs du vecteur *y*, on calculera le vecteur *a* comme suit :

$$U \cdot a = y \qquad \text{I.21}$$

I.1.5.2.4 Représentation des paramètres de prédiction

Dans tout codage de parole à bas débit, il est nécessaire, pour transmettre les coefficients de prédiction linéaire, de les quantifier en utilisant un nombre restreint de bits. Cependant, les coefficients LP ne sont guère appropriés à une quantification directe à cause de leur intervalle de définition (dynamique) trop important. Leur codage nécessiterait en effet 8 à 10 bits par coefficient [12]. En outre, il n'assurerait pas la stabilité du filtre de synthèse LP, importante pour la qualité de parole synthétisée. Afin de palier ces difficultés, des paramètres mathématiquement équivalents aux coefficients $\{a_k\}_{k=0\cdots p}$ d'ordre *p* sont calculés.

Les lignes de raies spectrales LSP (*Line Spectrum Pairs*) sont l'alternative de représentation des coefficients de prédiction linéaire la plus utilisée [16].les paramètres LSP fournissent les informations nécessaires à une bonne représentation du modèle LPC dans le domaine fréquentiel en localisant les formants et les vallées du spectre du signal de parole. Ainsi, les LSP, qui ont la propriété d'être ordonnés, se prêtent à une quantification robuste et efficace des coefficients LP [17]. Pour déterminer les lignes de raies spectrales, la technique utilisant les polynômes de *Tchebycheff* est la plus courante.

Il a été précédemment démontré que le filtre d'analyse LP pouvait être exprimé en fonction des coefficients LP comme un polynôme $A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k}$ d'ordre *p*. ce dernier est décomposé en polynômes symétrique *p*(*z*) et antisymétrique *Q*(*z*) qui vérifient l'égalité suivante :

$$A(z) = \frac{P(z) + Q(z)}{2} \qquad \text{I.22}$$

Les deux polynômes *p*(*z*) et *Q*(*z*) peuvent donc s'écrire en fonction des coefficients de prédiction linéaire $\{a_k\}_{k=0\cdots p}$ comme suit :

$$P(z) = A(z) + z^{-(p+1)} A(z^{-1}) \qquad \text{I.23}$$

$$P(z) = A(z) - z^{-(p+1)} A(z^{-1}) \qquad \text{I.24}$$

Les polynômes *P*(*z*) et *Q*(*z*) possédant respectivement une racine connue pour $z = e^{j\Pi} = -1$ et $z = e^{j0} = 1$, ils peuvent être réécrits de manière simplifiée sous la forme :

$$P(z) = (1 + z^{-1})\, P'(z) \qquad \text{I.25}$$

$$Q(z) = (1 - z^{-1})\, Q'(z) \qquad \text{I.26}$$

Compte tenu que $P(z)$ et $Q(z)$ possède des racines complexes conjuguées sur le cercle unité , seuls les $p/2$ premiers coefficients sont nécessaires pour spécifier chaque polynôme.

Les racines des polynômes $P(z)$ et $Q(z)$ sont données par les fréquences de raies spectrales $\{\omega_i\}_{i=1\cdots p}$, qui représente les p angles de position des zéros sur le cercle unité.

$P(z)$ et $Q(z)$ ont les trois propriétés suivantes [13] :

• Toutes les racines (zéros) des polynômes $P(z)$ et $Q(z)$ se trouvent sur le cercle unité.

• Les zéros de $P(z)$ et $Q(z)$ sont entrelacés $(0 < \omega_1 < \omega_2 < \cdots < \omega_p < \pi)$.

• Le minimum de phase est garanti au filtre $A(z)$ si les deux premières propriétés sont satisfaites donc la stabilité du filtre de synthèse est facilement vérifiable.

Soong et *Juang* [17] ont montrés que si $H(z)$ est stable, où $A(z)$ est à phase minimale, alors les zéros des polynômes $P(z)$ et $Q(z)$ sont appelés les LSP. La figure I-11 illustre un exemple de localisation des racines des deux polynômes pour un ordre p pair. Notons que les racines en 0 et π ne sont pas représentées.

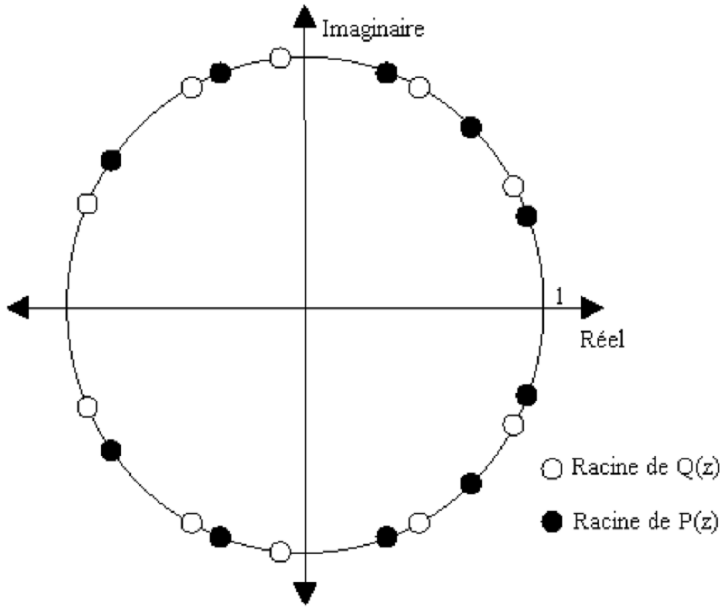


Figure I-11 Localisation possible des racines pour $P(z)$ et $Q(z)$ d'ordre pair

En fait, les polynômes $P'(z)$ et $Q'(z)$ pouvant s'écrire en fonction de cosinus des fréquences de raies spectrales, l'estimation des coefficients $\{\omega_i\}_{i=1\cdots p}$ revient à estimer les p racines de $P''(\omega)$ et $Q''(\omega)$:

$$P'(\omega) = P'(z)\big|_{e^{j\omega}} = 2e^{-j\frac{p}{2}\omega} P''(\omega) \quad \text{I.27}$$

$$Q'(\omega) = Q'(z)\big|_{e^{j\omega}} = 2e^{-j\frac{p}{2}\omega} Q''(z) \quad \text{I.28}$$

Pour rechercher les zéros des deux polynômes $P''(z)$ et $Q'(z)$, *Kabal* et *Ramachandran* [18] ont proposé un algorithme, utilisant les polynômes de *Tchebicheff*, qui pourra être facilement implémenté en temps réel sur un processeur.

I.1.5.3 Considérations pratiques

Pour bien mener l'analyse LPC, il faut choisir :

• La fréquence d'échantillonnage f_e.

• La méthode d'analyse et l'algorithme correspondant.

• L'ordre *p* de l'analyse LPC.

• Le nombre d'échantillons par tranche *N* et le décalage entre tranches successives *L*.

Le choix de la fréquence d'échantillonnage est fonction de l'application visée et de la qualité du signal à analyser :

• 8 kHz pour les signaux téléphoniques.

• 10 kHz pour les applications de reconnaissance.

• 16 kHz pour les applications de synthèse.

L'ordre de prédiction *p* est choisi de façon à ce qu'il permette de bien représenter toute la séquence du signal parole; l'ordre *p* est fonction de la fréquence d'échantillonnage, on estime en général qu'une paire de pôles est nécessaire par 1Khz de bande passante.

Lorsque la fréquence d'échantillonnage f_e est exprimée en échantillons/sec, une période de 1ms correspond à f_e/1000 échantillons.

A la fréquence d'échantillonnage de 8 kHz, la valeur correspondante de *p* doit être au moins égale à 8. Elle trouve d'ailleurs une justification expérimentale dans le fait que l'énergie de l'erreur de prédiction diminue rapidement lorsqu'on augmente *p* à partir de 1, pour tendre vers une asymptote au voisinage de ces valeurs : il devient inutile d'augmenter encore l'ordre, puisqu'on ne prédit rien de plus.

De plus la durée des trames d'analyse et leur décalage sont souvent fixés inférieur à 30ms.Les valeurs choisies sont liées au caractère quasi-stationnaire du signal parole.

Enfin, comme vu précédemment dans la méthode d'autocorrélation, pour compenser les effets de bord, on multiplie en général préalablement chaque tranche d'analyse par une fenêtre de pondération *w(n)*, la plus souvent utilisées est celle de *Hamming* (équation (I.7)).

I.1.5.4 Pondération perceptive

Le principe de l'analyse par synthèse, utilisée pour déterminer le signal d'excitation, est de minimiser un critère d'erreur entre une séquence du signal de parole *s* et sa modélisation paramétrique $\hat{s}$ (figure I-8). L'excitation optimale sera par exemple établie par minimisation de l'erreur quadratique moyenne, définie comme l'espérance mathématique $E\left\{(s(n)-\hat{s}(n))^2\right\}$. Cependant, ce critère n'étant pas bien adapté à notre système auditif, une correction peut être intégrée pour palier cet inconvénient. Dans de nombreux cas, le critère d'erreur est modifié par une pondération perceptive dont le but est de contrôler la répartition fréquentielle du bruit, qui s'étend généralement sur tout le spectre.

Le caractère physiologique de l'oreille possède la propriété intéressante d'enregistrer les sons d'une manière fréquentielle. De cette propriété découle la pondération perceptive qui a pour fonction de répartir la puissance du bruit d'une fréquence à une autre. Cette répartition est efficace lorsque la distribution de la puissance du bruit est augmentée au niveau des formants. Dans ce cas, le bruit est masqué par la puissance du signal et l'écoute sera plus agréable. Le nouveau critère, appelé critère d'erreur subjectif, est défini comme la minimisation de l'erreur $\left[(s(z)-\hat{s}(z))\frac{A(z)}{A_\gamma(z)}\right]^2$ quadratique avec le filtre de pondération perceptive donné par :

$$\frac{A(z)}{A_\gamma(z)}=\frac{1+\sum_{k=1}^{p}a_k z^{-k}}{1+\sum_{k=1}^{p}a_k \gamma^k z^{-k}} \qquad \text{I.29}$$

Le filtre de pondération est construit à partir du filtre de prédiction à court terme *A*(*z*). Ce choix rend la répartition spectrale du bruit proportionnelle à celle du signal de parole. Toutefois le système auditif ne réagissant pas avec la même sensibilité pour toutes les fréquences, il est nécessaire d'introduire un coefficient γ pour contrôler au mieux la répartition de la puissance du bruit. En effet, dans les régions d'anti-formant, la puissance du signal étant faible, le bruit, est à l'inverse, fortement audible [9].

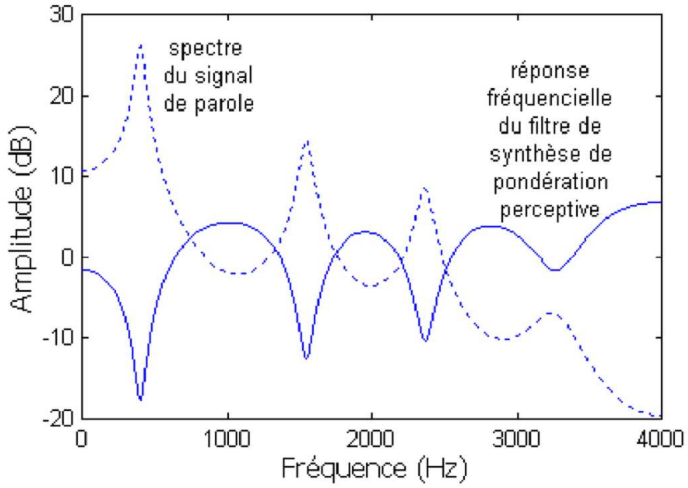


Figure I-12 : Réponse fréquentielle du filtre de pondération perceptive

La figure I-12 donne l'exemple d'un spectre de signal de parole et de la réponse en fréquence du filtre de synthèse de pondération perceptive associée. Le fait d'atténuer le rapport signal à bruit entre le signal d'excitation et le bruit au niveau des formants et de l'augmenter dans les régions d'anti-formant peut être ainsi interprété comme un effet masquant pour l'oreille humaine.

I.2 Principe de quantification

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

La quantification est le processus de substitution des échantillons d'un signal analogique par des valeurs arrondies prises parmi un nombre fini de valeurs possibles [13].

La quantification peut être *scalaire* ou *vectorielle* selon que les signaux sont à une ou plusieurs dimensions.

I.2.1 Quantification scalaire

Dans la quantification scalaire (*QS),* chaque échantillon du signal d'entrée est quantifié séparément des autres échantillons. Comme l'illustre la figure I-13, un échantillon *x* du signal d'entrée est spécifié par l'indice *i* s'il se trouve dans l'intervalle suivant :

$$I_i : \{x_i < x \le x_{i+1}\} \qquad i = 1, \cdots, N \qquad \text{I.30}$$

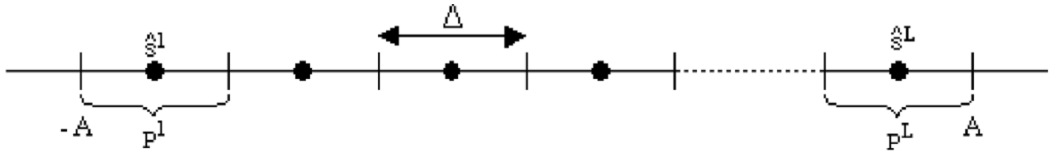


Figure I-13 Quantification scalaire

Tous les échantillons situés dans l'intervalle I_i seront remplacés par une valeur y_i appelée *niveau de reconstruction* ou *représentant*. Si les N intervalles distincts I_i sont de même longueur la quantification est dite uniforme sinon elle est non uniforme.

I.2.2 Quantification vectorielle

La quantification vectorielle (*VQ*) est l'extension de la quantification scalaire à un espace multidimensionnel.

Nous appellerons quantificateur vectoriel de dimension *m* à *N* niveaux une application *Q* qui, à un vecteur d'entrée $x = \{x_1, x_2, \cdots, x_m\}$, fait correspondre une valeur approchée *y* choisie dans un ensemble fini de *N* éléments $y = \{y_i, i = 0,1, \cdots, N-1\}$

L'ensemble *y* est un dictionnaire de *N* représentants. En posant $R = \log_2(N)$, nous dirons que les vecteurs d'entrées sont quantifiés sur *N* niveaux et codés avec *R* bits.

Contrairement à la quantification scalaire, un quantificateur vectoriel peut fonctionner avec un débit fractionnaire [13].

I.3 Techniques de codages de la parole

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

De nombreux travaux relatifs aux algorithmes de codage tendent à maximiser le compromis entre l'efficacité, le coût et la qualité de transmission des systèmes de communication en fonction des débits disponibles. Les techniques de codage numérique adaptent donc leur débit de transmission aux capacités du canal.

Les algorithmes de codage de parole basiques peuvent être divisés en trois classes distinctes [20] :

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

• Codage paramétrique (parametric coding).

• Codage hybride (hybrid coding).

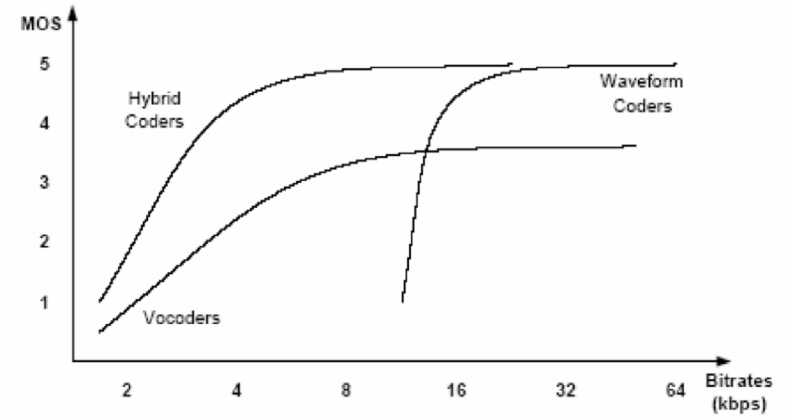


Figure I-14 : Comparaison de la qualité de codage de parole [20]

I.3.1 Le codage de forme d'onde

Les codeurs de formes d'ondes n'utilisent aucune connaissance à priori sur la façon dont le signal est généré, ils sont conçus pour être indépendant du signal à coder et sont relativement simples à mettre en œuvre. Ils produisent une qualité acceptable jusqu'à des débits de 16 Kbits/s. en deçà, la qualité du signal reconstruit se dégrade rapidement. Le signal reconstruit est sans doute le plus proche du signal original.

Le débit de codage est généralement élevé. En utilisant les propriétés de corrélation du signal, il est possible de diminuer ce débit jusqu'à une certaine limite. En dessous de 16kbits/s la qualité se dégrade et la réduction de débit en bande étroite (2.4-4.8 kbits/s) est peu envisagée.

L'algorithme de codage le plus simple est le codage appelé *PCM* (*Pulse Coded Modulation*) qui revient seulement à échantillonner un signal analogique à une fréquence de 8 kHz et à quantifier (c'est-à-dire à convertir des valeurs réels en valeurs de précision finie) les échantillons sur 8 bits, il correspondant à la norme G.721, il est utilisé pour coder la voix dans le réseau téléphonique avec un débit de 64 kbits/s.

Le codage *PCM* est à la famille de codages différentiels qui est basé sur l'observation des échantillons successifs d'une source audio qui sont fortement corrélés. Il semble donc judicieux d'encoder non pas les échantillons eux même mais la différence entre les échantillons successifs. On peut citer :

- Le codage DPCM (*Differential PCM*).
- Le codage ADPCM (*Adaptive Differential PCM*).
- Le codage ADM (*Adaptive Delta Modulation*).

I.3.2 Le codage paramétrique

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

I.3.3 Le codage hybride

L'utilisation des techniques de codage de formes d'ondes conduit à une excellente qualité de la parole mais induit un débit assez élevé. 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 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 de formes d'ondes 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 de forme d'onde modélisent respectivement le conduit vocal et le signal d'erreur résiduel.

A partir des années 80, l'intérêt pour les codeurs CELP (*Code-Exited Linear Prediction*), qui détermine une forme d'onde optimale du signal d'erreur en utilisant l'analyse par synthèse, ne cesse d'augmenter. Ces codeurs sont basés sur les algorithmes de codage de la parole les plus actuellement utilisés dans la téléphonie sans fil. Le standard G.729 de l'IUT est un codeur CELP qui produit une qualité téléphonique de parole (toll quality) de la parole à 8 kbits/s [21].

I.4 Qualité des codeurs

Pour conclure ce chapitre, nous allons énumérer les différents critères couramment utilisés pour juger et classer les méthodes de codage tels que :

- Débit de transmission.
- Qualité de la parole.
- Complexité de calcul et d'implémentation.
- Robustesse face aux erreurs.
- Délai de codage.

L'une des considérations les plus importantes dans tout codage de parole est la qualité du signal reconstruit, l'estimation de la qualité d'un codeur est un problème complexe. Une première approche consiste à utiliser une mesure objective de la ressemblance qui existe entre le signal original et le signal reconstitué. Cette méthodologie se situe dans le domaine des tests dits « objectifs ». Ils s'appliquent très bien aux codeurs de bonne qualité et font plutôt appel à la théorie du signal qu'aux connaissances sur la parole.

Lorsque l'on cherche une évaluation plus fine des codeurs, il faut faire appel à la dimension subjective de la qualité de la parole. Etant donné la part de subjectivité qui est présente dans l'appréciation d'un individu, il faut utiliser des procédures de test très élaborées. L'évaluation d'un codeur à l'aide de tests subjectifs est une opération délicate qui est généralement confiée à des laboratoires spécialisés.

D'autres critères sont couramment utilisés pour juger et classer les méthodes de codage

I.4.1 Mesure de distorsion subjective

L'évaluation subjective est obtenue par des tests d'écoutes du signal de parole où la qualité de la parole est mesurée par l'intelligibilité spécifiquement définie par le pourcentage de mots correctement écoutés et avec une sonorité naturelle. La perception des caractéristiques de la parole tend à changer considérablement entre les différents auditeurs,

mais ces essais restent toutefois utiles pour pointer sur différents aspects déterminés unanimement.

Il existe trois types de mesures subjectives [21] de la qualité généralement utilisées :

• Le test DRT (*Diagnostic Rhyme Test*)

• Le test DAM (*Diagnostic Acceptability Measure*)

• Le test MOS (*Mean Opinion Score*)

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

Tableau I-1 : Qualité avec la mesure MOS

I.4.2 Mesure de distorsion objective

Le système auditif de l'être humain est l'estimateur le plus adéquat de la qualité et des performances d'un codeur de la parole. Il permet de préciser l'intelligibilité et la sonorité naturelle des sons. Bien que, Les tests d'écoute subjectifs donnent une bonne évaluation pour les codeurs de la parole, ils peuvent exiger beaucoup de temps et sont non conformé. Les mesures objectives peuvent donner une estimation immédiate de la qualité perceptuelle de la parole [22], elles utilisent des fonctions ou des critères mathématiques pour comparer les formes d'ondes codées et originales telles que des mesures de distorsion ou de gain.

Les mesures objectives de distorsions peuvent être calculées aussi bien dans le domaine temporel que fréquentiel [21].

Les performances d'une mesure objective résident dans sa corrélation avec la mesure subjective correspondante (qualité ou intelligibilité).

Les mesures de distorsion sont classifiées comme suit [4] [13] :

• Domaine temporel (*RSB* et RSB_{seg}).

• Domaine fréquentiel (*distorsion spectrale*).

I.4.2.1 Domaine temporel

- Rapport Signal sur Bruit :

Si $\{S(n)\}_{n=0\cdots N_T}$ sont les N_t échantillons du signal parole original et $\{\tilde{S}(n)\}_{n=0\cdots N_t}$ sont les N_t échantillons du signal parole codé dans le *RSB* à la forme suivante :

$$RSB = 10\log\left(\frac{\sum_{n=0}^{N_t-1} S(n)^2}{\sum_{n=0}^{N_t-1}\left[S(n)-\tilde{S}(n)\right]^2}\right) \qquad (dB) \qquad \text{I.31}$$

Le *RSB* donne une valeur après avoir traité tout le fichier, donc il n'y a pas moyen de retrouver les instants où les divergences ont été enregistrées. De plus le *RSB* est dominé par la portion de forte énergie (tranches voisées), alors que le bruit a un effet perceptuel plus important sur les portions de faibles énergies.

- Rapport Signal sur Bruit Segmenté :

Le RSB_{seg} mesuré en dB, est la moyenne du *RSB* calculé sur de courts intervalles de temps du signal parole. Le RSB_{seg} calculé sur N_F trames de longueur N_S est donné par :

$$RSB_{seg} = \frac{1}{N_F}\sum_{i=0}^{N_F-1} 10\log\left(\frac{\sum_{j=0}^{N_S-1} S(N_S\, i+j)^2}{\sum_{j=0}^{N_S-1}\left[S(N_S\, i+j)-\tilde{S}(N_S\, i+j)\right]^2}\right) \qquad (dB) \qquad \text{I.32}$$

Le RSB_{seg} est meilleur que le *RSB*. Cependant, les tranches de silences renvoient de grandeurs négatives, biaisant de la sorte le résultat final. Ce problème peut être résolu en éliminant dans le calcul de la distorsion les trames de silence.

I.4.2.2 Domaine fréquentiel

La distorsion spectrale est définie comme étant la racine carré de la moyenne au carrée des différences entre le logarithmique décimal du spectre LPC original et le logarithme décimal du spectre LPC quantifier. La définition mathématique est comme suit :

$$DS_i = \sqrt{\frac{1}{F_e}\int_0^{F_e}\left[10\log_{10}\frac{S_i(f)}{\tilde{S}_i(f)}\right]^2 \partial f} \qquad (dB) \qquad \text{I.33}$$

Où F_e est la fréquence d'échantillonnage, $S_i(f)$ et $\tilde{S}(f)$ sont les spectres de la trame i donnés par :

$$S_i(f) = \frac{1}{A_i(e^{j2\pi f/F_e})} \quad \text{I.34}$$

$$\tilde{S}(f) = \frac{1}{\tilde{A}_i(e^{j2\pi f/F_e})} \quad \text{I.35}$$

Ou, $A_i(z)$ et $\tilde{A}(z)$ sont respectivement, les polynômes *PL* original et quantifié vus plus haut, pour la trame i, au lieu de l'intégration, une sommation des coefficients obtenus après application de la *TFD* (*Transformée de Fourier Discrète*) aux coefficients LPC, peut être utiliser pour calculer DS_i. La distorsion devient donc :

$$DS_i = \sqrt{\sum_{k=n_0}^{n_1-1}\left[10\log\frac{S_i\left(e^{j2\pi k/N}\right)}{S_i\left(e^{j2\pi k/N}\right)}\right]^2} \qquad (dB) \quad \text{I.36}$$

Une distorsion spectrale moyenne (la moyenne des distorsions spectrales calculées pour toutes les trames) de 1 dB est habituellement acceptée. Cependant, selon *Atal* et *Paliwal* les conditions de transparence spectrale (pas de distorsion audible) établies expérimentalement sont les suivantes :

- La moyenne *DS* inférieur à 1dB.
- Le nombre de trames ayant DS_i dans l'intervalle 2-4 dB est inférieur à 2%.
- Pas de trames ayant DS_i supérieur à 4 dB.

I.4.3 Mesure de distance euclidienne LSP pondérée

Cette distance a été développée dans le but d'optimiser la quantification des paramètres LSP, elle a la forme suivante :

$$d_{LSF} = \sum_{i=1}^{p}\left[c_i w_i(\omega_i - \tilde{\omega}_i)\right]^2 \quad \text{I.37}$$

Où c_i et w_i sont les poids du $i^{ème}$ coefficient LSP ω_i, et p l'ordre du filtre LP. Pour un filtre d'ordre 10, les poids fixes c_i sont donnés par :

$$c_i = \begin{cases} 1.0 & pour \quad 1 \le i \le 8 \\ 0.8 & pour \quad i = 9 \\ 0.4 & pour \quad i = 10 \end{cases} \quad \text{I.38}$$

Ces poids sont utilisés pour donner plus d'importance aux basses fréquences par rapport aux hautes fréquences. Ceci est justifié par le fait que l'oreille humaine est plus sensible aux basses fréquences qu'aux hautes fréquences. Les poids adaptatifs ωi sont utilises pour accentuer les régions de l'enveloppe spectrale $S(e^{j\omega})$ à forte énergie (formants). Ces poids sont données par :

$$w_i = \left[S(e^{jw})\right] \qquad \text{I.39}$$

On *r* est une constante empirique qui contrôle le degré de pondération, empiriquement *r* = 0.15. Une pondération plus simple a été proposée [23], elle a la forme suivante :

$$w_i = \frac{1}{\omega_i - \omega_{i-1}} + \frac{1}{\omega_{i+1} - \omega_i} \quad \text{où } \omega_0 = 0 \quad \text{et} \quad \omega_{p+1} = \pi \qquad \text{I.40}$$

Les mesures dans le domaine perceptuel sont basées sur les modèles d'audition humaine. Le signal est transformé vers un domaine adéquat de telle manière qu'on puisse exploiter les effets de masquage psycho-acoustique. Parmi les mesures perceptuelles les plus utilisées nous pouvons citer : PESQ (*Perceptual Evaluation of Speech Quality*) et EMBSD (*Enhansed Modified Bark Spectrum Distorsion*).

L'EMBSD estime la distorsion perceptuel d'un signal en le comparant au signal original dans le domaine des sons forts (*loudness domain*) tout en tenant compte du seuil de masquage de bruit modifié et du modèle cognitif basé sur le post-masquage.

II – Transmission de la voix sur les réseaux IP

Les communications via les réseaux informatiques représentent à l'évidence un phénomène en forte croissance, exponentielle depuis plusieurs années, dans le domaine des nouveaux moyens de communication. Depuis 1995, il est apparu que la transmission de la voix sur Internet pouvait se développer à grande échelle [24]. En effet, avec l'augmentation continue de la vitesse des microprocesseurs et le développement des techniques de traitement du signal, il est devenu réaliste de faire transiter de la voix, au même titre que des données informatiques, sur le réseau Internet, Un réseau informatique n'est pas a priori le support idéal pour assurer le transport de la voix en temps réel, cependant, le développement du web et son faible coût d'utilisation engagent d'ores et déjà de nombreux acteurs sur des applications de téléphonie IP (*Internet Protocol*), qui dispose d'un potentiel important de nouvelles fonctionnalités. La téléphonie sur Internet apparaît donc comme une des évolutions majeures dans le domaine des télécommunications.

II.1 Généralités

Tout d'abord, il faut préciser que le terme téléphonie sur Internet ou téléphonie IP correspond à la téléphonie utilisant la communication par paquets [24], que le réseau informatique soit Internet ou Intranet.

Plusieurs études prévoient une importante croissance du marché de la VoIP (*Voice over IP*), la principale raison est inhérente à la mise en oeuvre de la transmission de la voix par paquets qui utilise mieux les liaisons de télécommunications que la technique de commutation de circuits qui dédie un circuit de bout en bout à chaque communication téléphonique sans tenir compte des temps morts de la conversation. De plus, en téléphonie sur IP, une compression de l'information numérique, qui fait passer la voix numérisée d'un débit standard de 64 kbits/s à un débit de moins de 10 kbits/s en général, est pratiquée pour réduire l'occupation spectrale du réseau. L'idée est d'unifier le transport des informations, voix et/ou données, autour du protocole IP. Etant donné l'engouement actuel pour Internet et les investissements des opérateurs et des fournisseurs d'accès, la téléphonie IP apparaît bien comme une alternative aux réseaux téléphoniques classiques [24].

En téléphonie classique, la transmission pose peu de problèmes grâce à la commutation de circuit. Un canal est réservé aux deux locuteurs et la qualité du signal est constante. Le signal analogique est numérisé par un codage PCM (Pulse Code Modulation), 8000 échantillons par seconde quantifiés sur 8 bits, conforme à la recommandation G.711 de l'Union Internationale des Télécommunications (UIT). Par conséquent, le signal en téléphonie sur réseau commuté correspond à un débit de 64 kbits/s. Les délais de codage et de transit nécessaires à ces opérations étant peu perceptibles par les utilisateurs, les conversations demeurent fluides et sans interruption. Sur Internet, rien de tel, la réservation de ressources n'existe pas. C'est un réseau à commutation de paquets dit asynchrone, sur lequel tous les trafics ont la même priorité. En apparence, toutes ces contraintes ne sont pas vraiment compatibles avec la téléphonie. Pour qu'un bon flux soit envisageable, il faut élaborer des algorithmes de compression capables d'adapter l'information au canal de transmission restreint qu'est Internet. Pour une communication en VoIP, le signal vocal doit être compressé à l'aide d'algorithmes beaucoup plus élaborés qu'en téléphonie classique. Ensuite, l'information à transmettre est découpée par une procédure de paquetisation, à raison de 20 à 30 millisecondes de parole par paquet, avant l'envoi sur le réseau IP. Les paquets d'informations, qui circulent sur Internet, empruntent des chemins différents et arrivent fréquemment dans le désordre. Les paquets sont alors stockés dans des mémoires tampons, ou

*buffer*s, pour être re-séquencés et permettre la décompression de l'information et sa transformation en signal sonore [24].

Des constructeurs et des opérateurs commencent maintenant à mettre en place des passerelles entre Internet et le réseau téléphonique classique et offrent une téléphonie sur Internet concurrente du service habituel. Cependant, la téléphonie sur Internet est encore loin de satisfaire aux exigences de qualité de service attendue. Sans doute est-il trop tôt pour savoir si la téléphonie sur IP peut supplanter le téléphone classique, mais notons que plusieurs études prévoient une importante croissance du marché de la VoIP (*Voice over IP*).

II.2 Principe de fonctionnement

Il s'agit de faire transiter de la voix humaine d'un interlocuteur vers un autre à travers le réseau Internet, tout en ayant le souci que le dialogue se passe sans rupture et avec un confort d'écoute très proche de la conversation en vis-à-vis. Pour ce faire, le temps de transport de la voix entre un émetteur et un récepteur doit être inférieur à 150 ms, avec une tolérance allant jusqu'à 400 ms, et que l'opération doit s'effectuer en duplex intégral.

Le principe consiste, à partir d'une numérisation de la voix, à comprimer le signal, à le découper en paquets de données et à les transmettre sur le réseau. A l'arrivée, les paquets transmis sont réassemblés, le signal de données obtenu est décompressé puis converti en signal analogique pour restituer le signal sonore à l'utilisateur. Les paquets d'informations sont acheminés par les noeuds du réseau constitués de routeurs, ils arrivent alors à destination dans un ordre pouvant être différent de celui de l'émission. A la réception, les durées de transmission variables devront être compensées pour reconstituer le signal numérique. C'est le principe même de la (*VoIP*).

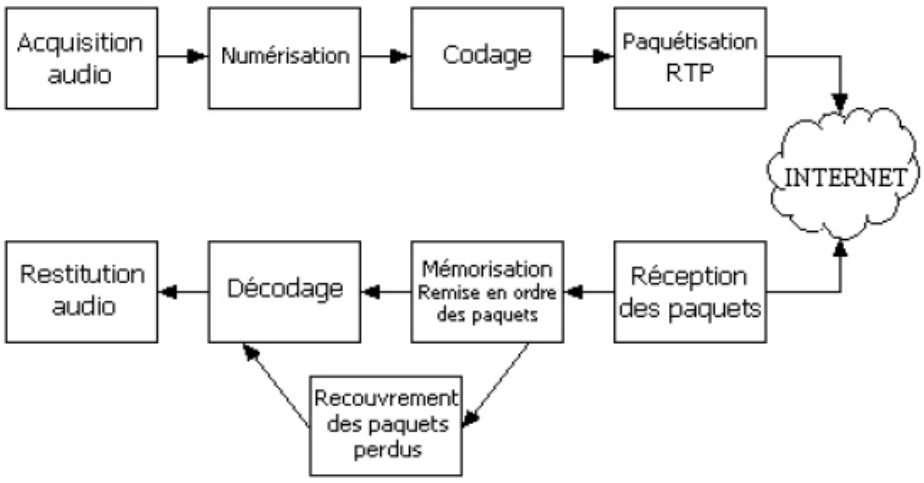


Figure II-1 : Principe de la transmission de la voix par paquets

II.2.1 Architecture des réseaux

Les règles à respecter pour transporter de l'information d'une extrémité à une autre d'un réseau s'appellent protocoles. L'ensemble des protocoles nécessaires pour réaliser une communication constitue une architecture [25].

Il existe plusieurs façons de représenter les protocoles dans une architecture. La plus courante consiste à les classer par couche. Les couches doivent êtres indépendantes les unes des autres, de telle sorte qu'il soit possible de modifier un protocole dans une couche sans avoir à modifier les protocoles dans les autres couches.

Trois grandes architectures coexistent dans le marché mondial actuellement [25]. La première provient de la normalisation de l'ISO (*International Standardisation Organisation*), que l'on appelle OSI (*Open Systems Inteconnection*), ou interconnexion de systèmes ouverts. La deuxième est fournie par l'environnement TCP/IP, utilisé dans le réseau Internet. La troisième a été introduite par l'ITU *(International Telecommunication Union*) pour l'environnement ATM (*Asynchronous Transfert Mode*).

II.2.1.1 Modèle OSI

Le modèle OSI est le premier à avoir été définie, c'est un modèle de référence en 7 couches (figure II-2), qui décrit le fonctionnement d'un réseau à commutations de paquets. Chacune des couches de ce modèle représente une catégorie de problème que l'on rencontre dans un réseau [25].

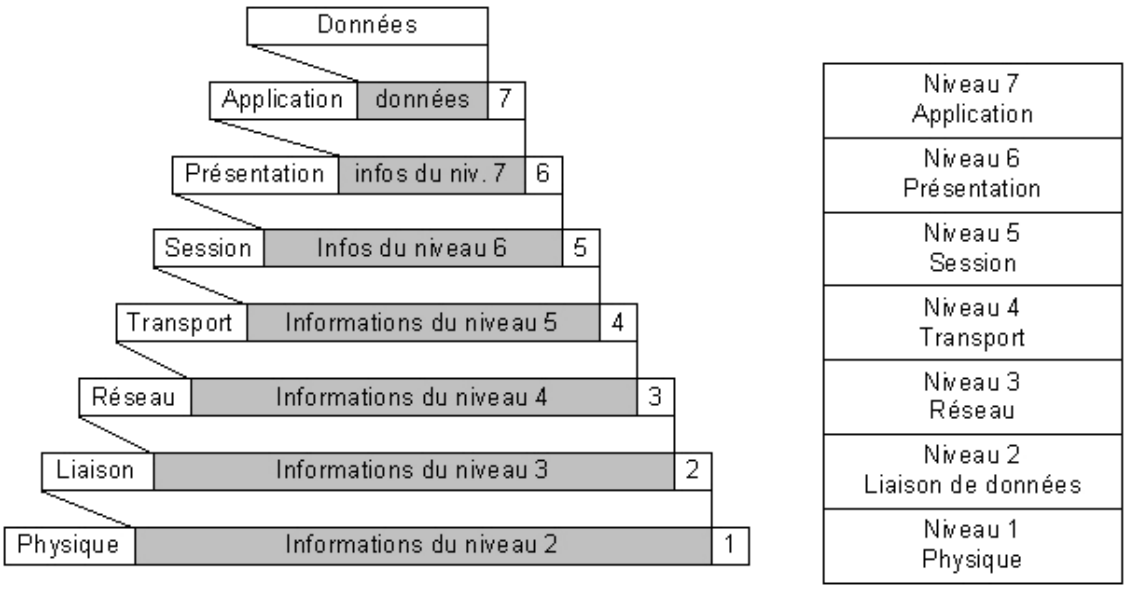


Figure II-2 : Modèle de référence OSI

- La 1ère couche (Niveau Matériel)

Dans cette couche, on s'occupe des problèmes strictement matériels. (Support physique pour le réseau).

- La 2ème couche (Niveau Liaison)

La trame est l'entité transportée sur les lignes physiques. Elle rassemble un certain nombre d'octets, regroupé pour être transporté simultanément. Le rôle de cette couche appelée aussi « niveau trame » consiste à envoyer un ensemble d'éléments binaires sur une ligne physique de telle façon que le récepteur puisse les récupérer correctement, et de reconnaître, lors de l'arrivée des éléments binaires, les débuts et fins de trames.

- La 3ème couche (Niveau Réseau)

Appelée aussi « niveau paquet », le rôle de cette couche est d'acheminer correctement les paquets d'information jusqu'à l'utilisateur final, en passant par les nœuds de transfert intermédiaire ou par des passerelles qui interconnectent deux ou plusieurs réseaux.

- La 4ème couche (Niveau Transport)

La couche transport doit permettre à la machine source de communiquer directement avec la machine destinatrice. On parle de communication de bout en bout (*end to end).* L'information à ce niveau est un « message ».

- La 5ème couche (Niveau Session)

Cette couche a pour rôle de fournir les services nécessaires à l'établissement d'une connexion entre deux machines distantes, de son maintien et de sa libération. Elle procède à l'organisation et la synchronisation de leur dialogue.

- La 6ème couche (Niveau Présentation)

A ce niveau on se préoccupe de la manière dont les données sont échangées entre les applications. En effet la couche présentation se charge de la syntaxe des informations que les applications se communiquent.

- La 7ème couche (Niveau Application)

Dans cette couche on trouve les applications qui communiquent ensemble (Courrier électronique, transfert de fichiers,...).

Découper les problèmes en couche présente des avantages :

- Lorsqu'on met en place un réseau, il suffit de trouver une solution pour chacune des couches.
- L'utilisation de couches permet également de changer de solution technique pour une couche sans pour autant être obligé de tout repenser.
- Chaque couche garantit à la couche qui lui est supérieur, que le travail qui lui a été confié a été réalisé sans erreur.

II.2.1.2 Modèle TCP/IP

Dans les années 70, la défense américaine décide, devant le foisonnement de machines utilisant des protocoles de communication différents et incompatibles, de définir sa propre architecture, qui devra transmettre des données entre des réseaux hétérogènes, quelque soit le support d'information utilisé (Hertzien, câble ou fibre optique), dans les conditions de transmission les plus difficiles (même en état de guerre) [26].

Cette architecture, TCP/IP, est à la source du réseau Internet. Elle est aussi adaptée aux réseaux privés, appelés Intranets. Ce modèle est régi par plusieurs protocoles qui se présentent sous la forme d'une architecture en couche illustrée dans la figure II-3.

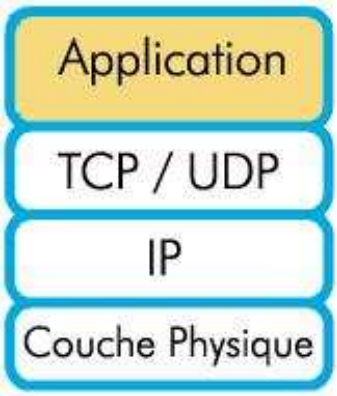


Figure II-3 : Modèle TCP/IP [26]

Les protocoles du modèle TCP/IP sont les suivants :

- **Protocoles RTP et RTCP**

C'est un protocole complémentaire destinés à prendre en compte les contraintes inhérentes à la transmission des paquets en temps réel (voix, images de visioconférence) par rapport aux applications de transfert simple de données. Le protocole RTP, standardisé en 1996, permet ainsi de reconstituer la base de temps des flux, de détecter les pertes de paquets et d'en informer la source dans le cas du mode connecté. Bien qu'autonome, le RTP peut être

complété par le RTCP. Ce dernier apporte à la source un retour d'informations sur la transmission et sur les éléments destinataires. Par exemple, un rapport peut regrouper des statistiques concernant la transmission telles que le pourcentage de perte, le nombre cumulé de paquets perdus ou la variation de délai de transmission, appelée gigue. Ces deux protocoles sont adaptés pour la transmission de données temps réel. Pour le transport de la voix, ils permettent une transmission correcte sur des réseaux bien ciblés tels que les réseaux ATM qui fournissent une qualité de service adaptée. Des réseaux bien dimensionnés, comme un Intranet, pourront aussi être adéquats. En revanche, les protocoles RTP et RTCP ne permettent pas d'obtenir des transmissions temps réel d'assez bonne qualité pour la VoIP. En effet, RTP ne procure pas de réservation de ressources sur le réseau, de fiabilisation des échanges (pas de retransmission automatique et de régulation automatique du débit) ou de garantie dans le délai de livraison.

- **Protocoles TCP et UDP**

S'intègrent dans une architecture multicouches de protocoles, supportant le fonctionnement de réseaux hétérogènes, et fournissant un service de transmission par paquet des données ou de voix.

Les applications nécessitant une transmission fiabilisée et ordonnée d'un flux de données utiliseront de préférence le protocole TCP. Lorsque deux processus désirent communiquer, leurs TCP respectifs négocient et établissent la connexion. Le protocole met ensuite en forme les données à transmettre afin de les transférer au protocole IP, qui les acheminera vers le TCP distant. Ce protocole doit bien sûr inclure les informations nécessaires à la reconstruction des données originales. Pour cela le TCP effectue un contrôle d'erreurs et traite les cas de données perdues, erronées, dupliquées ou arrivées dans le désordre à l'autre bout de la liaison Internet. Il fournit aussi au destinataire un moyen de contrôler le flux de données envoyé par l'émetteur.

Dans le cas d'une transmission de la voix, le protocole *Transport Control Protocol* (TCP) sera remplacé par le *User Datagram Protocol* (UDP), beaucoup plus simple à gérer étant donné qu'il ne fournit pas de contrôle d'erreurs. En effet, le protocole UDP n'opère pas de contrôle de transmission des données, contrairement au TCP, lors d'une communication établie entre deux machines. Il s'agit du mode non connecté dans lequel la machine émettrice envoie des données sans prévenir la machine réceptrice, qui reçoit les données sans envoyer d'avis de réception à la première. Ce protocole, plus adapté au transport temps réel et donc à la VoIP, ne garantit alors ni la délivrance du message ni son éventuelle réémission. La plus

grande rapidité de restitution de l'information se fera au détriment de la qualité de transmission.

- **Protocole IP**

Sa principale fonction est d'acheminer les paquets IP à travers l'ensemble des réseaux interconnectés, selon l'interprétation d'une adresse identifiant les équipements. Notons qu'une distinction doit être faite entre le nom, l'adresse et le chemin, qui désignent respectivement le terminal recherché, l'endroit où il se trouve et un chemin pour y accéder. Le protocole IP s'occupe essentiellement des adresses, qui sont transportées dans l'entête de chaque datagramme et exploitées par les équipements d'interconnexion pour réaliser le routage. C'est à des protocoles de niveau plus élevé que revient la tâche de lier des noms aux adresses. La tâche qui consiste à transcrire l'adresse en terme de chemin revient, quant à elle, au protocole de bas niveau. Ainsi, le protocole IP va assurer le traitement individuel et la transmission des paquets IP entre deux modules Internet, en fournissant un entête propre à chaque datagramme, considéré alors comme une entité indépendante. Cet entête contiendra l'ensemble des informations nécessaires à leur transfert, et disposera de mécanismes permettant l'adressage des services TCP/UDP quelle que soit leur position dans le réseau.

En plus de l'adressage et de l'acheminement en *best effort* des paquets, la couche IP réalise le multiplexage de protocoles ainsi que la destruction des datagrammes ayant transité trop longtemps sur le réseau. Cependant, le protocole IP ne garantit pas la réussite de l'acheminement, le contrôle d'erreur et de flux, ainsi que le séquencement des données.

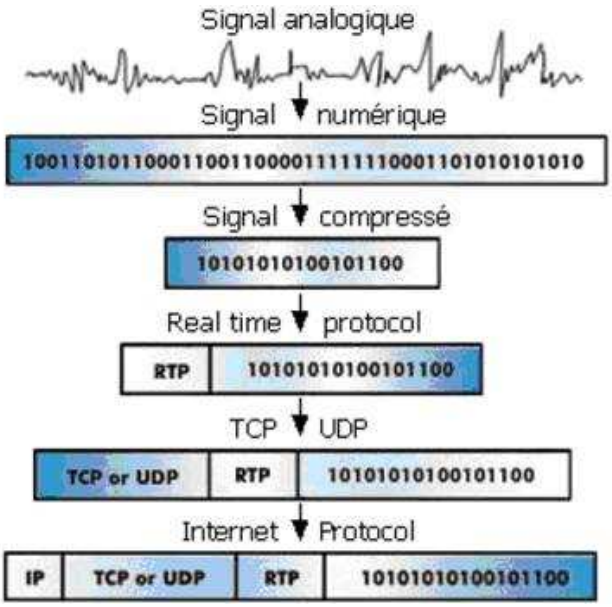


Figure II-4 : Mise en paquet de l'information [24]

II.2.1.3 Modèle UIT-T

Du fait que les réseaux de télécommunication de nouvelle génération doivent prendre en compte les applications multimédia, l'ITU a développé un nouveau modèle de référence adaptant des données différentes (voix, vidéo, texte, Internet,…) au mode de commutation à petites trames et à haut débit qu'utilisent les cellules ATM. Ce modèle n'est en réalité que l'adaptation du modèle OSI aux réseaux de télécommunication [25].

II.2.2 Acheminement de l'information dans les réseaux

Compte tenu de l'architecture d'Internet plusieurs routes peuvent être utilisées pour relier deux réseaux source/destination. Le trafic atteindra alors sa destination après avoir traversé plusieurs réseaux intermédiaires. Des protocoles créent des tables qui définissent le trajet optimal vers le récepteur, grâce à l'analyse des entêtes de paquets IP, en considérant plusieurs facteurs tels que la durée moyenne de transmission, la charge du réseau ou la longueur totale du message. Le routage consiste à transférer les datagrammes, en fonction des informations contenues dans la table de routage [26].

II.3 Qualité de service

Un des problèmes essentiels de la téléphonie sur réseau IP concerne la qualité de service, qui évalue le confort d'écoute. Plusieurs paramètres rentrent en jeux lors de cette évaluation, liés d'une part au retard de transmission et d'une autre à la perceptibilité même de la voix [27].

La notion de qualité de service appliquée aux communications prévoit l'établissement de deux listes soumises au réseau lors d'une demande de connexion. La première liste adresse les paramètres de qualité que l'on souhaite typiquement obtenir et maintenir. La seconde liste fixe les valeurs minimales acceptables pour cette qualité de service, c'est la tolérance. Si les valeurs minimales ne peuvent être fournies par l'un des réseaux traversés ou par l'entité distante, la demande de connexion est refusée.

II.3.1 Les niveaux de qualité

Pour faciliter l'analyse, on peut repérer trois niveaux de qualité (figure II-5) :

- **Niveau Q1**

C'est le seuil limite de compréhension, il demande un effort de concentration, mais la parole est reconnue. Il n'y a pas de notion de duplex et la conversation se déroule en mode alterné. Ce niveau correspond à un taux de perte de paquets compris entre 10% et 25%.

Au-delà, le signal sera trop dégradé, et par conséquent inaudible.

- **Niveau Q2**

C'est le seuil de clarté du signal. Il ne demande pas d'effort particulier pour être écouté confortablement. Certains craquements peuvent encore apparaître, mais sans nuire à la qualité globale. Il ne permet pas encore l'interactivité, car le temps de transmission est trop long et limite la conversation au *half duplex*.

Le taux de perte de paquets reste inférieur à 10%. Le signal sera généralement restauré par le terminal de réception.

- **Niveau Q3**

C'est la qualité « téléphonie » de type Télécom. La conversation est full duplex et le délai de transport reste inférieur à 400 ms. Les pertes de paquets se limitent à quelques pour cent facilement assimilables par le terminal.

Le niveau Q3 qui n'est pas suffisant pour rendre l'Internet comparable au réseau Télécom, sera difficile de le garantir sans faire appel à des procédures de réservation de ressources.

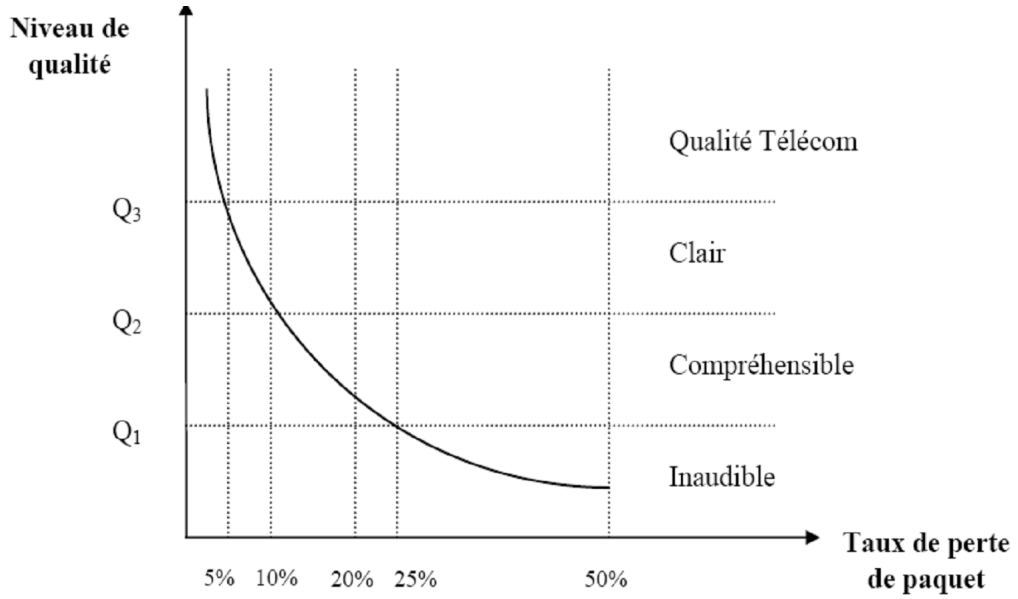


Figure II-5 : Qualité de service en fonction des pertes de paquets

II.3.2 Les facteurs affectant la Qualité de Service

II.3.2.1 Le Retard

Plusieurs paramètres peuvent induire un retard lors d'une transmission VoIP. Pour obtenir le décalage total entre l'émetteur et le récepteur, il faut sommer trois délais : le temps

de traitement par les terminaux d'extrémité, le temps d'acheminement et enfin le temps d'accumulation. En cas de congestion persistante, ces trois temps divergent ensemble.

II.3.2.1.1 Temps de traitement

" Le codage de la voix, dont la vitesse de réalisation dépend des algorithmes utilisés et de la puissance des processeurs qui les exécutent, introduisent actuellement un retard d'environ 25 millisecondes. Cette estimation tient compte du délai de trame et d'encodage qui induit un temps d'attente lié à la longueur de données contenues dans un paquet.

millisecondes selon la configuration d'utilisation.

II.3.2.1.2 Temps d'acheminement

retard, suivant le trafic sur le réseau.

, pour compenser la variation du délai de transmission, ou gigue (figure II-6), entraîne elle aussi un retard qui a pour conséquence d'accentuer encore l'aspect saccadé de la conversation. Le délai de transmission dépend aussi du nombre de routeurs traversés et de la charge de chacun d'eux.

II.3.2.1.3 Temps d'accumulation

La reconstitution de l'ordre des paquets et la compensation des fluctuations des délais de transmission sont estimés à plus de 50 millisecondes.

du signal introduisent un délai moyen de 25 millisecondes.

II.3.2.2 La Gigue

Dans la transmission par paquets, deux paquets émis par la même source à la même destination peuvent emprunter des chemins différents. Ceci est dû au fait que les paquets sont routés indépendamment sur le réseau IP. Deux paquets entre la même source et la même destination peuvent rencontrer différents traitements de retard et de congestion sur le réseau produisant ainsi une variation dans le retard complet rencontré par les paquets. Cette variation est appelée la gigue (*Jitter*). Pour prendre soin du retard gigue, un *buffer* est utilisé à la destination pour stoker les paquets reçus. Lorsque le *buffer* est plein, les paquets seront retardés en séquence avec un retard constant. Ainsi pour restituer un flux synchrone à l'arrivée, des *buffers* de compensation de gigue sont nécessaires mais ce stockage allonge le temps d'attente sur la communication.

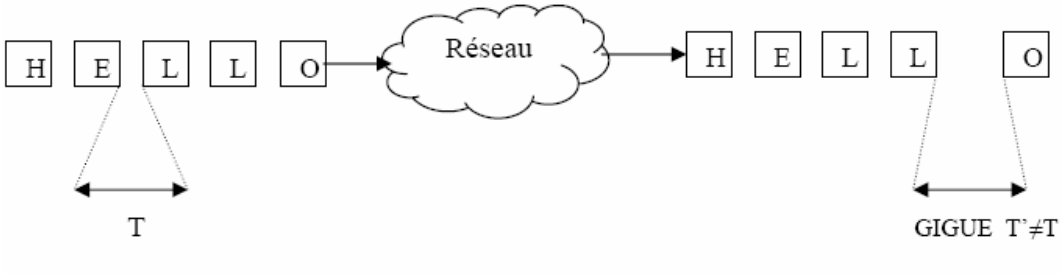


Figure II-6 : Schéma illustrant la GIGUE

II.3.2.3 La bande passante

Sans compression, la voix nécessite 64 Kbps de bande passante, avec compression on peut descendre jusqu'à 5 Kbps. Dans ce dernier cas la qualité du son est moins bonne et le temps de traitement pour la compression et la décompression au départ et à l'arrivée augmente ainsi le temps de latence.

II.3.2.4 Les Pertes de Paquets

Constitue un élément décisif dans la QoS. En effet la voix supporte bien les pertes de paquets par rapport à d'autres applications, néomoins le taux de pertes ne doit pas dépasser 20 %, pour que la voix garde un minimum d'intelligibilité. La retransmission des paquets erronés ou perdus est inutile car elle induirait un temps de latence trop important. Dans un réseau IP, lorsque le débit offert à une liaison excède durablement le débit maximal de cette liaison, la mémoire tampon correspondante élimine un certain nombre de paquets. En revanche, il existe dans les terminaux un mécanisme d'ajustement de la fenêtre aux pertes de paquets constatées.

Ce mécanisme ralentit l'émission de paquets en cas d'encombrement du réseau. Il s'agit là d'un mécanisme d'auto-limitation mis en jeu par les terminaux ; d'où la nécessité d'employé des techniques de masquage des paquets perdus. Dans l'état actuel d'Internet, la qualité de service n'est pas vraiment garantie et reste encore très inférieure à celle constatée sur les réseaux traditionnels de télécommunications [24].

Il existe, néanmoins, des niveaux d'intervention pour gommer partiellement quelques défauts. Tout d'abord, la vitesse sur le réseau Internet peut être augmentée en jouant sur la priorité des flux ou sur le débit, Par ailleurs, on peut agir directement sur le CoDec, en intégrant un algorithme de masquage des trames perdues, voir même dispersion de cette erreur dans le temps [24].

II.3.3 Les Techniques de Masquage des Paquets perdus

Vu l'impacte très néfaste qu'ont les pertes de paquets sur la qualité des transmissions des flux sonores, Des algorithmes de masquage des pertes PLC (*Packet Loss Concealement*) sont utilisés au niveau de l'émetteur ou du récepteur afin de combler les pertes de paquets. Ces techniques peuvent être divisées en deux classes basées respectivement sur l'émetteur (*sender-based*) et le récepteur (*receiver-based*), comme indiqué sur la figure II-7 [28][29].

II.3.3.1 Masquage Basé sur l'Emetteur

La figure II-7 rassemble les techniques de réparation basées au niveau de l'émetteur. Pour éviter les malentendus dans ce qui suit, nous avons distingué une trame de donnée d'un paquet de donnée. Une trame représente un intervalle du flux sonore. Un paquet peut contenir une ou plusieurs trames encapsulées afin d'être envoyées sur le réseau.

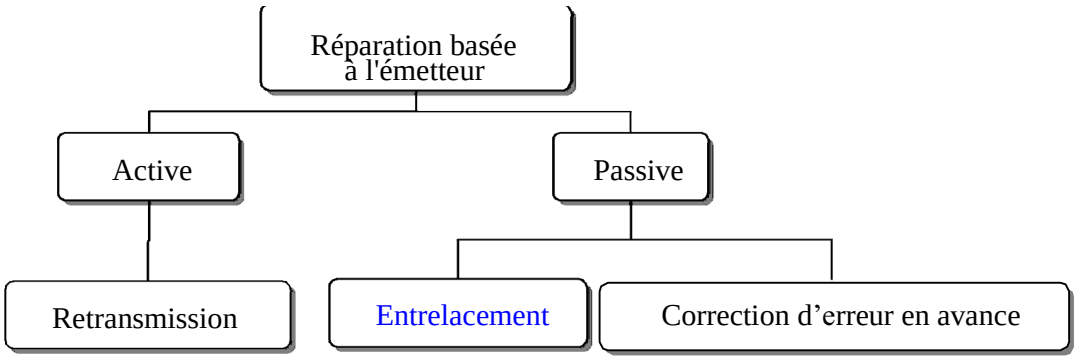


Figure II-7 : Classification des techniques de réparations basées à l'émetteur

II.3.3.1.1 Correction d'erreur en avance FEC (*Forward Error Correction*)

Le schéma de recouvrement repose sur l'addition de donnée de réparation au flux sortant de ces données, les paquets manquants peuvent être réparés le cas échéant la figure II-8.

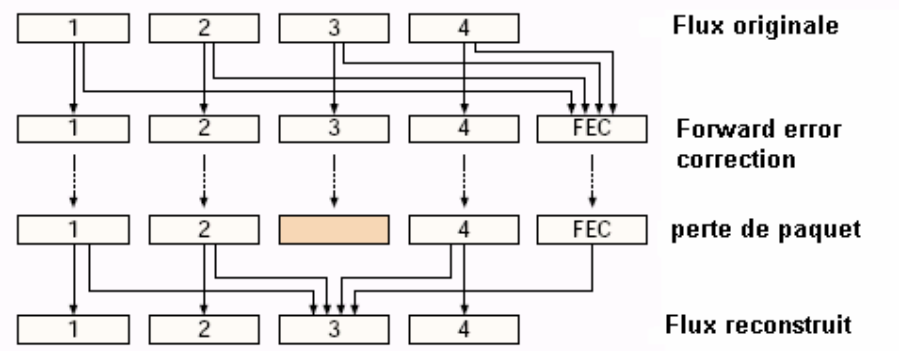


Figure II-8 : Exemple de FEC [30]

Plusieurs avantages découlent de cette méthode, nous pouvons citer la faible demande en ressource de calcul et la simplicité de l'implémentation. En contre partie, cette technique impose un retard supplémentaire, une augmentation de la bande passante et une difficile implémentation au niveau du décodeur [31].

II.3.3.1.2 Entrelacement

La technique d'entrelacement (*interleaving*) est très utile lorsque, les paquets contiennent plusieurs trames et le délai de bout en bout « *end to end* » n'est pas important [32]. Avant transmission du flux, les trames sont réarrangées de telle manière que celles, initialement, adjacentes se retrouvent séparées dans le flux transmis, puis remises dans leur ordre original au niveau du récepteur.

En conséquence, les effets d'effacement de paquets, sont dispersés. La figure II-9 illustre un exemple où chaque paquet contient 4 trames.

L'augmentation de latence constitue un sérieux inconvénient à l'utilisation de l'entrelacement dans des applications interactives. Alors que le maintient d'une bande passant stable avant et après son implémentation représente son avantage majeur.

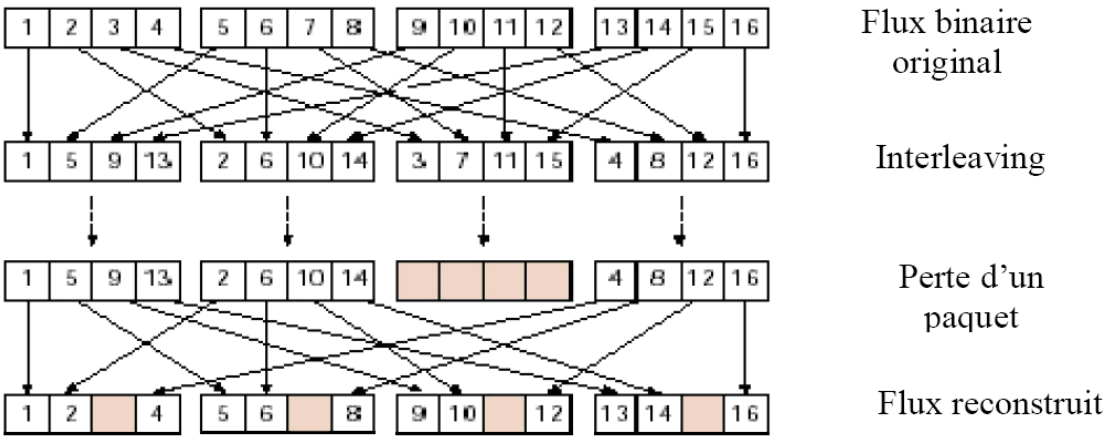


Figure II-9 : Exemple *d'Interleaving* [31]

II.3.3.1.3 Retransmission

Cette technique retransmet, simplement les paquets perdus, elle est difficilement applicable pour les applications interactives et pour lesquelles les délais de bout en bout sont réduits. Cependant pour des conditions de délai plus souple, cette méthode peut être implémentée.

II.3.3.2 Masquage Basé sur le Récepteur

Comme pour la réparation basée à l'émetteur, plusieurs techniques, de masquage d'erreur, initiées par le récepteur d'un flux sonore, ont été réalisées. Ces techniques peuvent travailler soient en tandems avec celles entreprises au niveau de l'émetteur, soient seules.

Le masquage d'erreur repose sur le principe de remplacer les paquets perdus par des paquets similaires aux originaux. Ceci reste possible du fait de la similarité à court terme du flux. La figure II-10 illustre les différentes techniques de masquage d'erreur.

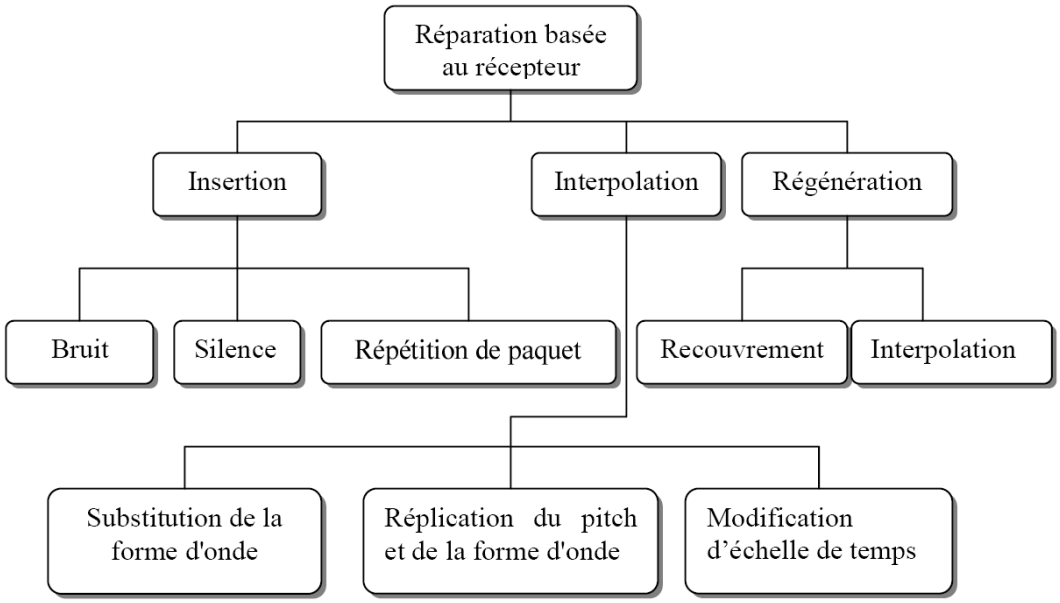


Figure II-10 : Classification des techniques de masquage d'erreur

II.3.3.2.1 L'Insertion

Cette technique de réparation génère un remplacement du paquet perdu en insérant une simple donnée de remplacement. Il est à mentionner que cette technique ne prend pas en compte les caractéristiques du signal, ce qui la rend simple à implémenter. La donnée de remplacement peut être de natures différentes, à savoir un silence, un bruit ou bien une version répétée de la dernière bonne trame reçue. Ces techniques sont faciles à implémenter, mais à l'exception de la technique répétitive, elles possèdent de faibles performances.

Substitution par un silence

Consiste à combler la lacune par un silence afin de maintenir la succession temporelle des paquets. Elle est efficace pour des paquets à longueurs courtes (< 4ms) et de faibles taux de perte (< 2 %). Ses performances se dégradent rapidement lorsque la taille des paquets augmente (la qualité est mauvaise pour des paquets d'une taille de 40 ms). Elle est

couramment utilisée dans les réseaux de communication audio. Toutefois, l'utilisation de ce type de substitution est répandue parce qu'il est simple à implémenter.

Substitution par un bruit

Puisque la substitution de pertes par un silence présente de mauvaises performances, une autre méthode a été introduite. Elle consiste à remplacer la trame perdue par un bruit de fond. En plus, une fois comparée au silence, l'utilisation du bruit blanc a donné une qualité subjective meilleure et une intelligibilité améliorée [32].

Répétition

Avec cette technique, les paquets perdus sont remplacés par la bonne donnée récupérée juste avant la perte.

II.3.3.2.2 L'Interpolation

Cette technique utilise un genre d'identification de paramètre et l'interpolation pour remplacer les paquets perdus. Elle est plus difficile à implémenter et requière plus de ressource de calcul que la méthode d'insertion, cependant, parce que cette technique prend en compte les changements des caractéristiques du signal, ses performances sont meilleures. Plusieurs techniques d'interpolation existent on en site:

Substitution de la forme d'onde

Cette technique met à profit le signal d'avant et, optionnellement, d'après perte pour trouver un signal convenable pour combler les pertes [31].

Réplication du pitch et de la forme d'onde

Cette méthode est une amélioration de la méthode précédente et semble donner de meilleurs résultats, elle utilise, en plus, un algorithme de détection du pitch des deux cotés d'une perte de paquet [31].

Modification d'échelle de temps (*Time scale modification*)

Cette technique permet au signal audio, des deux cotés d'une perte, d'être étiré sur toute la longueur de la perte. Malgré une demande, en ressource de calcul, importante, cette méthode semble travailler mieux que les deux méthodes précédente.

II.3.3.2.3 La Régénération

Les techniques de régénération profitent de la connaissance à priori de l'algorithme de compression des signaux audio pour récupérer les paramètres du CoDec. Par conséquent, le signal audio dans un paquet perdu peut être synthétisé. Ces techniques sont plus performantes en raison de la grande quantité d'informations utilisées dans la réparation.

II.4 Perspectives et enjeux économiques

Bien que le téléphone reste encore le principal mode de communication, le développement d'Internet nous oblige à nous interroger sur les bénéfices à tirer d'un réseau convergent intégrant la voix, la donnée et la vidéo. Selon diverses études, l'avenir de la téléphonie sur Internet semble radieux [33][34]. Pour certains, le seul frein à l'essor de la téléphonie sur Internet serait la qualité des communications. Or, comme celle-ci devrait aller en s'améliorant grâce à l'augmentation conjointe de la bande passante d'Internet, de la vite se de commutation et des algorithmes de compression et de masquage des pertes, la voix sur Internet ne peut que se développer. Une nouvelle approche, basée sur un choix de la qualité de service en fonction du client, est perçue. Le client choisirait alors son rapport prix/qualité en fonction de son type de communication (privée, professionnel, etc...). Cela étant, il est fort probable que l'utilisation par les entreprises de leurs réseaux intranet, pour y passer de la voix en plus des données, se développera de manière plus franche dans les années à venir, les intranets autorisant l'intégration des flux de données et des flux vocaux sur un même support tout en parvenant à administrer la bande passante. La téléphonie sur Internet devient alors qualitativement envisageable. Cette réflexion intéressera particulièrement les entreprises disposant de plusieurs sites répartis sur un périmètre pas trop étendu.

Pour se faire une idée des ordres de grandeur envisageables, certains spécialistes considèrent qu'actuellement le gain pour l'utilisateur en termes de prix devrait s'établir à environ 30 à 40% de réduction, par rapport à l'utilisation du réseau téléphonique classique. On est loin des rapports très importants annoncés quand on compare le prix d'une communication téléphonique longue distance, voire internationale à celui d'une communication locale mais il faut prendre en compte l'amortissement des matériels mis en place si l'on voulait raisonner en termes de coûts réels. Cependant, Internet étant considéré comme un média quasiment gratuit, la téléphonie sur Internet profite largement d'un état de fait qui durera tant qu'elle ne concurrencera pas vraiment la téléphonie classique. Dès lors que les passerelles sur Internet se développeront et que les parts de marché deviendront significatives, le législateur ne manquera pas d'imposer aux nouveaux opérateurs de s'acquitter des redevances d'interconnexion aux réseaux de téléphonie classique, et le coût de la communication téléphonique sur Internet s'en trouvera augmenté d'autant. De plus, force est de constater que le prix des appels internationaux en téléphonie classique diminue sans cesse, notamment grâce à l'utilisation optimale des infrastructures tel que la fibre optique [24].

Les enjeux liés aux systèmes de téléphonie sur IP deviendront de plus en plus importants alors que les offres de services se développent progressivement sur le marché [35]. Plusieurs études prévoient une croissance importante dans les années à venir (figure II-11) par rapport au trafic téléphonique sur le réseau filaire classique. En terme d'offre, de nombreux constructeurs commencent à offrir des systèmes (équipements et/ou logiciels) qui à terme permettront aux internautes d'ajouter une dimension vocale à leurs applications IP, aux entreprises d'interconnecter leurs autocommutateurs et leurs réseaux Intranet, aux fournisseurs d'accès ou de service Internet d'offrir des services de voix et de fax sur IP. En cas de généralisation de la transmission de la voix sur réseau IP, l'équation économique qui repose aujourd'hui sur le caractère marginal de ce type d'utilisation par rapport à la transmission de données serait sensiblement modifiée. Si un véritable basculement des transmissions vers "le tout paquet" est opéré, la voix s'alignerait sur cette évolution et l'IP s'imposerait alors naturellement. Cela conduirait à un redimensionnement des réseaux de transport IP évoluant en rapport avec le trafic.

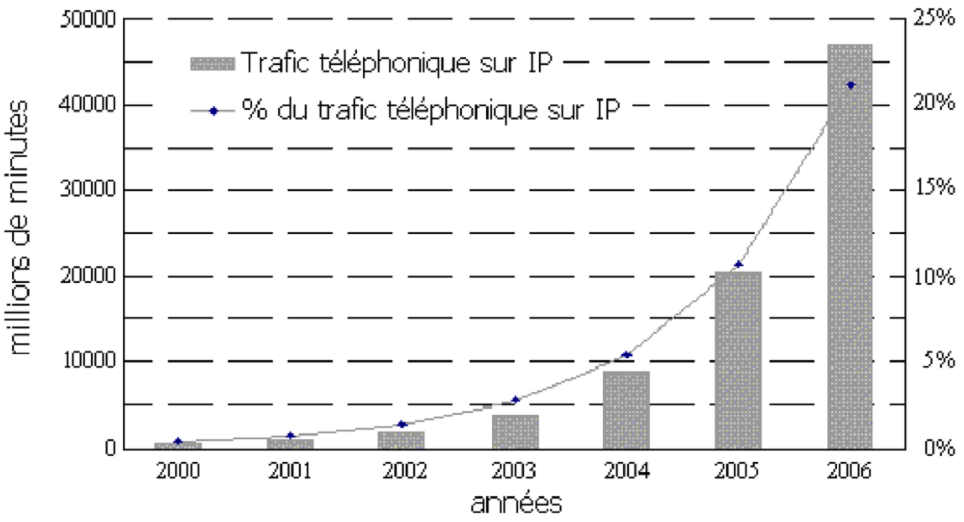


Figure II-11 : Evolution du trafic téléphonique international sur IP [33]

Pour conclure, les technologies évoquées étant dans l'ensemble assez matures pour transporter la Voix sur IP , le réseau et son dimensionnement restent malheureusement encore les points durs de la démarche VoIP. L'avenir de la téléphonie via Internet dépendra essentiellement de deux éléments, l'avenir des réseaux de communication et l'adoption de nouvelles technologies spécialement conçues pour la télécommunication en temps réel sur Internet.

III – Le standard G729

Le but du codage est de diminuer le débit nécessaire à la transmission des informations de synthétisation de la parole. Pour ce faire, des méthodes efficaces ont été proposées pour réduire le nombre de bits nécessaires pour coder le signal d'excitation. Les premiers codeurs prédictifs furent obtenus en quantifiant les échantillons du signal d'erreurs sur deux ou trois bits. Ce type de codage présentant un fort bruit de quantification, le débit n'a pu être réduit au-delà de 16 kbits/s sans perte de qualité importante.

C'est pourquoi la recommandation G.729 de l'ITU [36] décrit un algorithme pour le codage de signaux vocaux à 8 kbit/s au moyen de la prédiction linéaire à excitation par séquences codées à structure algébrique conjuguée CS-ACELP (*conjugate-structure-algebraic-code-excited-linear-prediction*).

Ce codeur est conçu pour fonctionner avec un signal numérique obtenu en effectuant d'abord un filtrage du signal analogique d'entrée dans la bande téléphonique (Recommandation G.712) puis en l'échantillonnant à 8000 Hz et en le convertissant en signal PCM linéaire à mots de 16 bits, qui est injecté dans le codeur. Inversement, on reconvertira le signal de sortie du décodeur en signal analogique [36][37]. Dans cette partie, nous allons présenter les fonctions principales du CoDec décrit par la norme G.729.

III.1 Description général du standard G.729

Le CoDec G.729 opère sur des trames vocales de 10 ms correspondant à 80 échantillons à raison de 8000 échantillons par seconde. Pour chaque trame, le codeur analyse les données d'entrée et extrait les paramètres du codage CELP, qui sont les coefficients du filtre de synthèse et le vecteur d'excitation.

La méthode utilisée, pour déterminer les coefficients du filtre et l'excitation, est appelée « analyse par synthèse : Le codeur cherche les paramètre de codage, en appelant la procédure de décodage dans chaque boucle de la recherche, et en comparant le signal décodé (le signal synthétisé) avec le signal original, les paramètres les plus proches de l'original sont choisis, codés, et puis transmis aux récepteurs, selon l'allocation des bits représentée sur le

Table III.1. Au niveau du récepteur, ces paramètres sont utilisés pour reconstruire le signal de parole original.

Paramètres	**Mot de code**	**Sous-trame1**	**Sous-trame2**	**Total par trame**
Paires de raies spectrales	L0, L1, L2, L3			18
Délai du dictionnaire adaptatif	P1, P2	8	5	13
Parité du délai tonal	P0	1		1
Index de dictionnaire fixe	C1, C2	13	13	26
Signe de dictionnaire fixe	*S*1, *S*2	4	4	8
Gains de dictionnaire (étape 1)	*GA*1, *GA*2	3	3	6
Gains de dictionnaire (étape 2)	*GB*1, *GB*2	4	4	8
Total				80

Tableau III-1 : Affectation des bits dans l'algorithme de codage CS-ACELP à 8 kbit/s

III.1.1 Codeur

Le codeur G.729 de type CS-ACELP, dont le principe est décrit par la suite, est un des plus adaptés au codage numérique de parole et de données multimédia.

Notons qu'avant tout traitement le signal de parole d'entrée, échantillonné et quantifié sur 16 bits, subit dans une procédure dite de prétraitement une normalisation et un filtrage $F(z)$ passe haut dont la fréquence de coupure du filtre est égale à 140 Hz.

$$F(z) = \frac{0.46363718 - 0.92724705z^{-1} + 0.46363718z^{-2}}{1 - 1.9059465z^{-1} + 0.9114024z^{-2}} \qquad \text{III.1}$$

Les réponses impulsionnelle et fréquentielle du filtre $F(z)$ sont données respectivement par la figure III-1 (a) et (b). En sortie de cette opération de prétraitement, le signal noté $s(n)$, sur lequel une fenêtre d'analyse de prédiction linéaire sera appliquée, est utilisé comme entrée de tous les blocs successifs du codeur.

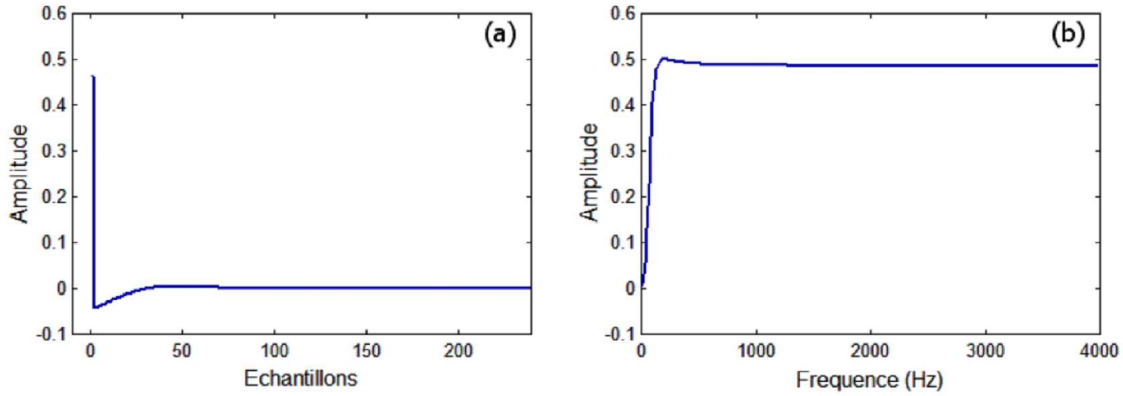


Figure III-1 : Réponse du filtre de prétraitement (a) impulsionnelle (b) fréquentiel

Avant toute analyse à court terme, le signal discret $s(n)$, en sortie du filtre de prétraitement, est segmenté par une fenêtre d'analyse de prédiction linéaire de largeur N_ω = 240 échantillons (fenêtre de Hamming), qui s'applique aux 120 échantillons issus des deux trames vocales passées, aux 80 échantillons issus de la trame vocale courante et aux 40 échantillons de la trame vocale future (figure III-2).

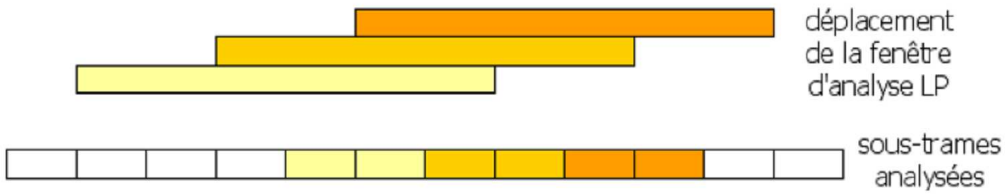


Figure III-2 : Procédure de fenêtrage en analyse (LP)

Le codeur G.729, dont les fonctions principales sont données en figure III-3, opère sur des trames de parole de 10 millisecondes qui correspondent à 80 échantillons numérisés sur 16 bits pour une fréquence d'échantillonnage de 8 kHz. Le signal de parole est analysé à chaque trame pour extraire les coefficients du filtre de Prédiction Linéaire (LP) du dixième ordre, ces derniers sont convertis en lignes de raies spectrales (LSP) et numérisés sur 18 éléments binaires par une quantification vectorielle prédictive à deux étapes.

Le signal d'excitation est choisi au moyen d'une procédure de recherche par analyse et synthèse, dite en boucle ouverte (qui détermine les paramètres en analysant directement le signal d'entrée), dans laquelle l'erreur entre les signaux de parole original et reconstruit est minimisée en fonction d'une mesure de distorsion pondérée. En effet dans le codeur G.729, une pondération perceptive permettra de masquer le bruit du signal et améliorera la qualité de restitution de la voix. Le signal résiduel d'énergie minimale, obtenu en sortie d'un filtre de pondération dont les coefficients sont déduits des paramètres de prédiction linéaire, est alors l'excitation optimale du filtre de synthèse LP. Ce qui permettra d'effectuer une estimation du délai tonal en boucle ouverte basée sur l'autocorrélation du signal vocal pondéré.

Par la suite, les paramètres d'excitation, tels que la période de pitch, les index ainsi que les gains des dictionnaires fixe et adaptatif, sont estimés à partir du signal d'erreur résiduelle de prédiction linéaire sur la base de sous-trames de 40 échantillons, soit 5 millisecondes [1].

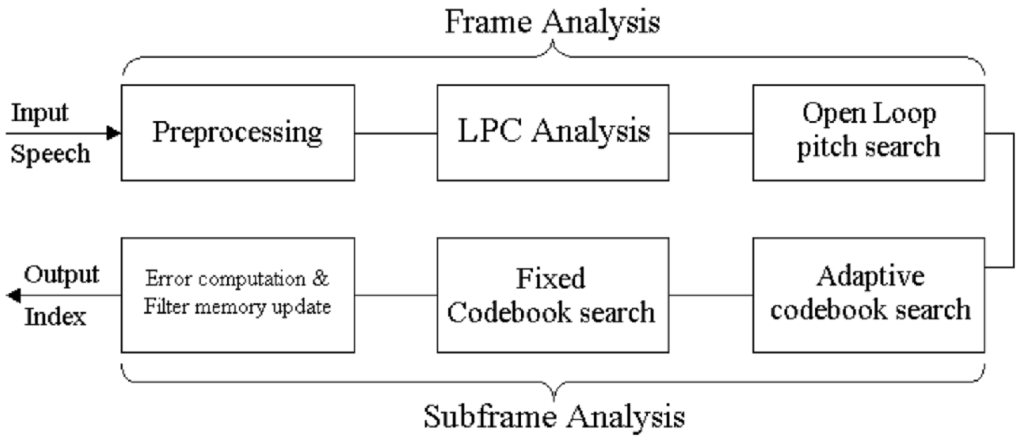


Figure III-3 : Principales procédures du codeur G.729

En fait, le codeur G.729 code le signal d'excitation en recherchant, dans deux dictionnaires ou *codebook*, les formes d'onde qui minimisent l'erreur quadratique entre elles et le signal d'erreur LP pondéré. Les mots de code, contenus dans ces dictionnaires définis par leur structure adaptative et stochastique, étant normalisés, un gain leur est associé de manière à modéliser au mieux les séquences du signal d'excitation.

L'excitation du signal vocal, est calculée pour chaque sous-trame de 5 ms (ce qui correspond à 40 échantillons PCM), et elle a deux composantes : contribution du *codebook* fixe et celle du *codebook* adaptatif. La contribution du *codebook* adaptatif modélise la corrélation à long terme des signaux vocaux, et elle est présentée par le délai tonal de la boucle fermée et un gain [36]. Le délai tonal de la boucle fermée est cherché autour du délai

tonal de la boucle ouverte en minimisant l'erreur quadratique pondéré entre le signal vocale originale et le signal reconstitué. La différence entre l'excitation trouvée, filtrée par le filtre de synthèse, et le signal original, est utilisée pour trouver la contribution du *codebook* fixe. Le vecteur et le gain du *codebook* fixe sont obtenus en minimisant l'erreur quadratique moyenne entre le signal d'entrée pondéré et le signal vocal reconstitué, en utilisant un train d'impulsions comme excitation. Le gain du *codebook* adaptatif et celui du *codebook* fixe sont conjointement quantifiés avec une quantification vectorielle (Ce type de quantification conjointe représente le *CS* « *Conjugate Structure* » dans le nom du CoDec).

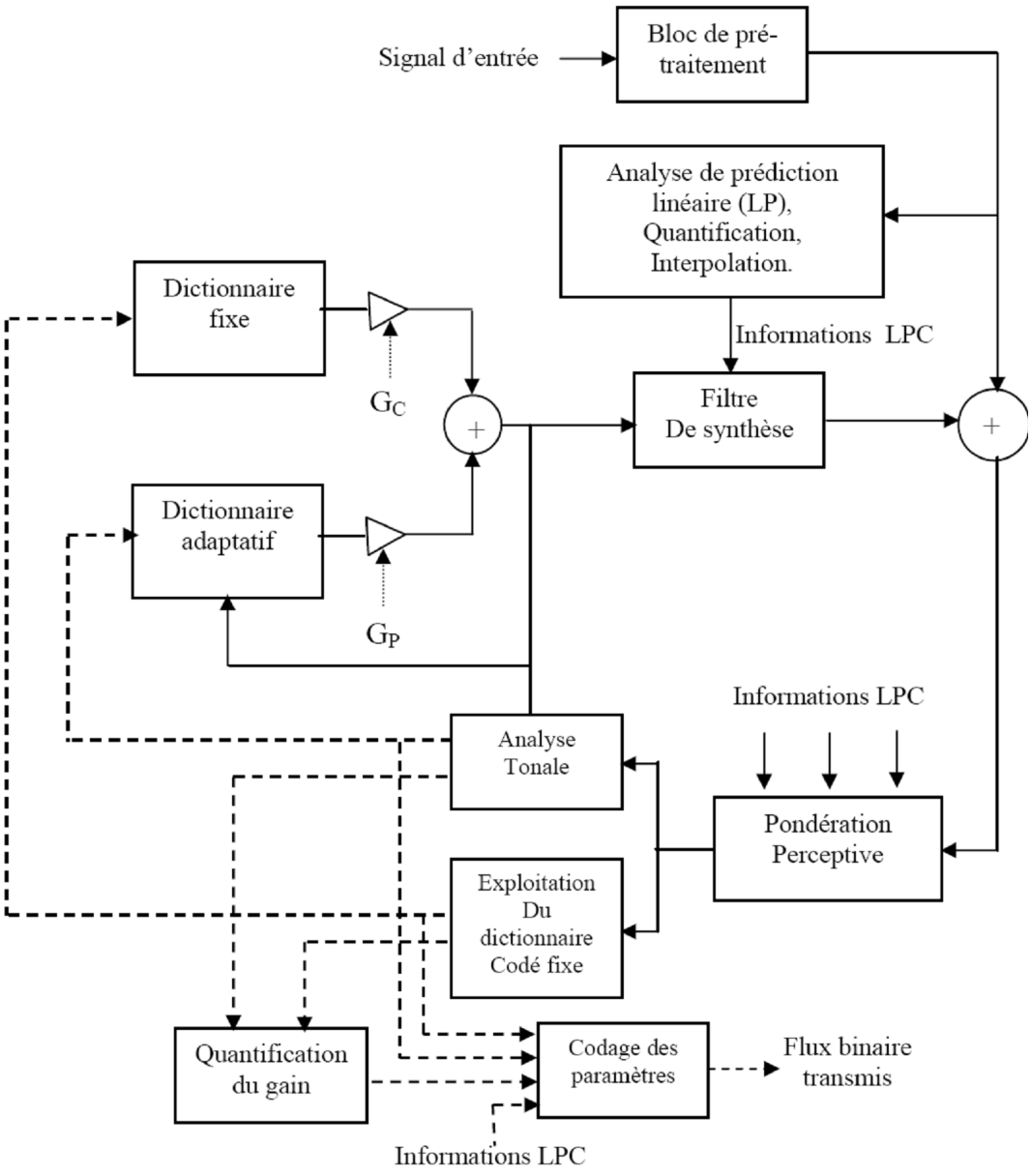


Figure III-4 : Principe du codeur CS-ACELP G.729 [1]

Pour une application temps réel l'utilisation d'un dictionnaire gaussien ou stochastique est pratiquement impossible du fait de la charge de calcul nécessaire pour déterminer les formes d'ondes optimales, d'où la nécessité de l'utilisation d'une recherche dans un *codebook* adaptatif, dont le contenu est mis à jour à l'aide de l'excitation codée passée. Les systèmes utilisant ce type de dictionnaires représentent l'excitation courante par une version antérieure et modulée de celle-ci, à l'aide généralement d'une période de *pitch* fractionnelle. Pendant cette recherche, une gamme de valeurs de chaque paramètre est testée et celle qui fournit la synthétisation la plus précise est choisie. Ce processus qui choisit, à partir du signal résiduel LP, les paramètres d'un dictionnaire de manière à optimiser la qualité du signal reconstruit, est désigné sous le nom de recherche en boucle fermée et fait suite à l'analyse LPC, dite en boucle ouverte, qui détermine les paramètres en analysant directement le signal d'entrée.

III.1.2 Décodeur

Le principe du décodeur est représenté sur la Figure III-5. Les index paramétriques sont d'abord extraits du flux binaire reçu. Ces index sont ensuite décodés pour obtenir les paramètres de codage correspondant à une trame vocale de 10 ms.

Ces paramètres sont les coefficients convertis en paires de raies spectrales (LSP), les deux délais tonals fractionnaires, les 2 vecteurs de dictionnaire fixe et les deux séries de gains par dictionnaire adaptatif et par dictionnaire fixe.

Notons que le décodeur sera relativement plus simple à mettre en œuvre, la plus part de ses fonctions est constituée d'algorithmes déjà rencontrés dans le cadre du codeur mais dans un ordre inverse :

• Les lignes de raies spectrales (*LSP*) sont interpolées et reconverties en coefficients de filtre de prédiction linéaire (LP) pour chaque sous trame de 5 millisecondes.

• L'excitation est construite par combinaison des codes vectoriels adaptatifs et fixes, normalisés par leur gain respectif.

• Le signal vocal est reconstitué par filtrage de l'excitation à travers le filtre de synthèse LP.

• La qualité du signal de parole est ensuite améliorée à l'aide d'un bloc de post-traitement, qui comprend un filtre adaptatif utilisant la sortie des filtres de synthèse à court et à long terme, suivi d'un filtre passe-haut et d'un échantillonneur-normalisateur.

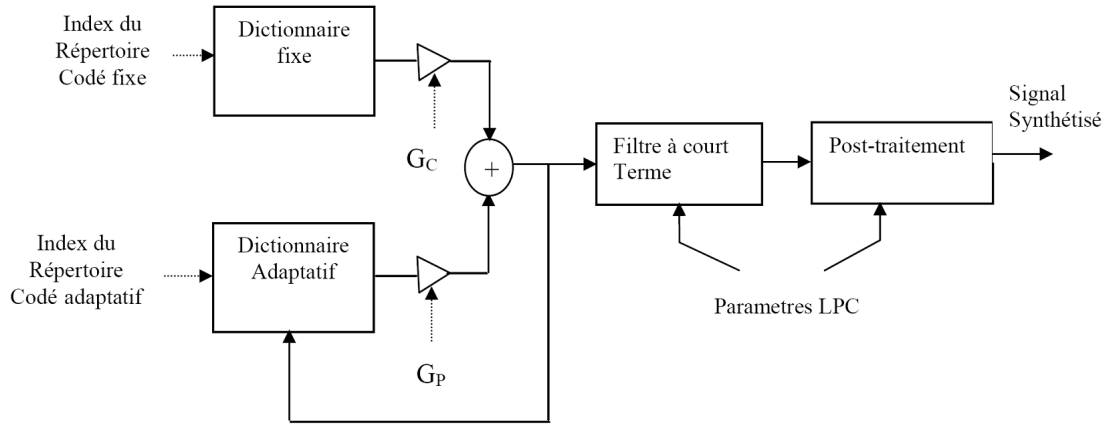


Figure III-5 : Principe du décodeur CS-ACELP G.729 [1]

III.2 Procédure de masquage des trames effacées du G729

Le réseau *IP* ne garantissant pas l'arrivée de tous les paquets envoyés, une procédure de masquage des erreurs a été incorporée dans le décodeur afin de réduire la dégradation dans le signal vocal reconstitué en raison d'effacements de trame dans le flux binaire, pour continuer la restitution en temps réel du signal de parole.

Ce processus de masquage des erreurs est fonctionnel lorsque la trame des paramètres du codeur (correspondant à une trame de 10 ms) a été identifiée comme étant effacée.

La stratégie de masquage consiste à reconstruire la trame actuelle sur la base de l'information déjà reçue. Cette méthode remplace le signal d'excitation manquant par un signal de caractéristiques similaires, tout en diminuant progressivement son énergie. Pour cela, on utilise un classificateur d'éléments voisés utilisant le gain de prédiction à long terme, qui est calculé dans le cadre de l'analyse par post-filtre à long terme. Celui-ci trouve le prédicteur à long terme pour lequel le gain de prédiction est supérieur à 3 dB [36]. Pour cela, on fixe un seuil de 0,5 pour le carré de la corrélation normalisée. Pour le processus de masquage d'erreur, une trame de 10 ms est déclarée « périodique » si au moins une sous-trame de 5 ms possède un gain de prédiction à long terme supérieur à 3 dB, et dans ce cas seul le dictionnaire de code adaptatif est utilisé et la contribution du dictionnaire de code fixe est mise à zéro, Le délai tonal est fondé sur la partie entière du délai tonal contenu dans la trame précédente. Ce délai est répété pour chaque trame successive, Sinon, la trame actuelle est considérée également comme « apériodique » et la contribution du dictionnaire de code adaptatif est mise à zéro, la contribution du dictionnaire de code fixe est construite par sélection aléatoire d'un index de dictionnaire et d'un index de signe.

Les étapes précises à suivre pour masquer une trame effacée sont les suivantes:

• *Répétition* des paramètres du filtre de synthèse (les LSF).

• Affaiblissement des gains du dictionnaire adaptatif et celui du dictionnaire fixe.

• Affaiblissement de l'énergie mémorisée par le prédicteur de gain.

• Production de l'excitation de remplacement.

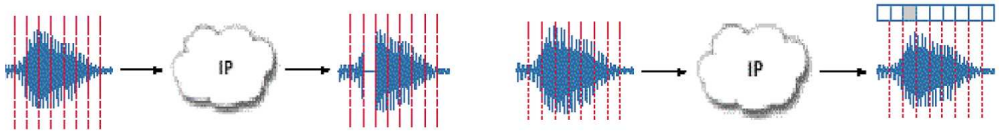


Figure III-6 : Masquage des paquets IP perdus

Le mécanisme de masquage de pertes adopté par ce codeur n'introduit aucun délai supplémentaire, parce que les paramètres de la trame perdue sont récupérés à partir des bonnes trames antérieures reçues, cependant, ce codeur quantifie les paramètres LSF par une méthode prédictive, donc l'utilisation d'un masquage prédictif peut causer une propagation des erreurs aux futures trames.

III.2.1 Répétition des paramètres du filtre de synthèse

Le filtre de synthèse pour une trame effacée utilise les paramètres de prédiction linéaire de la dernière bonne trame. Le registre du prédicteur à moyenne mobile des coefficients LSF contient les valeurs des mots de code. Etant donné que le mot de code n'est pas disponible pour la trame actuelle m, il est calculé à partir des paramètres LSF répétés et du registre de prédicteur précédent.

III.2.2 Affaiblissement des gains du dictionnaire adaptatif et fixe

Le gain de la contribution du dictionnaire fixe est fondé sur une version affaiblie du précédent gain. Il est donné par :

$$g_c^{(m)} = 0.9 g_c^{(m-1)} \qquad \text{III.1}$$

Où m est l'index de la sous-trame.

Le gain de la contribution du dictionnaire adaptatif est fondé sur une version affaiblie du précédent gain, et il est donné par :

$$g_c^{(m)} = 0.9 g_c^{(m-1)} \quad avec\ la\ \lim ite\ g_p^{(m)} < 0.9 \qquad \text{III.2}$$

III.2.3 Affaiblissement de l'énergie mémorisée par le prédicteur de gain

Le prédicteur de gain utilise l'énergie des vecteurs de code du dictionnaire fixe qui ont été précédemment sélectionnés, $c(n)$. Afin d'éviter des effets transitoires dans le décodeur, la mémoire du prédicteur de gain est rafraîchie dès que des trames normales sont reçues, au moyen d'une version affaiblie de l'énergie de la contribution du dictionnaire [36].

III.2.4 Production de l'excitation de remplacement

L'excitation utilisée dépend de la classification de périodicité. Si la dernière trame reconstituée a été classifiée comme étant périodique, la trame actuelle est également considérée comme périodique. Dans ce cas, seul le dictionnaire adaptatif est utilisé et la contribution du dictionnaire fixe est mise à zéro. Le délai tonal est fondé sur la partie entière du délai tonal contenu dans la trame précédente. Ce délai est répété pour chaque trame successive. Afin d'éviter une périodicité excessive, le délai est augmenté de 1 à chaque sous trame successive mais jusqu'à une limite de 143 [36]. Le gain de la contribution adaptative est fondé sur une valeur affaiblie selon l'équation (III.2).

Si la dernière trame reconstituée avait été classifiée comme étant apériodique, la trame actuelle est considérée également comme apériodique et la contribution du dictionnaire adaptatif est mise à zéro. La contribution du dictionnaire fixe est construite par sélection aléatoire d'un index de répertoire et d'un index de signe.

IV – Simulations et résultats

Lorsque des paquets de parole sont envoyés en temps réel à travers des réseaux IP, il n'y a aucune garantie de les recevoir dans une manière appropriée, ce qui est dû à la nature "*best effort*" des réseaux IP. Quand un ou plusieurs paquets sont perdus et aucun effort n'est fait pour les récupérer, la qualité perceptuelle de la parole reçue peut se détériorer considérablement. Plusieurs méthodes peuvent être proposées pour alléger cet effet et sont souvent classées en deux catégories: méthode basée sur le codeur et d'autres sur le décodeur.

La méthode d'entrelacement, basée principalement sur le codeur, fait en sorte que les pertes de paquets soit les plus imperceptibles possibles, en distribuant leur effet sur un intervalle important de temps, tandis que le G729 original où une perte d'un simple paquet, restant intégrale sera facilement perceptible, chose qui nuis à une écoute fluide et continue.

Une comparaison minutieuse, grâce à des outils normalisés, recommandés par les organisations internationales de télécommunication est faite, et les résultats sont présenté en graphes et tableaux, permettant de nuancer facilement les apports introduits au G 729 par la méthode d'entrelacement.

IV.1 Présentation de la méthode d'entrelacement

La méthode d'entrelacement (*Interleaving*) distribue l'effet de la perte de paquets dans le but de réduire sont impact sur la qualité du son transmis. L'information de parole est divisée en paquets de 80 bits, dans le G729 Original. Cette méthode consiste à diviser chaque paquet (trame) en sous trames (de 20, 10 et 5 bits), la première sous trame de chaque paquet est mise consécutivement dans un *buffer*, puis la deuxième et ainsi de suite, pour générer enfin un nouvel ensemble de trames de même taille et nombre, mais avec permutation.

La perte au niveau du réseau IP qui est d'une ou plusieurs trames, engendre au décodage un manque d'un nombre de bits très réduit, au niveau de chaque nouvelle trame reconstituée, ce qui rend cette perte imperceptible à l'écoute, contrairement au G729 original, où l'erreur de tout un paquet de 80 bit rend l'erreur perceptible, et son masquage difficile. Les figure IV-1 IV-2 illustre clairement ce principe.

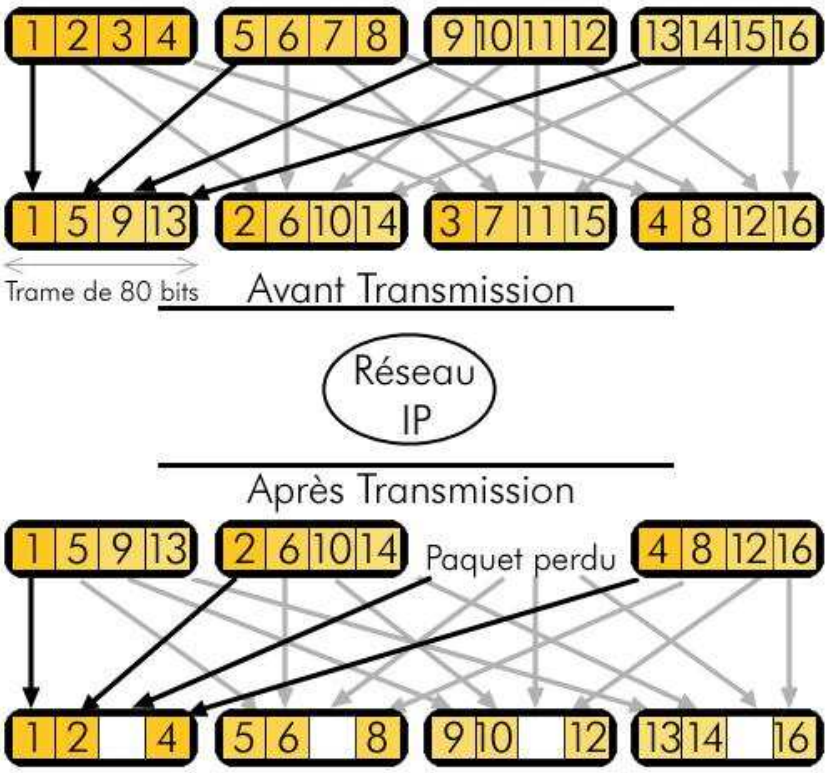


Figure IV-1 : Principe de la méthode d'Entrelacement

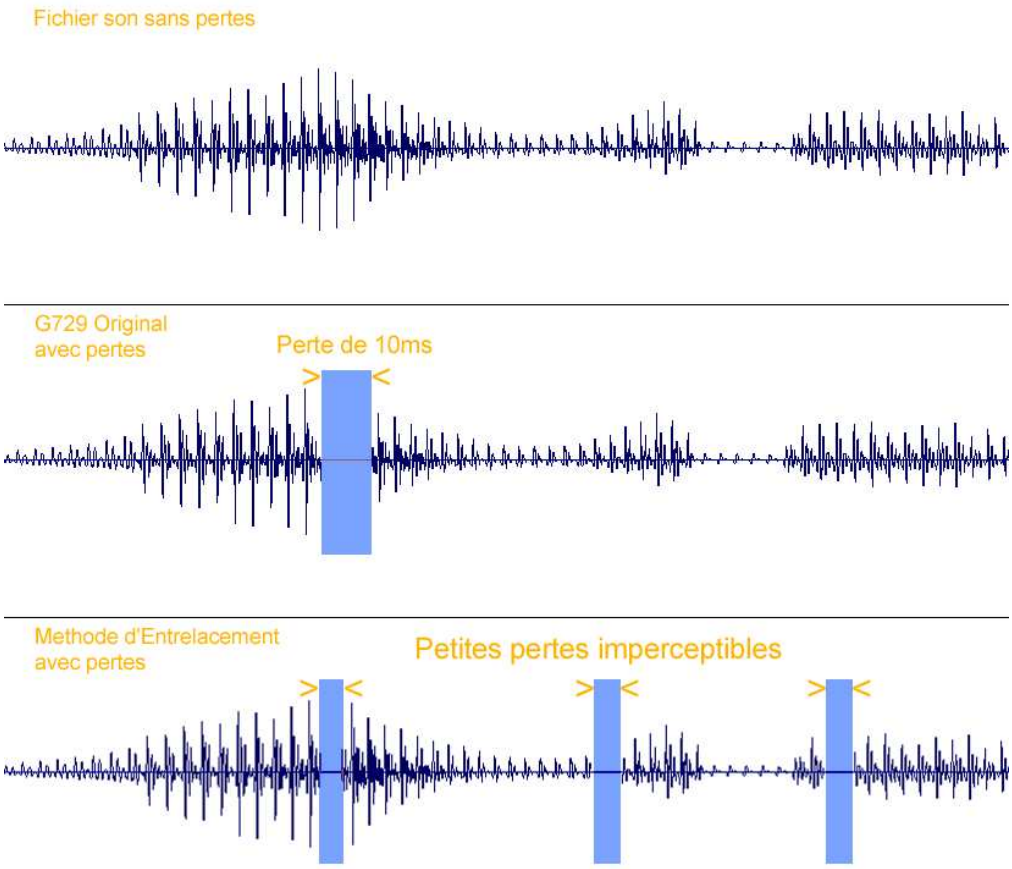


Figure IV-2 : Distribution de l'erreur par la méthode d'entrelacement

IV.1.1 Entrelacement de quatre trames

Le codeur accumule quatre trames successives, ensuite chaque trame sera divisée en 4 sous trames de 20 bits, qui seront ensuite mises chacune dans un *buffer,* le quatrième et dernier *buffer* se remplis au niveau de la quatrième trame reçue et divisée.

Les quatre nouvelles trames constituées sont transmises à la fois au décodeur, qui à son tour réordonne les paquets binaires avant des les transformer en paramètres de paroles.

IV.1.2 Entrelacement de huit trames

Il fonctionne selon le même principe, sauf que le paquet de 80 bits est cette fois divisé en 8 sous trames de 10 bits, et le cumule est de 8 trames successives.

IV.1.3 Entrelacement de seize trames

Le paquet de 80 bits est cette fois divisé en 16 sous trames de 5 bits, et le cumule est de 16 trames successives. Cette méthode est effectivement plus lente est le temps de calcul est plus grand, en espérant que l'erreur sera plus petite.

IV.2 Implémentation, simulation et résultats

IV.2.1 Implémentation

L'entrelacement de trames a été implémenté dans le code source du G729 de l'ITU, en C++ Builder, d'une part dans le codeur sous la forme d'une fonction effectuant l'entrelacement des trames avant d'être transformées en flux binaire, et d'une autre part dans le décodeur comme une autre fonction effectuant le dé-entrelacement du flux binaire, avant l'extraction des paramètres vocaux des trames réordonnées.

IV.2.2 Simulation

Une liaison IP est simulée, en introduisant un fichier parole, de la base de données TIMIT présentée si dessous, de format « .wav » échantillonnée à la fréquence de 8 kHz, au codeur (G729 original et avec entrelacement) qui donne à la sortie un fichier binaire, on fait subir à ce dernier une perte d'un certain nombre de trames, représentant les pertes des paquets IP.

Le fichier binaire ayant subit des pertes, est ensuite introduit au décodeur (G729 original et avec entrelacement) pour en extraire les informations permettant d'avoir en sortie le fichier « .wav » qui représente la parole transmise à l'autre coté de la liaison IP.

IV.2.2.1 Base de données vocales

Une bonne base donnée reste la condition nécessaire pour la validation d'un quelconque résultat trouvé. Dans nos travaux, nous avons utilisé une base de données mondialement reconnue, à savoir, « *The DARPA TIMIT Acoustic-Phonetic Continous Speech Corpus (TIMIT), Training and test data* ». Le corpus TIMIT, formé de paroles lues, a été conçu afin de fournir des données de parole, pour l'acquisition des connaissances acoustique-phonétique et pour le développement et l'évaluation de systèmes automatiques de reconnaissance de parole. Les 630 orateurs, provenant de régions pratiquant les 8 principaux dialectes des Etats-Unis ont participé à l'élaboration du corpus TIMIT. Chacun d'eux lit 10 phrases distinctes, totalisant ainsi, 6300 phrases.

Pour les résultats qui seront présentés par la suite nous avons utilisé deux fichiers parole de la base TIMIT, de même longueur, l'un d'une voix masculine et l'autre d'une voix féminine.

IV.2.2.2 Modèle du réseau IP

Nous avons employé un modèle simple de réseau appelé modèle de **Markov** à deux états pour modéliser le processus point à point de pertes des paquets sur le réseau *IP*. L'état 0 indique que le paquet précédent est reçu et l'état 1 qu'il est perdu.

Soit *p* la probabilité pour que le modèle du réseau abandonne un paquet sachant que le paquet précédent est livré, c'est à dire la probabilité de transiter de l'état 0 à l'état 1. Soit *q* la probabilité pour que le modèle du réseau abandonne un paquet sachant que le paquet précédent est abandonné, c'est à dire la probabilité pour que le modèle reste dans l'état 1. Cette probabilité est également connue comme la probabilité conditionnelle de perte.

Soient *P0* et *P1* les probabilités pour rester dans l'état 0 et l'état 1 respectivement :

$$P_0 = P_0.p + P_1.q$$

$$p_0 = \frac{1-q}{p+1-q} \quad p_1 = \frac{p}{p+1-q}$$

La probabilité pour qu'un paquet soit abandonné sans connaître si le paquet précédent est livré ou abandonné, c'est à dire. La probabilité de perte sans conditions est exactement la probabilité pour que le modèle du réseau soit dans l'état 1 (P1). La figure IV-3 présente le modèle de Markov avec ses probabilités de transition et le tableau IV-1 cite les taux de perte utilisés dans notre simulation.

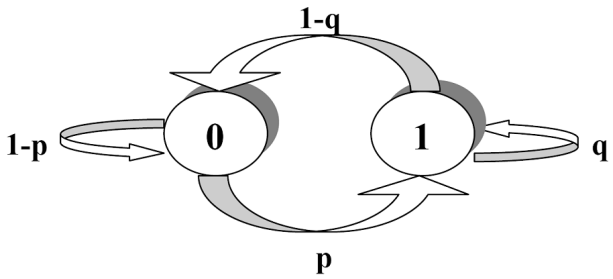


Figure IV-3 : Pertes de paquets modélisées par un processus aléatoire de Markov

Taux (%)	p	q
00	0.00	0.00
10	0.10	0.15
20	0.20	0.30
30	0.30	0.35
40	0.30	0.40

Tableau IV-1 : Les taux de pertes simulés

IV.2.3 Résultats

Une comparaison des performances, du CoDec supposé amélioré avec la méthode d'entrelacement, et du G729 original, est faite grâce à trois méthodes de test, la première calcul la distorsion spectrale entre l'entrée et la sortie, et les deux autres font une comparaison perceptuelle des fichiers son introduis et ceux plus tard décodés.

IV.2.3.1 Distorsion spectrale

Après extraction des LSF du fichier parole introduit, avant le codage, et de celui qui résulte du décodage, un algorithme calcule la distorsion entre les deux fichier LSF, et donne la distorsion moyenne en dB, et génère un fichier texte dans lequel il affiche les valeurs de distorsion spectrale pour chaque trame. A partir de ce fichier texte on peut calculer le pourcentage des distorsions qui dépassent 4 dB, et celle qui sont entre 2 et 4 dB, ce qu'on appelle *Outliers.*

IV.2.3.2 PESQ

Cette méthode représente un standard de l'ITU, pour comparaison objective de la qualité du son transmis de bout en bout d'une liaison à faible débit. Le *Perceptual Evaluation of Speech Quality* donne une valeur numérique entre 0 (aucune similitude) et 4,5 (fichiers son identiques), qui simule la perception humaine de la qualité de parole.

IV.2.3.3 EMBSD

Cet algorithme (*Enhansed Modified Bark Spectrum Distorsion*), estime également la distorsion perceptuelle du fichier son original et celui qui a subit un codage, perte de trames et décodage. L'EMBSD donne une valeur de 0 pour deux fichiers paroles identiques, et une valeur plus grande au fur et à mesure que la distorsion augmente.

IV.3 Présentation des résultats

Les résultats des trois méthodes précédemment expliquées : Distorsion spectrale, PESQ et EMBSD, seront présentés en tableaux de valeurs, avec leur graphes correspondants dessinés avec MATLAB, pour deux échantillons de voix, voix masculine et voix féminine :

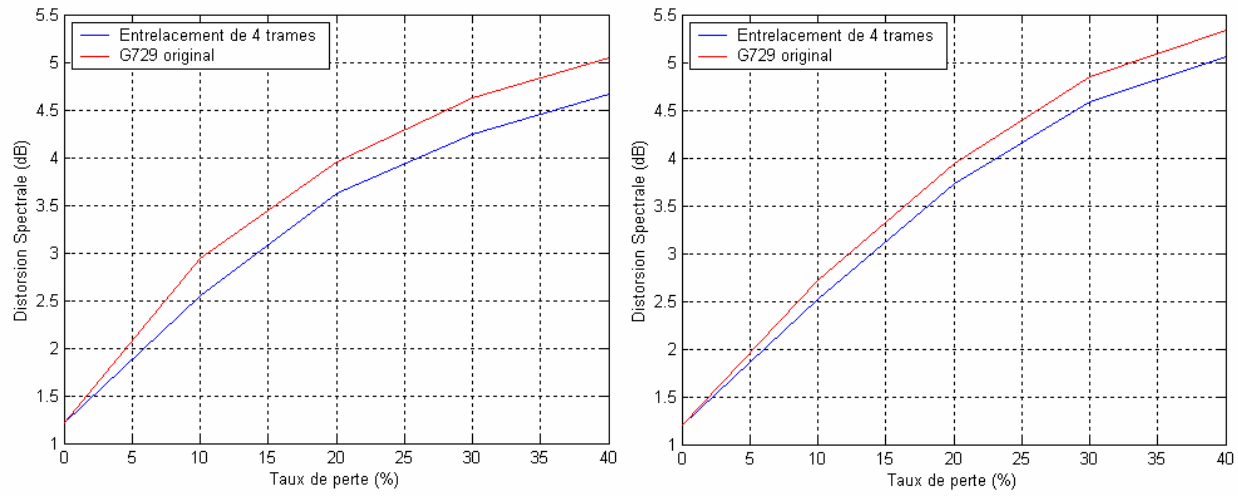
IV.3.1 Distorsion spectrale

Taux de pertes (%)	G729 Original			Entrelacement de 4 trames		
	Dis.Spect Moy (dB)	Outliers (%)		Dis.Spect Moy (dB)	Outliers (%)	
		2-4 dB	>4dB		2-4 dB	>4dB
0	1,22	6,95	0,05	1,22	6,95	0,05
10	2,94	8,51	2,54	2,54	7,14	1,95
20	3,96	35,05	23,85	3,63	28,80	18,94
30	4,63	45,65	26,25	4,25	40,34	26,25
40	5,05	47,80	26,30	4,67	44,11	26,30

Tableau VI-2 : Distorsion spectrale de l'Entrelacement de 4 trames pour une voix masculine

Taux de pertes (%)	G729 Original			Entrelacement de 4 trames		
	Dis.Spect Moy (dB)	Outliers (%)		Dis.Spect Moy (dB)	Outliers (%)	
		2-4 dB	>4dB		2-4 dB	>4dB
0	1,21	6,70	0,20	1,21	6,70	0,20
10	2,72	23,85	9,95	2,52	19,63	8,83
20	3,94	39,40	14,80	3,73	32,27	13,96
30	4,85	45,85	17,85	4,59	41,38	16,48
40	5,34	47,80	21,30	5,06	44,18	19,71

Tableau VI-3 : Distorsion spectrale de l'Entrelacement de 4 trames pour une voix féminine



Voix masculine Voix féminine

Figure IV-4 : Distorsion spectrale de l'Entrelacement de 4 trames

On peut constater une amélioration allant jusqu'à 0,40 dB pour une voix masculine, et 0,28 dB pour une voix féminine, avec plus de distorsion dans la voix féminine.

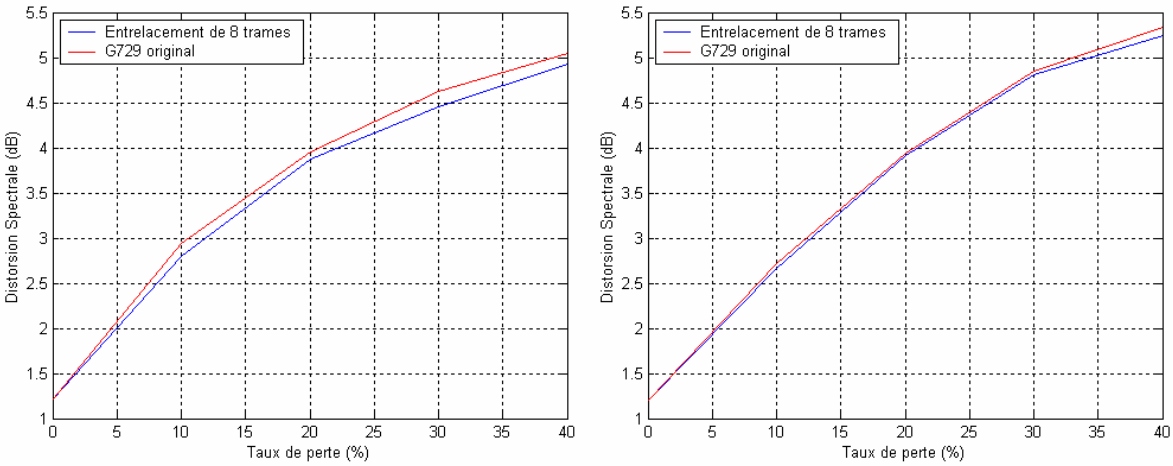
Pour un entrelacement de 8 Trames, les résultats sont les suivants :

Taux de pertes (%)	G729 Original			Entrelacement de 8 trames		
	Dis.Spect Moy (dB)	Outliers (%)		Dis.Spect Moy (dB)	Outliers (%)	
		2-4 dB	>4dB		2-4 dB	>4dB
0	1,22	6,95	0,05	1,22	6,95	0,05
10	2,94	8,51	2,54	2,79	7,16	2,56
20	3,96	35,05	23,85	3,88	7,40	23,10
30	4,63	45,65	26,25	4,45	29,50	24,12
40	5,05	47,80	26,30	4,93	46,54	26,25

Tableau VI-4 : Distorsion spectrale de l'Entrelacement de 8 trames pour une voix masculine

Taux de pertes (%)	G729 Original			Entrelacement de 8 trames		
	Dis.Spect Moy (dB)	Outliers (%)		Dis.Spect Moy (dB)	Outliers (%)	
		2-4 dB	>4dB		2-4 dB	>4dB
0	1,21	6,70	0,20	1,21	6,70	0,20
10	2,72	23,85	9,95	2,67	22,65	9,93
20	3,94	39,40	14,80	3,92	39,15	14,80
30	4,85	45,85	17,85	4,81	44,61	17,85
40	5,34	47,80	21,30	5,25	47,80	20,31

Tableau VI-5 : Distorsion spectrale de l'Entrelacement de 8 trames pour une voix féminine



Voix masculine Voix féminine

Figure IV-5 : Distorsion spectrale de l'Entrelacement de 8 trames

L'amélioration de dépasse pas 0,18 dB pour une voix masculine, et toujours moins pour la voix féminine avec un maximum d'amélioration de 0,9 dB.

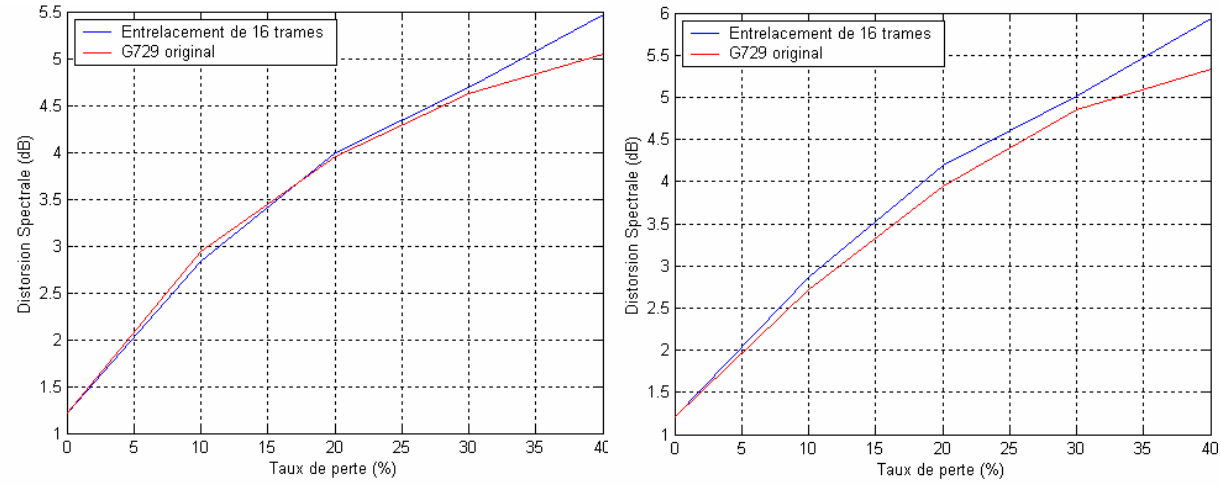
Pour un entrelacement de 16 Trames, les résultats sont les suivants :

Taux de pertes (%)	G729 Original			Entrelacement de 16 trames		
	Dis.Spect Moy (dB)	Outliers (%)		Dis.Spect Moy (dB)	Outliers (%)	
		2-4 dB	>4dB		2-4 dB	>4dB
0	1,22	6,95	0,05	1,22	6,95	0,05
10	2,94	8,51	2,54	2,83	8,51	3,25
20	3,96	35,05	23,85	4,00	38,74	27,58
30	3,63	45,65	26,25	4,69	47,36	29,00
40	5,05	47,80	26,30	5,47	49,78	29,64

Tableau VI-6 : Distorsion spectrale de l'Entrelacement de 16 trames pour une voix masculine

Taux de pertes (%)	G729 Original			Entrelacement de 16 trames		
	Dis.Spect Moy (dB)	Outliers (%)		Dis.Spect Moy (dB)	Outliers (%)	
		2-4 dB	>4dB		2-4 dB	>4dB
0	1,21	6,70	0,20	1,21	6,70	0,20
10	2,72	23,85	9,95	2,86	25,54	11,62
20	3,94	39,40	14,80	4,19	40,20	15,35
30	4,85	45,85	17,85	5,02	46,18	19,21
40	5,34	47,80	21,30	3,93	47, 96	25,16

Tableau VI-7 : Distorsion spectrale de l'Entrelacement de 16 trames pour une voix féminine

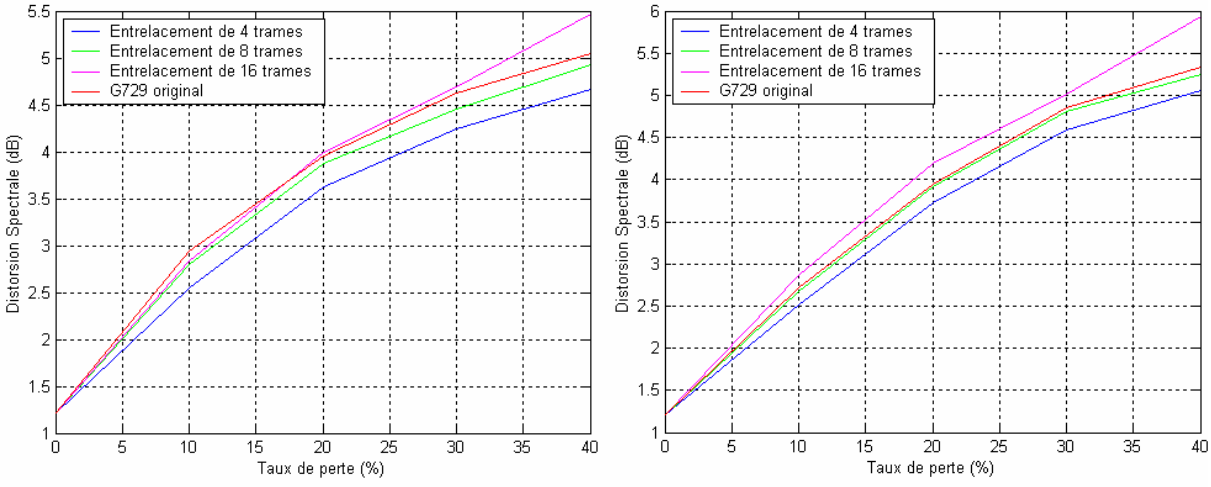


Voix masculine Voix féminine

Figure IV-6 : Distorsion spectrale de la méthode d'Entrelacement de 16 trames

L'entrelacement de 16 trames d'apporte aucune amélioration, mais il introduit plutôt une distorsion supplémentaire par rapport au G729 original, allant jusqu'à 0,42 dB pour une voix masculine. Cette distorsion est plus constatable pour une voix féminine et peut aller jusqu'à 0,59 dB.

En récapitulatif et comparaison des distorsions spectrales des trois méthodes d'entrelacement :



Voix masculine Voix féminine

Figure IV-7 : Distorsion spectrale de la méthode d'Entrelacement de 4, 8 et 16 trames

En comparant les résultats d'entrelacement de 4, 8 et 16 trames, on peut constater que l'entrelacement de 4 trames apporte une nette amélioration par rapport au deux autre méthodes, et c'est la seule méthode entre les trois qui pourra valoir le coût d'une implémentation.

En comparant les résultats donnés par une voix masculine et une autre féminine, on remarque que la voix féminine subit plus de distorsion à cause des pertes de paquets. Par ailleurs l'amélioration que lui apporte la méthode d'entrelacement est moins perceptible que pour une voix masculine.

IV.3.2 PESQ

Les résultats du PESQ de la méthode d'entrelacement à 4, 8 et 16 trames, pour des simulations de pertes de paquets IP de 0,10,20,30,40 % sont représentés dans les tableaux IV- pour une voix masculines et IV- pour une voix féminine :

Taux de Pertes (%)	0	10	20	30	40
G729 Original	3,32	2,06	1,69	1,45	1,28
Entrelacement 4 Trames	3,32	2,12	1,80	1,54	1,39
Entrelacement 8 Trames	3,32	2,08	1,58	1,47	1,27
Entrelacement 16 Trames	3,32	2,01	1,53	1,32	1,19

Tableau VI-8 : Valeurs du PESQ de la méthode d'Entrelacement pour une voix masculine

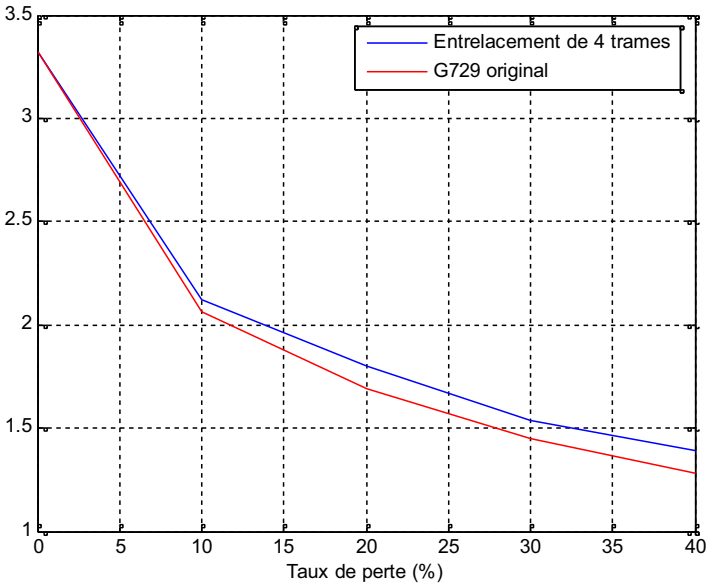


Figure IV-8 : PESQ de l'entrelacement de 4 trames pour une voix masculine

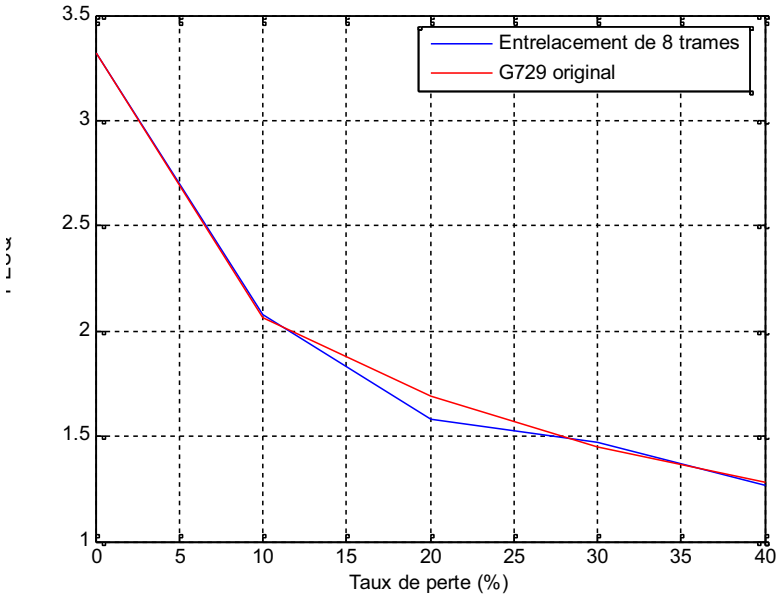


Figure IV-9 : PESQ de l'entrelacement de 8 trames pour une voix masculine

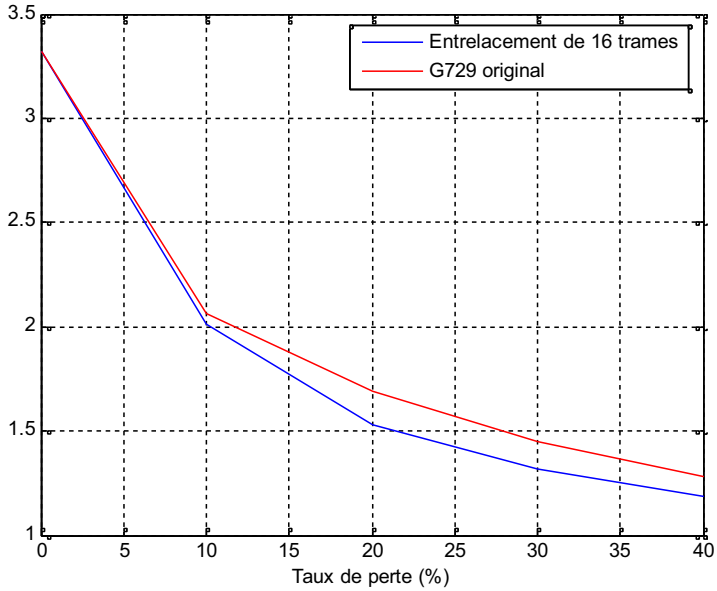


Figure IV-10 : PESQ de l'entrelacement de 16 trames pour une voix masculine

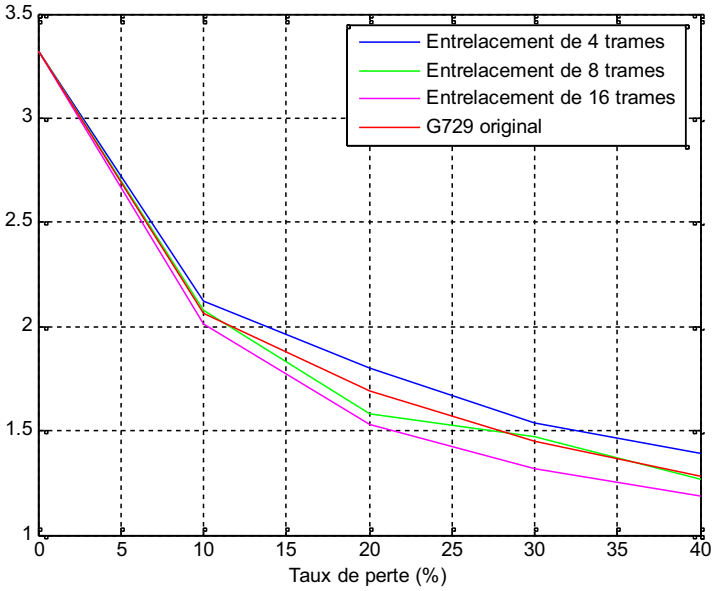


Figure IV-11 : PESQ de l'entrelacement de 4, 8 et 16 trames pour une voix masculine

La méthode d'entrelacement de quatre trames, pour une voix masculine, améliore le PESQ du G729 d'une valeur allant jusqu'à 0,11, tandis que l'amélioration de la méthode de 8 trames est presque inexistante. La méthode de 16 trames diminue plutôt la valeur du PESQ à une valeur au dessous de celle du G729 original.

Les résultats du PESQ pour une voix féminine sont représentés dans le tableau IV ci-dessous :

Taux de Pertes (%)	0	10	20	30	40
G729 Original	3,48	1,83	1,29	1,21	1,13
Entrelacement 4 Trames	3,48	1,98	1,33	1,37	1,19
Entrelacement 8 Trames	3,48	1,87	1,28	1,21	1,02
Entrelacement 16 Trames	3,48	2,01	1,13	1,02	0,93

Tableau VI-9 : Valeurs du PESQ de la méthode d'Entrelacement pour une voix féminine

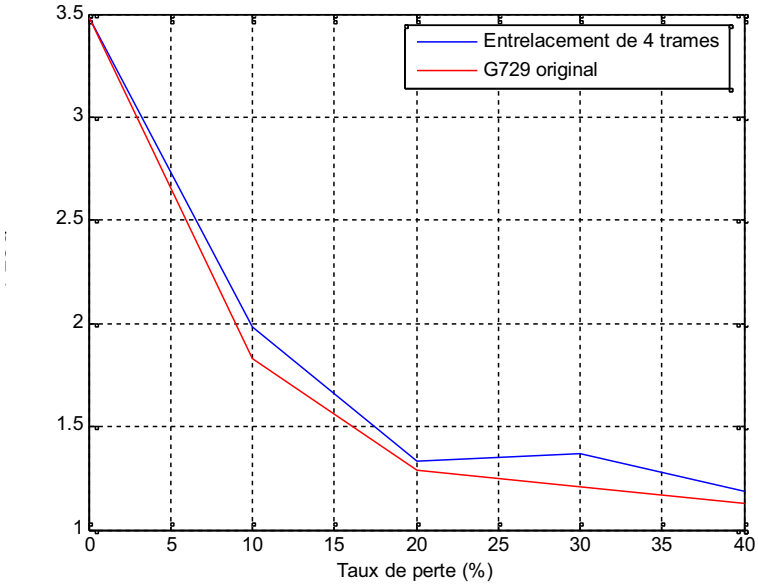


Figure IV-12 : PESQ de l'entrelacement de 4 trames pour une voix féminine

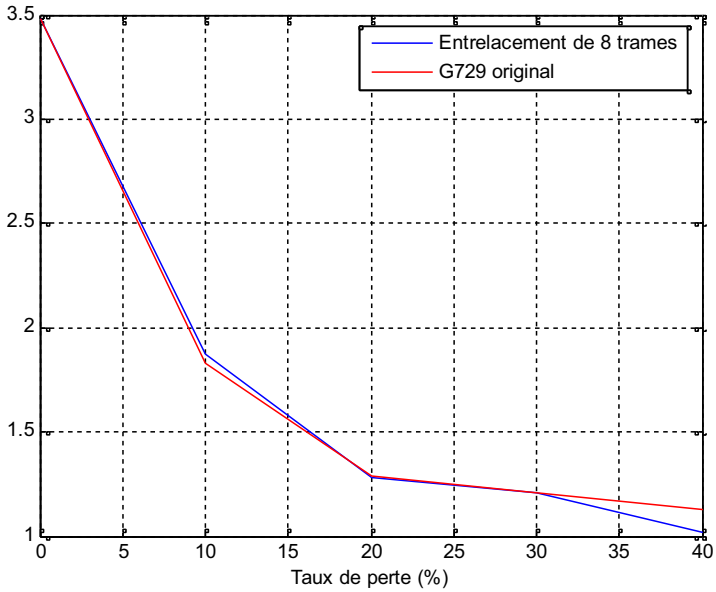


Figure IV-13 : PESQ de l'entrelacement de 8 trames pour une voix féminine

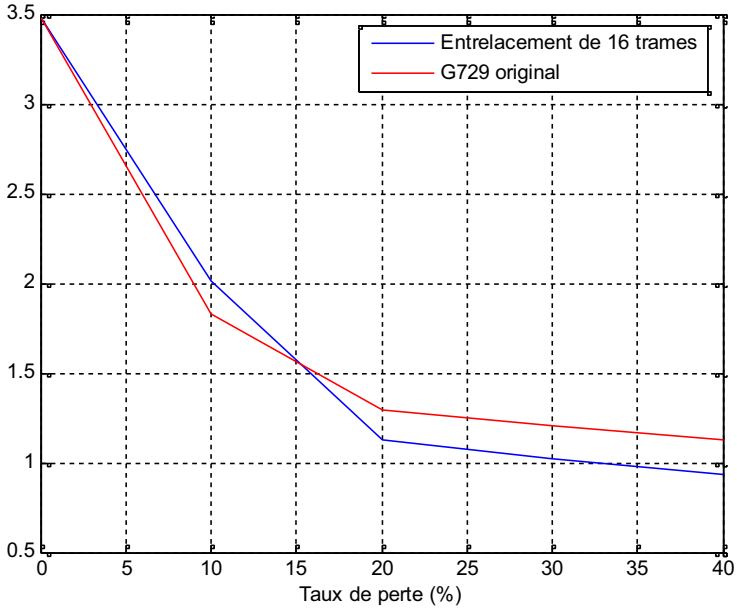


Figure IV-14 : PESQ de l'entrelacement de 16 trames pour une voix féminine

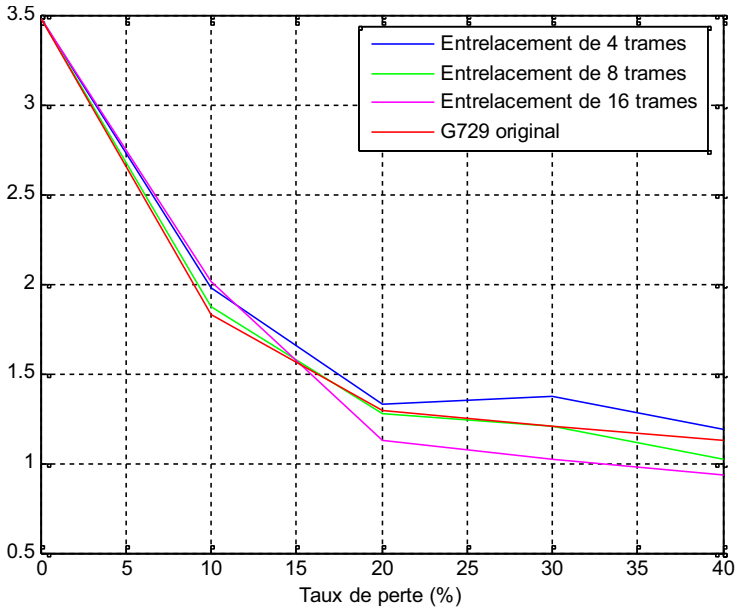


Figure IV-15 : PESQ de l'entrelacement de 4, 8 et 16 trames pour une voix féminine

Pour une voix féminine également, il n'y a que la méthode de quatre trames qui améliore d'une manière constatable les valeurs du PESQ, mais avec plus de distorsion. En effet à partir de 15% de perte le PESQ ne dépasse plus 1,5 ; tandis que pour une voix masculine le PESQ ne descend au dessous de 1,5 qu'à partir de 20 à 25 % de pertes.

IV.3.3 EMBSD

Les résultats de l'EMBSD de la méthode d'entrelacement à 4, 8 et 16 trames, pour des simulations de pertes de paquets IP de 0,10,20,30,40 % sont représentés dans les tableaux IV- pour une voix masculines et IV pour une voix féminine :

Taux de Pertes (%)	0	10	20	30	40
G729 Original	1,37	4.80	8.44	12.18	12.29
Entrelacement 4 Trames	1,37	4.76	7.12	8.12	10.38
Entrelacement 8 Trames	1,37	4.80	7.72	9.50	10.09
Entrelacement 16 Trames	1,37	4.94	8.94	11.66	11.02

Tableau VI-10 : Valeurs de l'EMBSD de la méthode d'Entrelacement pour une voix masculine

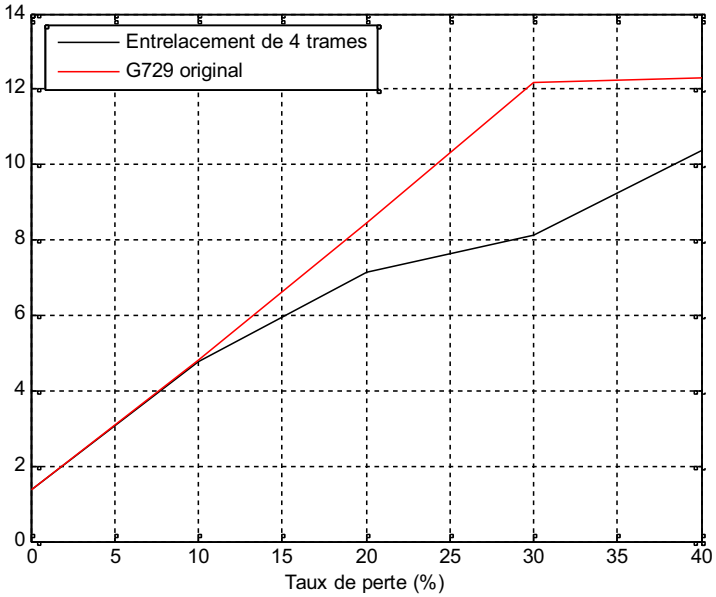


Figure IV-16 : EMBSD de l'entrelacement de 4 trames pour une voix masculine

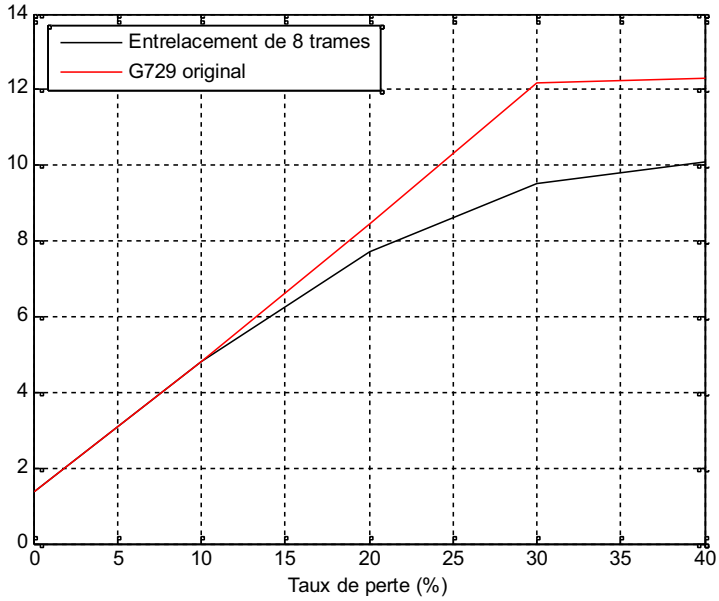


Figure IV-17 : EMBSD de l'entrelacement de 8 trames pour une voix masculine

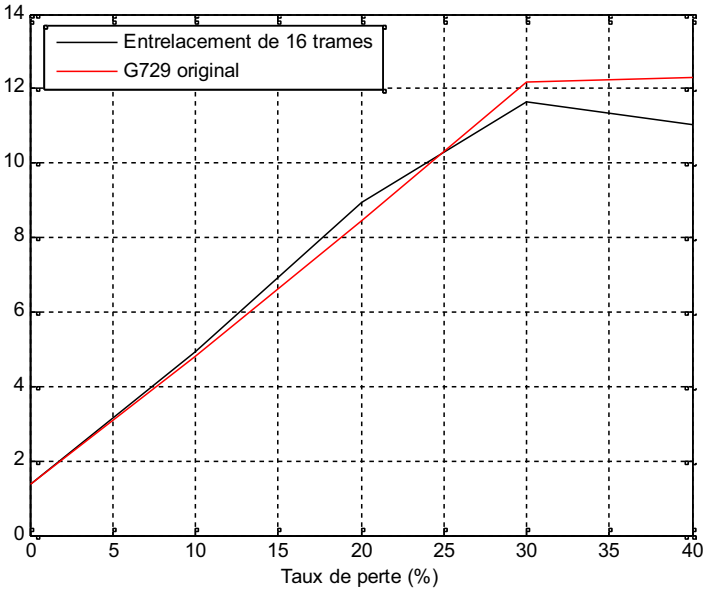


Figure IV-18 : EMBSD de l'entrelacement de 16 trames pour une voix masculine

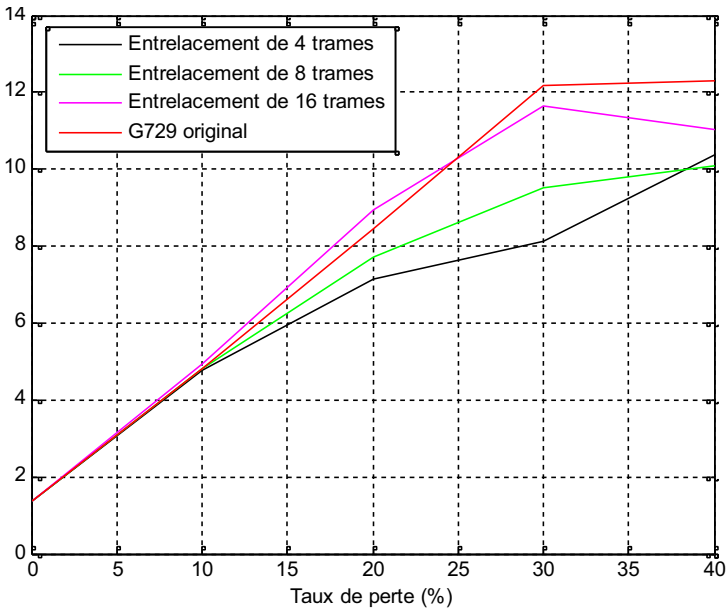


Figure IV-19 : EMBSD de l'entrelacement de 4, 8 et 16 trames pour une voix masculine

Pareil pour l'EMBSD que pour le PESQ, la méthode d'entrelacement de quatre trames, pour une voix masculine, améliore l'EMBSD du G729 d'une valeur allant jusqu'à 4,02, tandis que l'amélioration de la méthode de 8 trames elle ne dépasse pas 2,68. La méthode de 16 trames augmente plutôt la valeur de l'EMBSD à une valeur au dessous de celle du G729 original.

Les résultats de l'EMBSD pour une voix féminine sont représentés dans le tableau IV- :

Taux de Pertes (%)	0	10	20	30	40
G729 Original	0,24	8.11	10.93	12.91	15.16
Entrelacement 4 Trames	0,24	6.75	8.17	9.03	11.20
Entrelacement 8 Trames	0,24	7.03	9.58	10.81	14.59
Entrelacement 16 Trames	0,24	10.72	10.73	13.00	14.87

Tableau VI-11 : Valeurs de l'EMBSD de la méthode d'Entrelacement pour une voix féminine

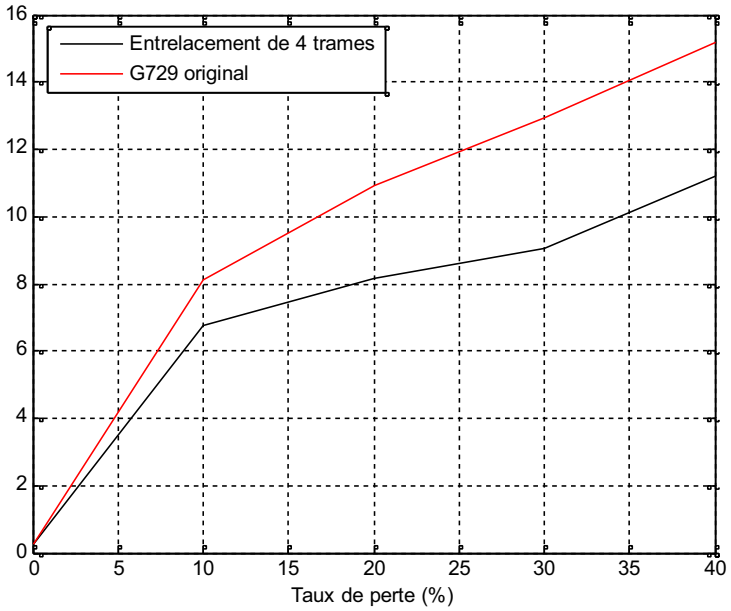


Figure IV-20 : EMBSD de l'entrelacement de 4 trames pour une voix féminine

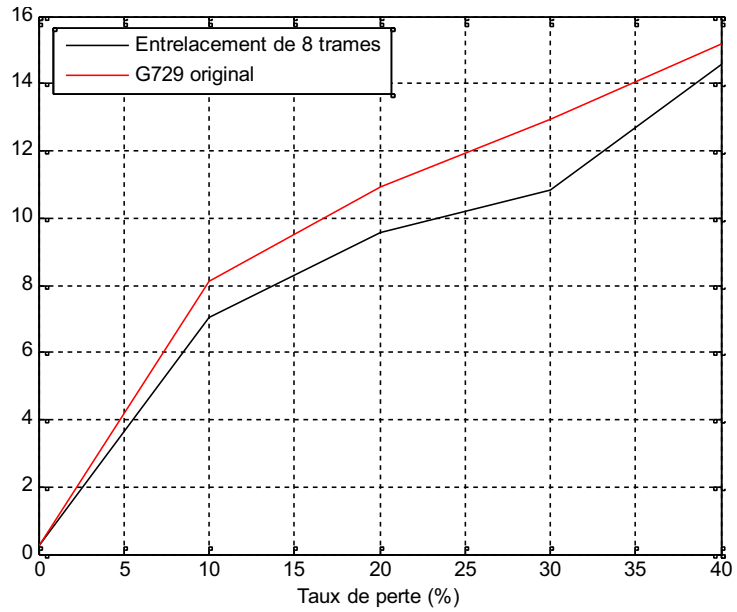


Figure IV-21 : EMBSD de l'entrelacement de 8 trames pour une voix féminine

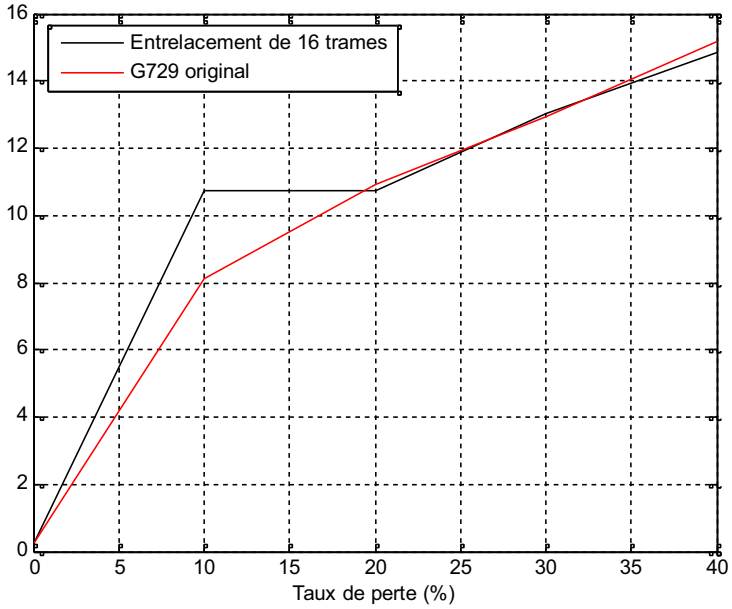


Figure IV-22 : EMBSD de l'entrelacement de 16 trames pour une voix féminine

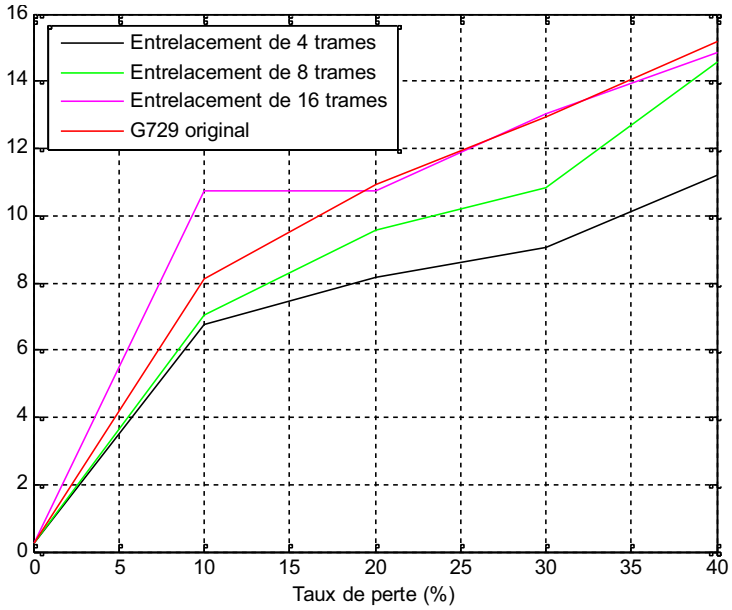


Figure IV-23 : EMBSD de l'entrelacement de 4, 8 et 16 trames pour une voix féminine

Pour une voix féminine également, il n'y a que la méthode de quatre trames qui diminue d'une manière constatable les valeurs de l'EMBSD. Par ailleurs l'EMBSD peut atteindre une valeur de plus de 15, tandis que pour la voix masculine, sa valeur maximale était de 12,29. Chose qui confirme que la voix féminine a moins d'immunité aux distorsions que la voix masculine.

IV.4 Analyse et interprétations

- L'amélioration du CoDec est nettement plus grande, pour l'entrelacement de 4 trames que celui de 8 trames et de 16 trames. Cette dernière a moins de performances que le G729 original, bien que la longueur de la sous trame a diminué et par conséquent celle de la perte, et cela est du à :
  - Le code de correction (masquage) d'erreur incorporé à l'origine dans le G729, ne fonctionne correctement que pour un nombre important de bits perdus, ce qui fait qu'une perte de 10 ou 5 bits au lieu de 80 bits, engendre plus de distorsion à la correction que l'amélioration qu'elle est sensé apporté. Pour une sous trame de 20 bits (méthode des 4 trames) le compromis apporte une considérable amélioration, car la correction d'une erreur de 20 bits est faisable, avec une distorsion peu nuisible.
  - Avec un cumul de 8 ou 16 trames, le signal dépasse le délai de transmission, il ne répond donc pas à la contrainte du temps réel qui est le critère le plus important de la VoIP. Par ailleurs un cumul de 4 trames est à peine dans la limite de stationnarité de 40ms.
  - Dans 1/8 ou 1/16 de trame, il y a très peu d'échantillons pour que le signal soit représentatif, ce qui influe sur sa stabilité.
  - En plus du rendement médiocre, voir même nul, des méthodes d'entrelacement de 8 et 16 trames, elle demandent beaucoup plus de calcul qu'un entrelacement de 4 trames.

- La voix féminine subit plus de distorsion à la perte que la voix masculine, et son amélioration par entrelacement est moins importante, à cause de ses caractéristiques. En effet la voix féminine est généralement caractérisée par une fréquence plus élevée, ce qui rend un petit segment de voix féminine plus informatif qu'en voix masculine, est sa perte induit un plus grand manque de données. Ce qui est illustré par la figure IV-24.

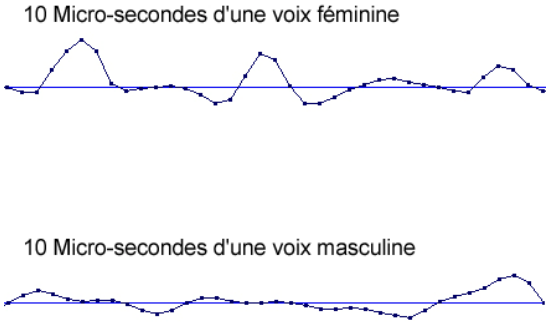


Figure IV-24 : Segment de 10 Micro-secondes de voix féminine et masculine

Conclusion

Dans le réseau IP, il y a un trafic excessif de paquets, des retards et des pertes de paquets qui peuvent détériorer la qualité de la communication.

Bien que la téléphonie *IP* soit très économique le publique hésite encore à l'utiliser en raison des éventuelles dégradations dues aux retards et pertes de paquets lors de la transmission. Ces deux facteurs affectent considérablement la qualité de la parole restituée.

Même si la perte de paquets est irrémédiable, son effet peut toutefois être réduit au minimum : il existe quelques solutions palliant à la dégradation due à ces pertes de données pour poursuivre la restitution du signal parole en temps réel.

Dans les fonctions du CoDec décrit par la norme G.729, il existe une procédure de masquage des erreurs. Dans le but d'améliorer ses performances, nous avons implémenté une autre technique de dissimulation de trames perdues basée sur l'entrelacement.

Cette technique distribue l'effet des pertes de paquets pour réduire leur impact sur la qualité de la communication tout en l'étendant sur des périodes plus longues : les informations d'une même trame de parole sont distribuées et envoyées dans différents paquets après entrelacement et seront par la suite réordonnées à la réception dans leur forme originale.

Ainsi, la perte d'un paquet de la trame entrelacée à l'émission occasionne de multiples petites pertes dans la trame reconstruite à la réception qui causent une distorsion moins audible que celle de la perte d'un paquet et de ce fait n'affectent pas l'intelligibilité de la parole.

L'avantage de cette méthode réside dans son adéquation au réseau IP car la probabilité de perte d'un seul paquet est plus grande que celle de la perte de plusieurs paquets successifs. De plus, elle permet de garder un débit constant.

Son inconvénient est le délai supplémentaire requis pour le cumul des trames et le temps de traitement qui affecte négativement la restitution du signal en temps réel.

Après implémentation de cette méthode, nous avons évalué ses performances à l'aide de la distorsion spectrale et constaté qu'elle avait renvoyé de bons résultats : elle nous a permis d'améliorer la plage de distorsion (de 0.20 à 0.28 dB pour une voix féminine et de 0.38 à 0.40 dB pour une voix masculine) obtenue par rapport au standard.

Pour mieux valider nos résultats, nous avons utilisé une mesure perceptuelle. Les résultats de celle-ci nous ont montré que l'amélioration est nettement meilleure pour les voix masculines.

Enfin, en ce vaste et passionnant domaine, marqué par des évolutions constantes, d'autres investigations demeurent naturellement indispensables.

Parmi ces perspectives d'avenir, nous citerons tout particulièrement :

- L'implémentation de l'algorithme d'interpolation de l'excitation avec celui de la dissimulation basé sur l'entrelacement.
- La confirmation des résultats obtenus par un test d'écoute.
- L'implémentation de l'algorithme final sur un chip (DSP).

Bibliographie

[1] D. O'Shaughnessey, *"Speech Communication - Human and Machine"* Addison-Wesley Publishing Company, 1987.

[2] K.Bartkova, *"Production, description, perception du signal vocal"*, CNET, Lannion, Janvier 1996.

[3] S. Saoudi, *"Introduction à la compression de la parole"*, Support ENST Bretagne, Novembre 1999.

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

[5] D.G. Rowe, *"Techniques for Harmonic Sinusoidal Coding"*, July 1997.

[6] J. R. Deller Jr., J. H. L. Hansen, and J. G. Proakis, *"Discrete-Time Processing of Speech Signals"*, New York IEEE Press, 2000.

[7] S. Saito, K. Nakata, *"Fundamentals of Speech signal Processing"*, Academic Press, 1995.

[8] M. Xie et D.Berkani. "*Amélioration des performances des codeurs de parole*"Août 1997.

[9] C. Laot, *"CELP Code Excited Linear Prediction"*, Master ENST Bretagne, 1992.

[10] C. Papacostantinou, *"Improved Pitch Modelling for Low Bit-Rate Speech Coders"*, Master Thesis, McGill University, Montreal, Canada, August 1997.

[11] N. Moreau, *"Techniques de compression des signaux"*, Collection technique et scientifique des télécommunications, 1995.

[12] John Makhoul, *"Linear Prediction A Tutorial Review"* Proc. of the IEEE, Vol. 63, No. 4, April 1975.

[13] F.Merazka, "*Techniques de codage de la parole applications aux LSPs et aux systèmes VoIP*", Thèse de Doctorat d'État, Présenté a l'École National Polytechnique Alger 2004.

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

[15] S. Haykin,*"Adaptive Filter Theory"*. Upper Saddle River, New Jersey Prentice Hall, Third ed., 1996.

[16] F. Itakura, *"Line Spectrum Representation of Linear Predictive Coefficients of Speech Signals"* J. Acoust. Society America, vol. 57, p. S35, April 1975.

[17] F.K. Soong, B.-H. Juang, ”*Line Spectrum Pair (LSP) and Speech Data Compression*” Proc. ICASSP’84, pp. 1.10.1-1.10.4, San Diego, California, March 1984.

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

[19] W. Pereira,”*Modifying LPC Parameter Dynamics to Improve Speech Coder Efficiency*”, Master Thesis, McGill University, Montreal, canada, Septembre 2001.

[20] Alexis Pascal Bernard, "*Source-Channel Coding of Speech*", Master of Science in Electrical Engineering University of California Los Angeles,1998.

[21] F.Merazka, "*quantification des paramètres LSF*", Thèse de Magistèr, a l’École National Polytechnique Alger 1997.

[22] S. Wang, A. Sekey, and A. Gersho, "*An objective measure for predicting subjective*".

[23] R. Laroia, N Phambo, and N,Favardin,”*Robust abs=d efficient quantization of speech LSP parameter using structured vector quantizer*”, in Proc.IEEE Int. Conf on acoustics , speech , and Sig.processing(Toronto, Canada) ,may 1991 pp 641-644.

[24] G.Madre, ”*Application de la transformée en nombres entiers à l’étude et au développement d’un codeur de parole pour transmission sur réseaux IP*”,Université de Bretagne Occidentale, octobre 2004.

[25] Pujolle, ”*Les réseaux*”, Eyrolles, 2003.

[27] S. Pracht, D. Hardman,”*Voice quality in converging telephony and IP networks*”, www.ednmad.com, September 2000.

[28] N. Jayant and S.W. Christensen, “*Effect of packet losses in waveform coded speech and improvements due to an odd-even sample-interpolation procedure*”, IEEE Trans.Commun.

[29] W. R. Erhart and J. D. Gibson, “*A speech packet recovery technique using a model based tree search interpolator*”, IEEE workshop of speech coding for telecommunications, Sainte-Adele, Quebec, Canada pp. 13-15, Oct. 1992.

[30] P.Kroon and B.S. Atal, "*Predictive coding of speech using analysis-by-synthesis techniques*", in Advances in *Speech Signal Processing* S. Furui and M.M. Sondhi, Eds New York Markel- Dekker, pp 141-164. *1991*.

[31] C. Perkins, O. Hodson, and V. Hardman,”*A Survey of Packet-Loss Recovery Techniques for Streaming Audio”*, IEEE Network , Volume 12 Issue 5, pp. 40 –48, Sept.-Oct. 1998.

[32] Moo Young Kim and Renat Vafin,”*Packet-Loss Recovery Techniques For VoIP”*, Dept. of Speech, Music, and Hearing Royal Institute of Technology (KTH).

[33] J.C. Merlin, P. Fritz et C. Malet, *"La téléphonie sur Internet*", Rapport du Ministère de l 'Economie, des Finances et de l'Industrie, Octobre 1998.

[34] B. Munch,*"IP Telephony - Today/Tomorrow/Ever?*", Ericsson Australia, July 1998.

[35] S. Krawczyk, *"La téléphonie sur IP Convergence, Enjeux et Composants*", Analyse IDC pour Cisco Systems, 2002.

[36] ITU-T Recommandation. G.729, *"Coding of Speech at 8 kbits/s using Conjugate Algebraic Code-Excited Linear Prediction (CS-ACELP)*", June 1995.

[37] Romain Trilling"*Codage Large Bande de la Parole Par Encapsulation du Codeur ITU-G729(CS-ACELP)*" mémoire de maîtrise en sciences appliquées Spécialité génie informatique Sherbrook(Québec), Canada –Août 1998.