République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Ecole Nationale Polytechnique

Département d'Electronique
Laboratoire du signal & communications

Mémoire de Magister

En électronique
Option : signal et communications

Présenté par : Mustapha MEDJBER
Ingénieur à UST Blida

Thème

Amélioration du standard G.729 -8Kb/s par la méthode de modification de l'échelle temporelle (WSOLA)

Soutenu devant le Jury:

Mr. M.Trabelsi, Maître de conférences à L'E.N.P	Président
Mr. D.BERKANI, Professeur à l'ENP	Rapporteur
Mme. F.MERAZKA, Maître de conférences à L' USTHB	Rapporteur
Mme. M.GUERTI, Maître de conférences à L'E.N.P	Examinateur
Mme. L.Hamami, Maître de conférences à L'E.N.P	Examinateur
Mr. R. Zergui, Chargé de cours à L'E.N.P	Examinateur

Année universitaire 2006-2007
Ecole Nationale Polytechnique, 10, Avenue Hacène Badi El-Harrach-Alger

Dédicaces

Je dédie ce travail :

A ma mère et mon père, qui m'ont aimé et assuré mon départ dans ma vie, J'espère qu'ils trouveront dans ce travail toute ma reconnaissance et tout mon amour.

A mes grands parents.

A ma très chère sœur Sarah.

A mes très chères frères : Hassen, Salah, Farid, Yacine et Khaled.

A mes très chères belles-sœurs Nessrine et Ibtissam.

A la mémoire de mon grand père Ahmed et mon oncle Mohamed.

A tous les membres de ma famille.

A mes amis.

A tous ceux qui ont contribué un jour à notre éducation et formation.

Remerciements

Ce travail a été réalisé au Laboratoire du signal & communications de l'Ecole Nationale Polytechnique.

J'ai eu le plaisir et la chance de travailler sous l'orientation de Monsieur le Professeur Daoud BERKANI qui m'a fait profiter en toute sympathie de sa grande compétence tant sur le plan expérimental que théorique. Je le remercie sincèrement.

Je tiens à exprimer toute ma reconnaissance à Madame Fatiha MERAZKA qui m'a proposé ce sujet d'actualité et a dirigé ce travail de magister pendant deux ans. Je lui suis également très reconnaissant pour m'avoir donné tous les outils et la documentation nécessaire ainsi que ses conseils précieux pendant toutes ces années.

Je remercie le Docteur Mohamed TRABELSI,le Docteur Mhania GUERTI, le Docteur Latifa HAMAMI et Monsieur Rachid ZERGUI pour l'intérêt qu'ils ont manifesté à l'égard de cette thèse en acceptant de faire partie du jury.

Enfin je remercie toutes les personnes ayant participé de loin ou de près à ce travail, qu'ils trouvent ici l'expression de ma profonde gratitude.

ملخص:

إن المكالمات عن طريق شبكة الانترنت(réseau IP), تتم عن طريق إرسال الحزمات. على مستوى الاستقبال يتم ضياع بعض الحزم و هذا راجع لسبب التأخر, الازدحام أو لأخطاء الإرسال. هذا الضياع في الحزم يؤدي إلى رداءة في نوعية الصوت عند الاستقبال.

إن الهدف من هذا البحث هو تحسين الرامزة الكلامية ITU-T G.729 المستعملة في مجال إرسال الصوت عن طريق الانترنت(VoIP). هذا التحسين يخص استرجاع الحزم الضائعة.

أكدت التجارب أن آلية استرجاع الحزم الضائعة المقترحة في الرامزة الكلامية G.729 أظهرت عجزا و أعطت نوعية رديئة للكلام و خاصة عندما يكبر عدد الحزم الضائعة, هذه الرداءة ترجع إلى انتشار الأخطاء و هذا حتى بعد استقبال حزم صحيحة.

سنقترح آلية فعالة لاسترجاع الحزم الضائعة للرامزة الكلامية G.729 سنقوم بتغيير السلم الزمني(TSM) و هذا باستعمال تقنية الإضافة بالتداخل للأمواج المتشابهة(WSOLA) من اجل إعادة إنشاء إشارة تنبيه الحزم الضائعة.

إن تقييم نوعية الكلام عن طريق تجارب ذاتية(PESQ) و غير ذاتية(EMBSD) تثبت أن الآلية المقترحة تحسن بصفة معتبرة نوعية الكلام المعاد إنشاؤه.

الكلمات المفاتيح :

ITU-T G.729، الكلام عن طريق الانترنت، استرجاع الحزم الضائعة، تغيير السلم الزمني، الإضافة بالتداخل للأمواج المتشابهة،EMBSD, PESQ.

Résumé

La transmission de la voix sur l'Internet (réseau IP) se fait par paquets. Au récepteur, certains paquets manquent dû aux délais, à la congestion ou aux erreurs de transfert. Cette perte de paquets dégrade la qualité de la voix au niveau du récepteur d'un système de transmission IP.

L'objectif de notre travail est d'améliorer le codeur de la parole ITU-T G.729 dédié à la transmission de la voix sur IP (VoIP). Cette amélioration concerne la récupération des trames perdues. Les tests ont montré que le mécanisme de récupération des trames perdues du G.729 présente une incapacité et donne une mauvaise qualité de la parole surtout lorsque le taux de perte augmente, cette dégradation est due à la propagation de l'erreur même après réception des bonnes trames. Nous proposons un mécanisme efficace de récupération des trames perdues pour le Codeur G.729. Nous effectuons une modification de l'échelle temporelle (Time-Scale Modification TSM) en utilisant la technique de l'ajout en chevauchement des similarités des ondes (WSOLA) pour reconstruire le signal d'excitation des trames perdues. L'évaluation de la qualité de la parole par des tests subjectifs (PESQ) et tests objectifs (EMBSD) démontre que la méthode proposée améliore considérablement la qualité de la voix reconstruite.

Mots clés :

ITU-T G.729, VoIP , Récupération des trames perdues (PLC), Modification de l'échelle temporelle (TSM), Ajout en chevauchement des similarités des ondes (WSOLA), PESQ, EMBSD.

Abstract

Voice-over-IP (VoIP) uses packetized transmission of speech over the Internet (IP network). However, at the receiving end, packets are missing due to network delay, network congestion (jitter) and network errors. This packet loss degrades the quality of speech at the receiving end of a voice transmission system in an IP network.

The goal of our work is to improve the ITU-T G.729 speech CoDec frequently used in transmission of the voice on IP (VoIP).This improvement relates to the packet loss concealment. The tests showed that the mechanism of packet loss concealment (PLC) embedded in G.729 has an incapacity and gives a bad speech quality especially when the rate of loss increases, this degradation is due to the propagation of the error even after reception of good packets. We propose a mechanism of packet loss concealment for the G.729 coder. We perform a Time-Scale Modification (TSM) using a Waveform similarity overlap-add (WSOLA) technique to reconstruct the excitation signal of the lost frames. The evaluation of speech quality by using subjective tests (PESQ) and objectives tests (EMBSD) shows that the proposed method improves considerably the quality of the reconstructed speech.

Key words:

ITU-T G.729, Voice over IP(VoIP), Packet loss concealment, Time-Scale Modification(TSM), Waveform similarity overlap-add (WSOLA), PESQ, EMBSD.

Sommaire

Liste des abréviations

ACELP	Algebraic Code Excited Linear Prediction
ADM	Adaptive Delta Modulation
ADPCM	Adaptive Differential Pulse Code Modulation
AMDF	Average Magnitude Difference Function
APC	Adaptive Predictive Coding
AR	Auto-Regressive.
ARMA	Auto-Regressive Moving Average
ARQ	Automatic Repeat reQuest
CCITT	Comité Consultatif International de la Téléphonie et la Télégraphie
CELP	Code Excited Linear Prediction
CoDec	Coder / Decoder
CS-ACELP	Conjugate Structure-Algebraic Code Excited Linear Prediction
dB	Decibel
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DSP	Digital Signal Processor
EMBSD	Enhanced Modified Bark Spectral Distortion
FEC	Forward Error Correction
FER	Frame Error Rate
IEEE	Institute of Electrical and Electronic Engineers
IETF	Internet Engineering Task Force
IP	Internet Protocol
ITU	International Telecommunications Union
LAN	Local Area Network
LAR	Log-Area Ratios
LP	Linear Prediction
LPC	Linear Predictive Coding
LSF	Line Spectral Frequency
LSP	Line Spectral Pair
MA	Moving Average

MGCP	Media Gateway Control Protocol
MIPS	Millions d'Instructions Par Seconde
MOS	Mean Opinion Score
OLA	OverLap-Add
OSI	Open Systems Interconnection
PCM	Pulse Code Modulation
PDF	Probability Density Function
PESQ	Perceptual Evaluation of Speech Quality
PLC	Packet Loss Concealment
PSTN	Public Switched Telephone Network
QoS	Quality of Service
RNIS	Réseau Numérique à Intégration de Services
RSVP	Ressource reSerVation Protocol
RTP	Real-time Transport Protocol
SD	Spectral Distortion
SIP	Session Initiation Protocol
SNR	Signal to Noise Ratio
SNRseg	SNR Segmental
SOLA	Synchronous Overlap and Add.
SQ	Scalar Quantization
SVQ	Split Vector Quantization
TCP	Transmission Control Protocol
TDHS	Time Domain Harmonic Scaling
TD-PSOLA	Time Domain Pitch Synchronous Overlap-Add
TFCT	Transformée de Fourier à Court Terme
TIMIT	Texas Instruments- Massachusetts Institute of Technology
TSM	Time-Scale Modification
UDP	User Datagram Protocol
ULP	Uneven Level Protection
VAD	Voice Activity Detector
Vocoder	Voice Coder
VoIP	Voice over IP
VQ	Vector Quantization
WSOLA	Waveform Similarity Overlap-Add

Liste des figures

Liste des Tableaux

Introduction

Dans les systèmes numériques modernes, le signal parole est représenté sous forme numérique (séquence d'éléments binaires, bits), il est nécessaire de représenter le signal par un nombre minimum de bits possible. Ainsi, pour le stockage de données, réduire le nombre de bits signifie l'économie de la mémoire. Pour les transmissions, réduire le débit binaire signifie l'économie de la bande passante. Il est donc nécessaire d'utiliser un algorithme efficace de compression de la voix. Le traitement qui permet d'effectuer une telle opération est appelé codage.

Durant ces trois dernières décennies, on a vécu deux principales évolutions technologiques, la première consiste en l'utilisation de la représentation numérique du signal parole. La seconde est le développement et le déploiement des réseaux à commutation de paquets. Ces deux technologies ont commencé à converger vers les années 80 par des recherches et des expériences sur la paquetisation de la voix. Ce n'est qu'en 1992 que les premières recherches sur la transmission de la voix sur Internet (VoIP) ont été faites. Récemment, on a vu le développement des architectures de téléconférence sur Internet incluant des protocoles de base pour des applications en temps réel.

La motivation pour la convergence est, cependant, la création d'une infrastructure de communications unique et intégrée offrant des services avancés et maniables pour satisfaire la demande d'utilisateur d'accéder à tous les services d'une plateforme unifiée avec un coût très réduit (intégration voix et données).

La VoIP offre plusieurs avantages par rapport aux réseaux téléphoniques classiques (PSTN), nous pouvons citer, entre autres, l'intégration voix et données en un seul réseau, l'optimisation et l'efficacité d'exploitation de la bande passante, la diminution des tarifs et des coûts des communications, la facilité d'administration, de supervision et de maintenance des réseaux IP et l'augmentation et l'amélioration des services... etc.

Pour qu'elle devienne une alternative crédible aux réseaux téléphoniques traditionnels (PSTN), le système VoIP doit offrir la même fiabilité et qualité de voix. Une bonne qualité de la voix de bout en bout dans les réseaux à commutation de paquets dépend principalement des facteurs dits facteurs de qualité de service (QoS). Ces facteurs ne sont pas garantis par le réseau Internet qui fournit un service d'acheminement des paquets avec meilleur effort «Best-Effort », nous pouvons citer le CoDec de la voix, le retard de bout en bout, la gigue et la perte des paquets.

Dans un système VoIP, au niveau du récepteur, certains paquets peuvent manquer, à cause des délais, à l'encombrement ou aux erreurs de transfert. La perte de paquets dégrade la qualité de la voix et se traduit par des ruptures au niveau de la conversation et une impression de hachure de la parole. Il est, par conséquent, indispensable de mettre en place un mécanisme de dissimulation de perte de paquets. Plusieurs algorithmes de masquage des pertes de paquets PLC (Packet Loss Concealment) sont utilisés aussi bien au niveau de l'émetteur qu'au niveau du récepteur.

Il existe plusieurs techniques de masquage (PLC) basées sur l'émetteur, ces techniques sont généralement plus efficaces mais plus complexes. On peut distinguer deux types de techniques, celles qui ajoutent une redondance de contrôle et celles qui n'ajoutent pas. Les méthodes qui ajoutent la redondance nécessitent une très large bande passante ou un long retard de bout en bout, parmi ces méthodes, on distingue celles qui envoient des paquets dupliqués, celles qui envoient avec les paquets courants des paquets précédents codés avec un faible débit, celles qui envoient des bits de correction d'erreurs sur les paquets en utilisant la méthode de correction d'erreurs en aval (FEC Forward Error Correction), ou encore celles qui utilisent la requête de répétition automatique (ARQ Automatic Repeat Request). Les méthodes qui n'ajoutent pas de redondance, utilisent une redondance inhérente dans la trame de voix au niveau de la source. Une méthode typique entrelace les échantillons de la voix dans des paquets distincts et reconstruit les paquets perdus par interpolation en utilisant les échantillons survivants voisins.

Les techniques de masquage basées sur le récepteur consistent à produire des remplacements semblables aux paquets originaux perdus. Il existe trois catégories de méthodes de dissimulation : l'insertion, l'interpolation et la régénération.

Objet de ce mémoire

Le standard G.729 utilise le codage prédictif, dans ce type de codage, la perte des paquets cause une perte de synchronisation entre le codeur et le décodeur. Donc, les erreurs ne se produisent pas seulement dans les trames perdues, mais se propagent aussi dans les trames suivantes, jusqu'à ce que le décodeur soit resynchronisé avec le codeur.

Les études ont montré que la perte d'un seul paquet peut être bien dissimulée par le décodeur du G729 mais lorsque le nombre de paquets perdus augmente de un à deux ou plus, l'énergie du signal d'erreur augmente considérablement et le MOS (Mean Opinion Score) décroît **[1]**.

Notre travail consiste en l'amélioration du CoDec G.729 de l'UIT par implémentation d'une nouvelle technique de dissimulation des trames perdues basée sur le récepteur. Cette technique consiste en la modification de l'échelle du temps (TSM Time Scale Modification) (changer la durée du signal parole sans changer la période du fondamental « pitch») du signal de parole basée sur la méthode WSOLA (Waveforme Similarity Ovelapp-Add) afin de compléter les paquets perdus **[2]**. La modification de l'échelle du temps a pour objet d'étirer le signal en préservant le même débit de parole afin de générer les échantillons perdus. Ainsi, Un bon algorithme basé sur la modification de l'échelle du temps est celui qui produit un signal avec des caractéristiques sonores naturelles. Avec la méthode WSOLA les propriétés du signal de parole sont préservées telles que l'intelligibilité, la qualité tonale et la reconnaissance du locuteur.

Plan du document

Nous avons divisé notre document en six chapitres, après une introduction générale, nous donnons au premier chapitre des généralités sur le codage de la parole, nous consacrons, dans un second temps, un chapitre à la prédiction linéaire en codage de la parole, dans le troisième chapitre, nous abordons la transmission de la voix sur le réseau IP, la quatrième partie de ce document est dédiée à la présentation du codeur G729, le cinquième chapitre traite les techniques de modification de l'échelle du temps et la méthode WSOLA, la dernière partie expose l'évaluation et les résultats relatifs à l'amélioration du codeur G729, enfin, nous terminerons notre travail par une conclusion.

Chapitre 1
Généralités sur le codage de la parole

1.1 Introduction

La parole est le moyen de communication par excellence entre les êtres humains. C'est pourquoi le traitement de la parole est devenu aujourd'hui une composante fondamentale des sciences de l'ingénieur. Le traitement de la parole est divisé en trois disciplines, la reconnaissance, soit du locuteur soit de la parole, la synthèse, et le codage de la parole. Dans notre travail on se limite au codage de la parole qui permet une transmission ou un stockage de la parole avec un débit réduit.

Dans le codage de la parole, il est crucial de comprendre la physiologie de la phonation, les propriétés de base du signal parole et sa perception, en effet, la représentation source-filtre des codeurs actuels tels que les Vocodeurs n'est qu'une modélisation du système phonatoire humain. L'existence d'une corrélation entre les échantillons du signal parole traduite par une redondance dans l'information a permis aux codeurs de réduire considérablement leurs débits. Ce chapitre regroupe les notions fondamentales sur la parole, sa perception, sa production et ces propriétés **[3,4]**.

1.2 La production de la parole

1.2.1 L'appareil vocal

L'appareil vocal, ou système phonatoire, comprend quatre éléments fondamentaux fonctionnant en étroite synergie pour produire des signaux acoustiques.

1- La soufflerie : elle est constituée par les poumons et la trachée artère.

2- Le vibrateur : est représenté par le larynx, qui est un ensemble de muscles et de cartilages mobiles qui entourent une cavité située à la partie supérieure de la trachée, il contient les ***cordes vocales*** qui sont en fait deux lèvres symétriques, elles peuvent fermer complètement le larynx, en s'écartant progressivement, elles déterminent une ouverture triangulaire appelée ***glotte***.

3- Le corps sonore : est formé du pharynx, de la cavité buccale et de la cavité nasale.

4- Le système articulateur : il intègre un ensemble d'organes mobiles, le voile du palais, la mâchoire inférieure, la langue et les lèvres **[5,6]**.

1.2.2 Mécanisme de phonation

Les poumons, actionnés par les muscles du thorax et de l'abdomen génèrent l'énergie nécessaire à la production de sons, en poussant l'air à travers la trachée artère, vers le larynx qui module l'air par le biais des cordes vocales. L'air ainsi modulé, ou onde glottique, est ensuite envoyé vers le conduit vocal formé par le pharynx et les différentes cavités, qui constitue un filtre dynamique ayant des paramètres variables commandés par le système articulateur qui intègre la mâchoire inférieure, la langue et les lèvres, ces dernières changent la taille et la forme du conduit vocal, le voile du palais ou *luette* sert comme commutateur pour la cavité nasale, lorsqu'elle est en position basse, la cavité nasale s'ajoute au conduit vocal. La parole résulte par la convolution de l'onde glottique avec la fonction de transfert de ce filtre (voir **Fig 1.1**).

Les sons peuvent être classifiés en terme de *phonèmes* ou unités acoustiques ayant une signification linguistique, ces phonèmes sont classés suivant la manière et la position des articulations qui se réfèrent au degré et à l'endroit des constrictions dans le conduit vocal.

Les voyelles sont généralement des sons voisés ayant des formants stables tout le temps, les sons tels que les fricatives, les occlusives et les nasales sont produits suivant la forme des constrictions du conduit vocal, les fricatives telles que /s/ ou /z/ sont produits lorsque l'air est forcé à travers des étroites constrictions, les occlusives tels que /b/,/d/,/g/,/p/,/t/ ou /k/ correspondent à une fermeture complète momentanée du conduit vocal, enfin, les consonnes nasales /m/,/n/ font intervenir les cavités nasales par abaissement du voile du palais **[7] et [8] p 2 et 3.**

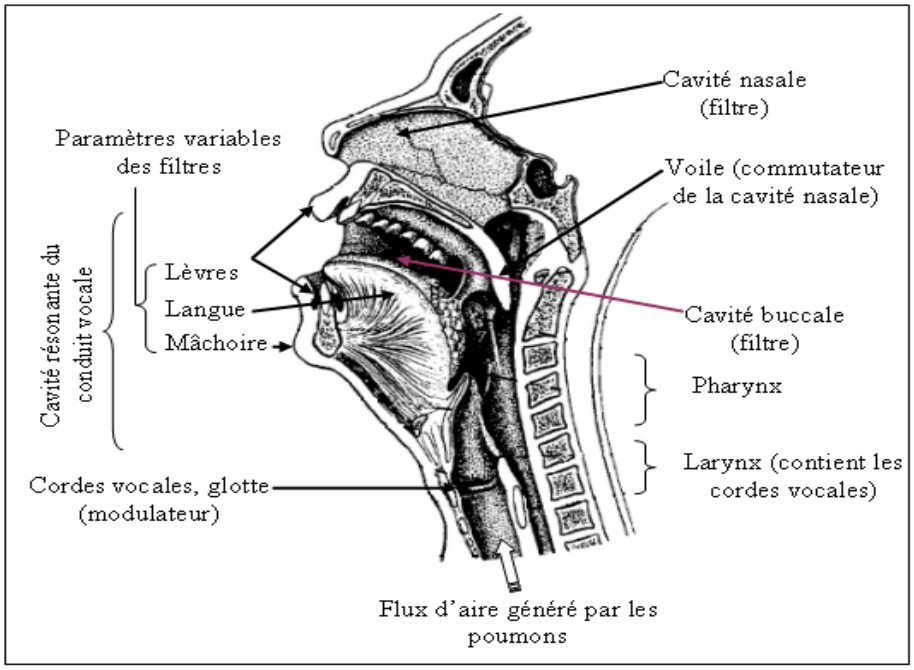


Fig 1.1 L'appareil phonatoire.

1.3 Classification des sons

Les signaux de parole peuvent être classifiés en deux catégories : signaux **voisés** caractérisés par des segments quasi-périodiques et d'énergie élevée tels que les voyelles, et signaux **non voisés** qui présentent généralement des segments de basse énergie tels que les consonnes.

Les sons voisés résultent, d'une vibration périodique des cordes vocales. Le larynx est d'abord complètement fermé, ce qui accroît la pression en amont des cordes vocales et force ces dernières à s'ouvrir, ce qui fait tomber la pression en permettant aux cordes vocales de se refermer. Des impulsions périodiques de pression sont ainsi appliquées au conduit vocal. Le taux auquel les cordes vocales s'ouvrent et se ferment s'appelle la **fréquence fondamentale** (dénotée par F_0) qui correspond physiquement au **pitch** perçu, sa valeur change avec la taille du conduit vocal. La fréquence fondamentale peut varier de 80 à 200 *Hz* pour une voix masculine, de 150 à 450 *Hz* pour une voix féminine, et de 200 à 600 *Hz* pour une voix d'enfant.

Les sons voisés présentent dans le domaine temporel un signal quasi-périodique tandis qu'ils présentent une structure harmonique dans le domaine fréquentiel, l'espacement entre les harmoniques est égal à la fréquence fondamentale F0 (pitch). L'enveloppe spectrale possède une structure formantique, elle est caractérisée par un nombres de pics (dénotés par Fi), chacun d'eux

est appelé **Formant,** Les trois premiers formants sont essentiels pour caractériser le spectre vocal. Les formants d'ordre supérieur ont une influence plus limitée. La structure formantique est attribuée au conduit vocal qui agit comme un filtre ayant comme résonances les pôles de la fonction de transfert ou formants, et comme anti-résonances les zéros de la fonction de transfert **[9]** et **[10].**

Les sons non voisés sont le résultat du passage du flux d'air par une étroite constriction au niveau du conduit vocal causant des turbulences, c'est-à-dire du bruit. Contrairement aux sons voisés, les sons non voisés ne présentent pas de structure périodique. Ils peuvent être modélisés par un bruit blanc filtré par le conduit vocal. La structure fine du spectre est, de ce fait, sensiblement la même sur tout le spectre. Notons que contrairement aux sons voisés, leur énergie est plus concentrée dans les hautes fréquences. La figure **Fig 1.2** présente des segments de sons voisés et non voisés avec leurs spectres correspondants.

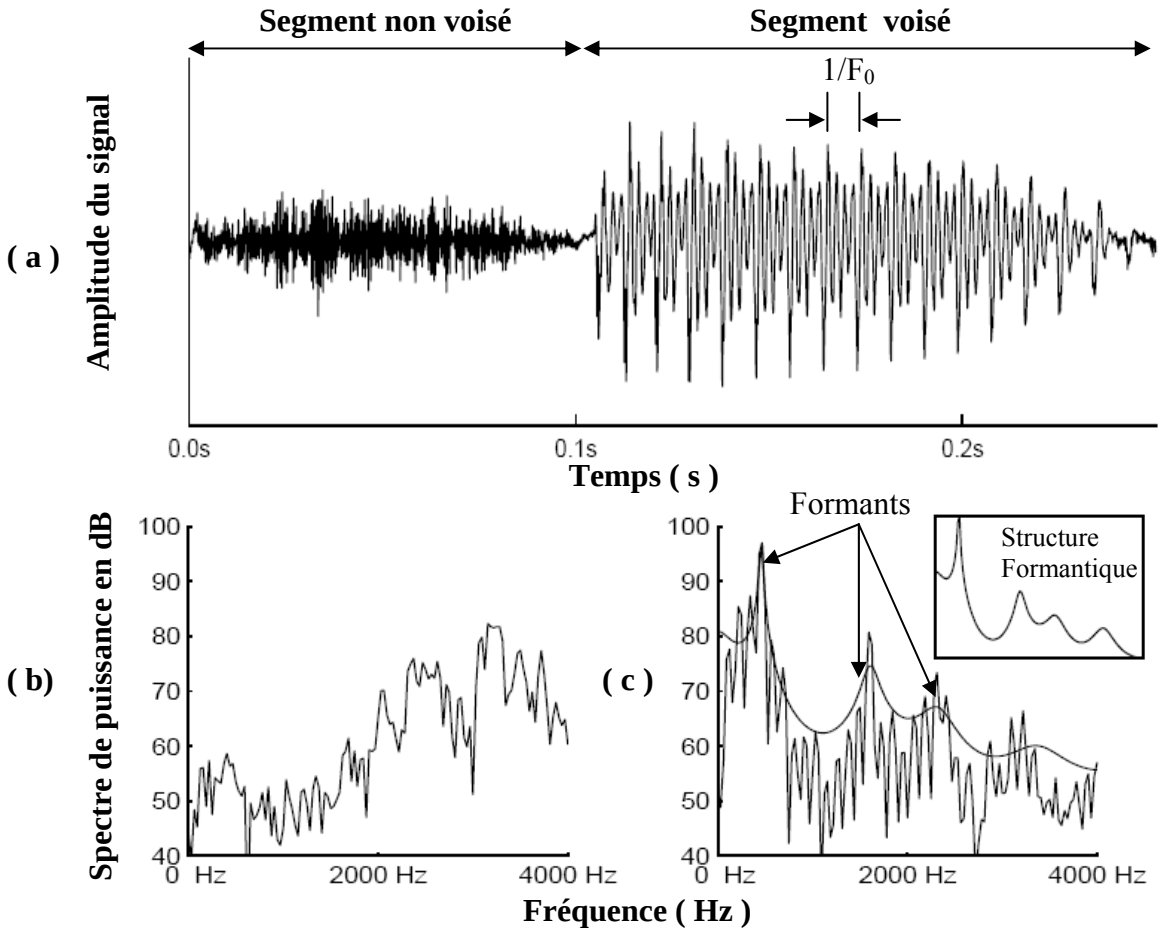


Fig 1.2 Représentation dans le domaine temporel et fréquentiel des segments de sons voisés et non voisés. (a) Deux segments de voix dans le domaine temporel, l'un non voisé et l'autre voisé. (b) Spectre de puissance pour un segment de 32 ms non voisé commençant à 50 ms. (c) Spectre de puissance et sa structure formantique correspondante pour un segment de 32 ms voisé commençant à 150 ms.

1.4 Le Modèle de Production de la Parole

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

Fant a proposé en 1960 un modèle de production qui spécifie qu'un signal voisé peut être modélisé par le passage d'un train d'impulsions *u(n)* à travers un filtre numérique récursif de type tous-pôles. On montre que cette modélisation reste valable dans le cas des sons non voisés, à condition que *u(n)* soit cette fois-ci un bruit blanc. Le modèle final est illustré à la figure **Fig1.3**. Il est souvent appelé modèle auto régressif (AR), parce qu'il correspond dans le domaine temporel à une régression linéaire de la forme

$$S(n) = G \cdot u(n) + \sum_{i=1}^{p} - a_i S(n-i) \qquad (1.1)$$

où *u(n)* et *p* sont respectivement le signal d'excitation et l'ordre du système. Chaque échantillon est obtenu en ajoutant un terme d'excitation à une prédiction obtenue par combinaison linéaire des *p* échantillons précédents. Les coefficients du filtre $\{a_i\}$ sont appelés coefficients de prédiction et le modèle AR est souvent appelé modèle de prédiction linéaire LP (linear prediction). Les paramètres du modèle AR sont : la période du train d'impulsions (sons voisés uniquement), la décision sur le son voisé/non voisé, le gain *G* et les coefficients du filtre 1/A(z), appelé filtre de synthèse.

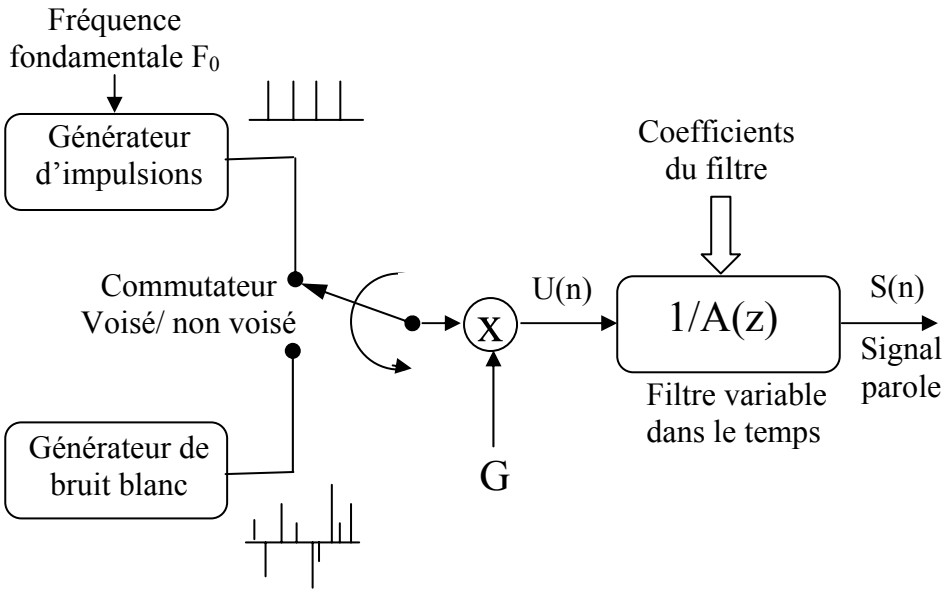


Fig 1.3 Modèle simplifié de production de la parole.

Par analogie entre le modèle physique et le modèle mathématique, ont peut donner les relations d'équivalences suivantes **[11]**:

Conduit vocal	$\Longleftrightarrow$	1/A(z), le filtre LP
Le flux d'air	$\Longleftrightarrow$	Le signal d'excitation $u(n)$ ou signal résiduel ou encore signal d'erreur de prédiction
Vibration des cordes vocales	$\Longleftrightarrow$	voisé
Période de vibration des cordes vocales	$\Longleftrightarrow$	$T = 1/F_0$, période du pitch
Fricatives et plosives	$\Longleftrightarrow$	non voisé/voisé
Volume d'air	$\Longleftrightarrow$	G , le gain

Le problème de l'estimation d'un modèle AR, souvent appelée analyse *LP* revient à déterminer les coefficients d'un filtre tous-pôles dont on connaît le signal de sortie, mais pas celui de l'entrée. Il est par conséquent nécessaire d'adopter un critère, afin de faire un choix parmi l'ensemble infini de solutions possibles. Le critère généralement utilisé est celui de la minimisation de l'énergie de l'erreur de prédiction.

1.5 Considérations pratiques de l'analyse LP

Pour mener à bien une analyse LP, il faut bien choisir :

- la fréquence d'échantillonnage *fe*.
- la méthode d'analyse et l'algorithme correspondant.
- l'ordre *p* de l'analyse LP.
- le nombre d'échantillons *N* par fenêtre d'analyse et le décalage entre fenêtres successives *L*.

Le choix de la fréquence d'échantillonnage est fonction de l'application visée et de la qualité du signal à analyser. Pour la téléphonie, on estime que le signal garde une qualité suffisante lorsque son spectre est limité à 3400 Hz et l'on choisit *fe* = 8000 Hz. Pour les techniques d'analyse, de synthèse ou de reconnaissance de la parole, la fréquence peut varier de 6000 à 16000 Hz. Par contre pour le signal audio (parole et musique), on exige une bonne représentation du signal jusqu'à 20 kHz et l'on utilise des fréquences d'échantillonnage de 44.1 ou 48 kHz. Pour les applications multimédia, les fréquences sous-multiples de 44.1 kHz sont de plus en plus utilisées: 22.5 kHz, 11.25 kHz.

L'ordre d'analyse conditionne le nombre de formants que l'analyse est capable de prendre en compte. On estime en général que la parole présente un formant par 1 kHz de bande passante, ce qui correspond à une paire de pôles pour *Ap(z)*. Si on y ajoute une paire de pôles pour la modélisation de l'excitation glottique, on obtient les valeurs classiques de *p=10, 12, et 18* pour *fe=8, 10 et 16 kHz* respectivement. On 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 autour de ces valeurs : il devient inutile d'encore augmenter l'ordre, puisqu'on ne prédit rien de plus.

La durée des fenêtres d'analyse et leur décalage sont souvent fixés à 30 et 10 ms respectivement. Ces valeurs sont liées au caractère quasi-stationnaire du signal de parole. Enfin, pour compenser les effets de bord, on multiplie en général préalablement chaque fenêtre d'analyse par une fenêtre de pondération *w(n)* de type *fenêtre de Hamming* :

$$W(n) = 0.54 - 0.46\cos\frac{2\pi n}{N-1} \qquad \text{pour } 0 \le n \le N-1 \qquad (1.2)$$

1.6 Codage de la parole

Le codage de la parole consiste à réduire l'information contenue dans le signal parole tout en gardant une qualité satisfaisante du signal reconstitué.

Afin de réaliser un codage performant, assurant un débit minimum avec une bonne qualité, il est nécessaire de prendre en considération certains critères de performances :

1.6.1 Critères de performances dans le codage de la parole

- **Débit binaire :** On mesure le débit binaire d'une représentation digitale en bits par échantillon, ou bits par seconde (b/s) selon le contexte. Le débit en bits par seconde n'est que le produit de la fréquence d'échantillonnage et le nombre de bits par échantillon. La fréquence d'échantillonnage doit être au moins deux fois plus grande que la largeur de bande du signal correspondant. Dans le cas de la téléphonie, utilisant la bande de 3.2 KHz (200-3400 HZ), la fréquence d'échantillonnage utilisée est de 8 KHz **[12]**.

- **Qualité du signal :** La qualité de la parole peut être déterminée par des tests d'écoute qui calculent l'opinion moyenne des auditeurs. La qualité de la parole peut être aussi déterminée par des mesures objectives comme la prédiction du gain, la distorsion spectrale logarithmique, etc.

- **Complexité :** En réalité les algorithmes de codage sont exécutés sur des cartes DSP (Digital Signal Processor). Ces processeurs possèdent une mémoire de stockage et une vitesse (en MIPS Million Instructions per Second) limitées. Par conséquent, les algorithmes de codage de la parole ne doivent pas être complexes pour ne pas dépasser la capacité des cartes DSP modernes. D'autres mesures de complexité peuvent être signalées, telles que la taille physique du codeur ou du décodeur, son prix et sa consommation en puissance (en Watt ou en mW) qui constituent un important critère dans un système portable.

- **Retard de communication :** Ce retard représente la somme des délais algorithmique, délai de traitement, délai de transmission, et du délai de bufferisation. Ce retard n'est pas tolérable surtout pour les applications en temps réel, telle que la téléphonie. Pour remédier au problème de retard il faut réduire la complexité des algorithmes, améliorer les protocoles de communications et augmenter les performances des processeurs.

- **Sensibilité aux erreurs de canal** : Ce paramètre mesure la robustesse du codeur de la parole par rapport aux erreurs de canal, les erreurs qui sont souvent provoquées par la présence du bruit dans le canal, de la perte de paquets de signal et de l'interférence inter-symboles.

- **Largeur de bande du codeur** : Pour mieux exploiter la bande passante, il faut bien choisir la largeur de bande du codeur, en effet, on utilise des codeurs à bande étroite dans les applications où la haute qualité n'est pas exigée, pourvu que le signal soit intelligible, c'est le cas de la transmission téléphonique dont la largeur de bande varie de 200 à 3400 Hz. alors qu'on utilise des codeurs à large bande de 7 à 20 kHz dans les applications qui nécessitent une transmission audio de qualité supérieure.

1.6.2 Redondance d'information dans le signal parole

1.6.2.1 Rappel théorique

Considérons une source d'information constituée d'une suite de caractères x_i, ayant l'alphabet : X=[x1, x2, ...xL] . Si $p(x_i)$ est la probabilité associée à l'occurrence à priori de x_i, son apparition apporte une information :

$$I = -\log_2 p(x_i) \textbf{ [12]}. \qquad (1.3)$$

- L'entropie de la source ou quantité moyenne d'information est donnée par :

$$H(X) = \sum_{n=1}^{L} p(xi)I(xi) = -\sum_{n=1}^{L} p(xi)\log_2 p(xi) \qquad (1.4)$$

- L'efficacité du code est donnée par : $\eta = \frac{H(X)}{Lm \log D}$ (1.5)

où $H(X)$: l'entropie de la source

Lm : longueur moyenne des mots codes avec :

Log D : est la capacité du code où D représente le nombre de symboles de l'alphabet du code, pour le code binaire : $D = 2$ et $\log_2 D = 1$.

- La redondance ρ par définition égale à : ρ=1-η.

1.6.2.2 La redondance de la parole

Le signal vocal est caractérisé par une très grande redondance, en effet, pour la langue française, par exemple, dont la probabilité d'occurrence des phonèmes est connue, on obtient une entropie H=4.73 et si l'on admet que, dans la conversation courante, environ 10 phonèmes sont prononcés par seconde, l'information moyenne est inférieure à 50 bits/s (10 x 4.73 < 50). Ce chiffre est à comparer au débit binaire maximum admissible sur un canal téléphonique, qui fait aujourd'hui partie des connaissances de tout un chacun : la plupart des modems envoient des informations numériques sur nos lignes téléphoniques avec un débit binaire de 33.6 kbits/s. On en conclut que le signal vocal est extrêmement redondant. Cette redondance sera mise à profit par les techniques de *codage de la parole*, dont le but sera de diminuer le débit nécessaire au stockage ou à la transmission de la parole en gardant une qualité satisfaisante **[7].**

1.7 Classification des codeurs

Ces dix dernières années, un nombre important de codeurs de parole a été proposé et réalisé. Traditionnellement les codeurs de parole sont divisés en trois grandes classes, les codeurs en formes d'ondes, les codeurs paramétriques ou vocodeurs et les codeurs hybrides **[13,14]**.

Les codeurs en formes d'ondes opèrent à des débits hauts, néanmoins, ils fournissent une très bonne qualité de parole. Les codeurs paramétriques opèrent à de très bas débits mais produisent des signaux de qualité synthétique. Enfin, les codeurs hybrides combinent les deux techniques de codage, le codage en formes d'ondes et le codage paramétrique et fournissent une bonne qualité de parole pour des débits moyens.

1.7.1 Les codeurs en formes d'ondes

Ces codeurs essayent de reproduire la forme d'onde du signal d'entrée à coder. Ils sont conçus pour être indépendants du signal, ainsi, ils peuvent être employés pour coder une large variété de signaux. Généralement ils sont de faible complexité, ils fournissent des signaux de parole de bonne qualité à des débits au-dessus de 16 kbps. Le codage en formes d'ondes peut être effectué aussi bien dans le domaine temporel que dans le domaine fréquentiel.

1.7.1.1 Codeurs dans le Domaine Temporel

Les codeurs de formes d'ondes dans le domaine temporel réalisent le processus de codage sur des échantillons temporels du signal. Les méthodes de codage les plus connues dans le domaine temporel sont : le codage PCM (Pulse Code Modulation), le codage APCM (Adaptive Pulse Code Modulation), le codage DPCM (Differential Pulse Code Modulation), le codage ADPCM (Adaptive Differential Pulse Code Modulation), le codage DM (Delta Modulation), le codage ADM (Adaptive Delta Modulation) et le codage APC (Adaptive Predictive Coding). Dans ce qui suit, on décrit brièvement quelques schémas importants de codage dans le domaine temporel.

Les Codeurs PCM

C'est le plus simple type de codage de formes d'ondes. C'est essentiellement un processus de quantification échantillon par échantillon. N'importe quelle quantification scalaire peut être utilisée avec ce schéma, mais la forme de quantification la plus utilisée est la quantification logarithmique, dans laquelle on prend en considération la nature du signal parole qui possède une densité de probabilité proche d'une gaussienne, autrement dit, les faibles amplitudes du signal parole sont plus fréquentes par rapport aux grandes amplitudes (**Fig.1.4**). Deux variantes (incluses dans la norme CCITT G.711) **[15]** se sont répandues dans la téléphonie: La norme américaine (μ-Law), utilisée aux États-Unis et au Japon et la norme européenne (A-Law) utilisée dans le reste du monde et dans les communications internationales. Elles sont toutes deux des variations d'une correspondance exponentielle.

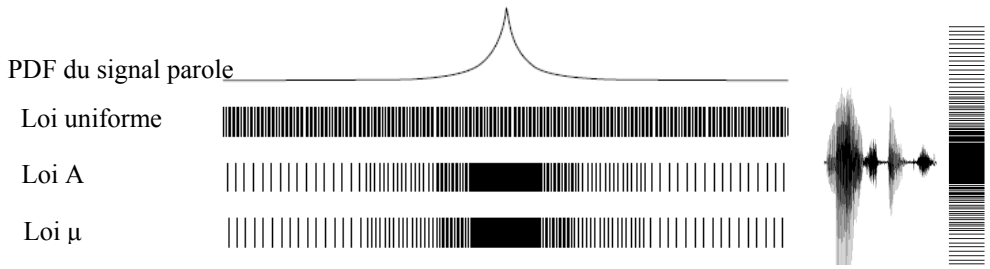


Fig 1.4 Codage PCM

Les Codeurs DPCM et ADPCM

La technique PCM ne fait aucune supposition sur la nature des formes d'ondes à coder, par conséquent, elle fonctionne bien pour des signaux différents de ceux de la parole. Cependant, en codant la parole, il existe une très forte corrélation entre les échantillons successifs obtenus. Cette corrélation peut être exploitée pour réduire le débit binaire. Une méthode simple de le faire est de transmettre uniquement la différence entre deux échantillons. Le signal différence possèdera alors une gamme dynamique plus réduite que le signal original, et peut être alors quantifié moyennant un nombre de niveaux de reconstitution plus réduit. Dans la méthode citée plus haut, l'échantillon précédent est utilisé pour prédire la valeur de l'échantillon présent. La prédiction sera améliorée si un bloc plus large de la parole est utilisé pour la prédiction. Cette technique est connue sous le nom de DPCM.

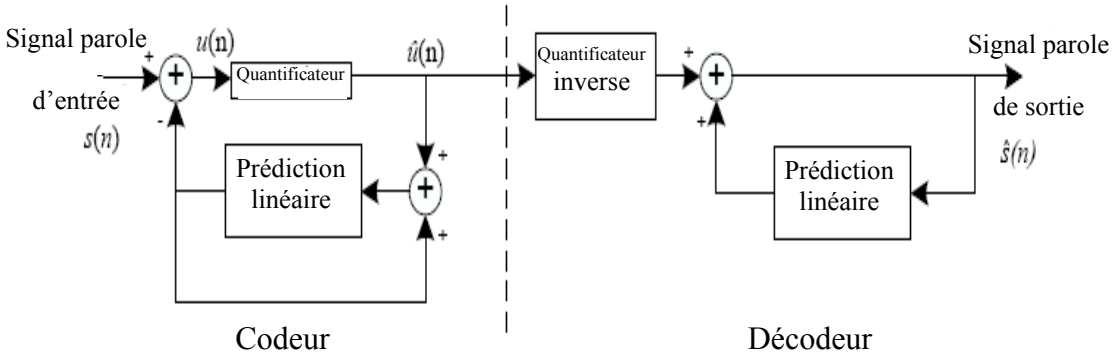


Fig 1.5 Schéma de principe du codeur DPCM : le codeur sur la gauche, et le décodeur sur la droite. Le quantificateur inverse, convertit les codes transmis en la valeur û(*n*).

Une version améliorée de la DPCM est la DPCM Adaptative (ADPCM) dans laquelle le prédicteur et le quantificateur sont adaptés aux caractéristiques locales du signal d'entrée. Il existe un bon nombre de recommandations de l'ITU basées sur les algorithmes ADPCM pour la bande étroite (fréquence d'échantillonnage de 8 kHz) et le codage audio comme le G.726 opérant à 40, 32, 24 et 16 kbits/s **[16]**. La complexité de l'ADPCM est légèrement faible.

1.7.1.2 Codeurs dans le Domaine Fréquentiel

Les codeurs de formes d'ondes dans le domaine fréquentiel divisent le signal en un nombre de composantes fréquentielles et code chacune d'elles séparément. Le nombre de bits utilisé pour coder chaque composante fréquentielle peut varier de manière dynamique. Les codeurs dans le domaine fréquentiel sont divisés en deux groupes : Les codeurs en sous bande (sub band coders) et les codeurs par transformée (transform coders).

Les Codeurs en Sous Bande

Les codeurs en sous bande emploient des filtres passe bande pour diviser le signal en un nombre de signaux passe bande (sub band signals) qui sont codés séparément. Au niveau du récepteur, les signaux en sous bande sont décodés et additionnés pour reconstruire le signal de sortie. L'avantage principal du codage en sous bande est que la quantification du bruit produit dans une bande est confiné uniquement dans cette bande. L'organisme ITU a standardisé en codage sous bande le codeur audio G.722, SB-ADPCM, qui code les signaux audio à large bande de 7 kHz échantillonnés à 16 kHz, pour une transmission à 48, 56 ou 64 kbits/s **[17]**.

Les Codeurs par Transformée

Cette technique transforme par bloc, un segment du signal d'entrée dans le domaine fréquentiel ou un domaine similaire. Le codage adaptatif est réalisé en attribuant plus de bits aux coefficients de transformation les plus importants. Au niveau du récepteur, le décodeur fait la transformation inverse pour obtenir le signal reconstruit. Plusieurs transformées comme la DFT (Discrete Fourier Transform) ou DCT (Discrete Cosine Transform) peuvent être utilisées.

1.7.2 Les codeurs paramétriques ou Vocodeurs

Les performances des codeurs paramétriques, connus aussi sous le nom de Vocodeurs, sont fortement dépendantes de la précision des modèles de production de la parole. Ces codeurs sont conçus spécifiquement pour des applications à bas débit et sont principalement destinés à maintenir une qualité satisfaisante de la parole. Les vocodeurs les plus efficaces sont basés sur la prédiction linéaire LP (Linear Prediction). Les détails de cette technique seront abordés dans le chapitre suivant. Une *qualité des communications* peut être obtenue à des débits inférieurs à 2 kbits/s avec les vocodeurs LP.

1.7.3 Les Codeurs Hybrides

Les codeurs hybrides sont conçus pour fournir une qualité aussi bonne à des débits relativement faibles ou moyens, ce sont donc, des codeurs intermédiaires entre les codeurs en formes d'ondes et les vocodeurs. Cependant, ces codeurs ont tendance à nécessiter un nombre d'opérations plus élevé. Virtuellement, tous les codeurs hybrides reposent sur l'analyse LP pour l'obtention des paramètres du modèle de synthèse. Les techniques de formes d'ondes utilisées pour coder le signal d'excitation et les modèles de production du pitch peuvent être incorporés pour améliorer les performances. A partir des années 80, l'intérêt pour les codeurs CELP (Code-Exited Linear Prediction) ne cesse d'augmenter.

Dans les codeurs CELP, l'analyse LP est utilisée pour obtenir le signal d'excitation. La modélisation du pitch est utilisée pour coder efficacement le signal d'excitation. Le standard G.729 de l'ITU est un codeur CELP qui produit une qualité téléphonique (toll quality) de la parole à 8 kbits/s **[18]**. Les codeurs de formes d'ondes par interpolation WI (Waveform Interpolation) modélisent le signal résiduel par des formes d'ondes caractéristiques qui peuvent être interpolées aussi bien dans le domaine temporel que fréquentiel pour la reconstitution du signal. Pour des débits inférieurs à 4 kbits/s, les codeurs WI donnent de meilleurs performances, comparés à d'autres codeurs opérant à des débits similaires. Cependant, les codeurs WI sont actuellement alourdis par leur complexité élevée et par leur retard (typiquement 40 ms).

Le tableau suivant présente un résumé des méthodes de codage les plus utilisées.

Standard (N)	Algorithm (N)	Complexity (MIPS)	Frame Size /lookahead(ms)	Compression	Bit rate (kb/s)	MOS (N)
G.711	PCM	0	0.125/0	1	64	4.10
G.726 G.727	ADPCM	1	0.125/0	4/2.7/2/1.6	16/24/32/40	3.85
G.722	SB-ADPCM	10	0.125/1.5	1.3/1.1/1	48/56/64	3.3
G.728	LD-CELP	30	0.625/0	4	16	3.61
G.729	CS-ACELP	20	10/5	8	8	3.92
G.729A	CS-ACELP	11	10/5	8	8	3.7
G.723.1	MPC-MLQ	16	30/7.5	10.2/12.1	6.3/5.3	3.9
GSM 06.10	RPE-LTP	10	20/0	4.9	13	3.5
IS-54	VSELP	24	20/5	8	8	3.54
IS-96	QCELP	20	20/5	7.5/16/32	8.5/4/2	–
FS-1016	CELP	30	–	13.3	4.8	3.0
FS-1015	LPC10E	15	–	26.7	2.4	2.4

Tableau 1.1 Tableau comparatif entre les différentes méthodes de codage **[14].**

1.8 Les différentes étapes suivies dans le processus de codage de la parole

Afin de coder la parole, plusieurs étapes sont nécessaires. Le signal subit tout d'abord un filtrage anti-repliement, puis un échantillonnage suivi d'une quantification et enfin le codage. L'**échantillonnage** est le processus de représentation d'un signal continûment variable par une séquence de valeurs. La **quantification** consiste à représenter approximativement chaque échantillon dans un ensemble fini de valeurs. Enfin, le **codage** consiste à assigner un numéro réel à chaque valeur.

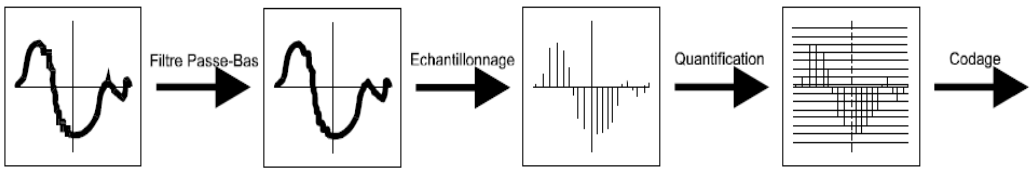


Fig 1.6 Etapes suivies pendant le processus de codage de la parole.

Filtrage anti-repliement

Avant l'échantillonnage, un filtre passe-bas de fréquence de coupure égale à la moitié de la fréquence d'échantillonnage est inséré pour éviter l'effet dénommé « repliement » ou « aliasing » postulé par le théorème de Nyquist-Shannon, ce filtre est appelé filtre « anti-repliement » ou « anti-aliasing ».

L'échantillonnage

L'échantillonnage transforme le signal à temps continu *x(t)* en un signal à temps discret *x(nTe)* défini aux instants d'échantillonnage, multiples entiers de la période d'échantillonnage *Te*, celle-ci est elle-même l'inverse de la fréquence d'échantillonnage *fe*. En ce qui concerne le signal vocal, le choix de *fe* résulte d'un compromis. Son spectre peut s'étendre jusque 12 kHz **[7]**. Il faut donc en principe choisir une fréquence *fe* égale à 24 kHz au moins pour satisfaire raisonnablement au théorème de Shannon **[12]**.

Cependant, le coût d'un traitement numérique, filtrage, transmission, ou simplement enregistrement peut être réduit d'une façon notable si l'on accepte une limitation du spectre par un filtrage préalable. C'est le rôle du filtre de garde, dont la fréquence de coupure *fc* est choisie en fonction de la fréquence d'échantillonnage retenue.

Pour la téléphonie, on estime que le signal garde une qualité suffisante lorsque son spectre est limité à 3400 Hz et l'on choisit *fe* = 8000 Hz. Pour les techniques d'analyse, de synthèse ou de reconnaissance de la parole, la fréquence peut varier de 6000 à 16000 Hz. Par contre pour les signaux audio (parole et musique), on exige une bonne représentation du signal jusqu'à 20 kHz et l'on utilise des fréquences d'échantillonnage de 44.1 ou 48 kHz. Pour les applications multimédia, les fréquences sous-multiples de 44.1 kHz sont de plus en plus utilisées: 22.5 kHz, 11.25 kHz.

1.9 La Quantification

La quantification est une partie intégrante dans le codage. C'est l'opération de discrétisation d'une ou plusieurs variables, C'est aussi l'approximation de la valeur instantanée exact d'un signal par une valeur voisine tirée d'une association de *N* valeurs discrètes.

Si on désigne par x une variable aléatoire, un quantificateur est un appareil qui fait associer à l'entrée x comprise dans un intervalle, une sortie y comprise dans le même intervalle. Donc la quantification est l'opération de substitution des échantillons d'un signal analogique par des valeurs arrondies prises parmi un nombre fini de valeurs possibles **[14]**.

La quantification peut être scalaire ou vectorielle selon que la variable x est à une ou plusieurs dimensions. On peut considérer la quantification scalaire comme étant un cas particulier de la quantification vectorielle lorsque la dimension de la variable x est un.

1.9.1 Quantification Scalaire (SQ)

La quantification scalaire attribue à une valeur d'entrée x une autre valeur plus proche appartenant à un ensemble fini prédéterminé, ou dictionnaire (codebook) de N valeurs $C=\{y_k \mid k=1,\ldots,N\}$ **[8]**. (1.6)

Le quantificateur divise le signal d'entrée en N intervalles I_k , de sorte que une entrée x appartenant à l'intervalle I_k soit codée avec la sortie y_k.

Pour coder un vecteur x de m dimensions par SQ, il faut quantifier chaque valeur de ce vecteur x_i indépendamment tel que :

$$x_q = y_{i,k} = Q_i(x_i),\ i=1,\ldots,m, \qquad (1.7)$$

La quantification scalaire introduit une erreur ou bruit $e = Q(x) - x$ (1.8), sur l'échantillon d'entrée x. la mesure de distorsion la plus répandue dans la conception SQ est l'erreur quadratique entre la valeur originale et la valeur quantifiée :

$$D(x,x_q) = |x-x_q|^2 = |e|^2. \qquad (1.9)$$

La performance d'un quantificateur scalaire est souvent évaluée en utilisant la moyenne de l'erreur quadratique :

$$D=\min E[d(x,x_q)] \qquad (1.10).$$

La quantification scalaire se divise en deux types, la quantification uniforme et la quantification non uniforme.

1.9.1.1 Quantification uniforme

Cette quantification est utilisée dans la conversion analogique-numerique, les intervalles de décision I_k ont la même valeur Δ et les niveaux de sortie y_k sont au milieu de ces intervalles, tels que :

$$\Delta = \frac{x_{max} - x_{min}}{N} \qquad (1.11)$$

$$I_k = \{x \mid x_k < x \le x_{k+1}\} \qquad (1.12)$$

$$y_k = x_{min} + (k - 0.5)\Delta, k = 1,\ldots, N, \qquad (1.13)$$

$$Q(x) = \{y_k | x \in I_k\} \qquad (1.14)$$

avec x_{min} et x_{max} sont respectivement le minimum et le maximum du signal d'entrée. L'opération de troncation ou d'arrondie des nombres réels en nombres entiers est un exemple de la quantification uniforme.

1.9.1.2 Quantification non uniforme

Malgré que les quantificateurs uniformes sont faciles à implémenter, ils n'offrent pas une bonne performance, on fait donc recours à la quantification non uniforme, qui consiste à varier les intervalles de décision I_k, en effet, on réduit la largeur des l'intervalles I_k dont les valeurs d'entrée ayant une grande probabilité d'apparition, tandis qu'on augmente celle ayant une faible probabilité d'apparition. La PCM loi A et loi μ sont les plus populaires des quantifications non uniformes.

1.9.2 Quantification vectorielle (VQ)

La quantification vectorielle (VQ pour Vector Quantization) est l'extension de la quantification scalaire à un espace multidimensionnel. La quantification vectorielle attribue un bloc ou un vecteur de valeurs d'entrée à un seul vecteur à partir d'un ensemble fini de vecteurs de sortie. Shannon a démontré que pour un débit donné, le codage de longs bocs d'information donne une très bonne performance de point de vu distorsion. La figure **Fig.1.7** illustre la structure de base d'un quantificateur vectoriel. Le quantificateur vectoriel ou l'encodeur, fait correspondre le vecteur d'entrée x à K dimensions à un symbole de canal, ou index i, qui sera transmis sur le canal. L'encodeur partitionne le vecteur d'espace multidimensionnel d'entrée en N régions tel que

$$P=\{R_1, R_2,\ldots, R_3\} \text{ où } R_i = \{x| \, d(x,y_i)\leq d(x,y_j), j\neq i\} \qquad (1.15)$$

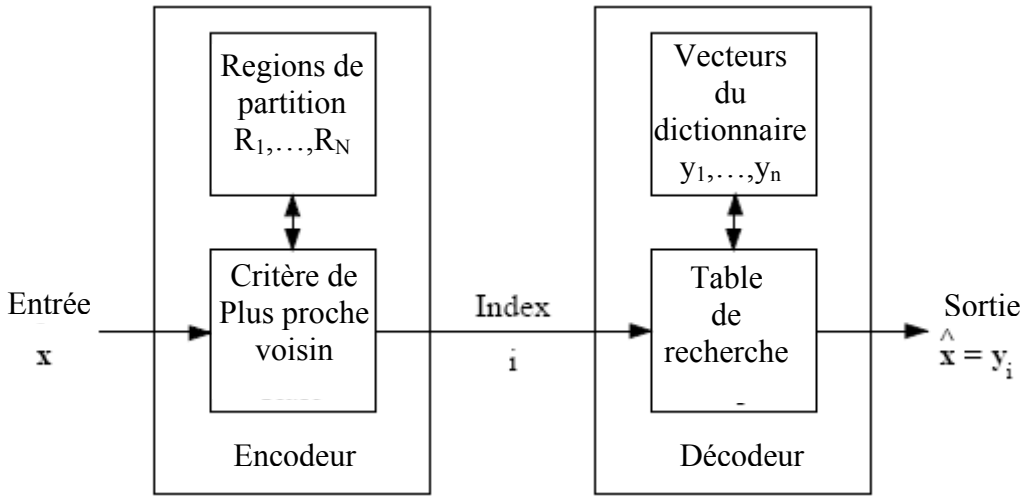


Fig 1.7 Modèle d'un quantificateur vectoriel.

Le vecteur yi est le vecteur de code associé avec la région Ri. L'index est choisi d'une manière à ce que x appartient à la partition Ri, le quantificateur inverse, ou décodeur, doit correspondre le symbole i à un code vecteur approprié $xq=y_i$ en utilisant une simple procédure de recherche.

1.9.2.1 Conditions sur la quantification vectorielle

Conditions d'optimalité

La performance de la quantification vectorielle dépend de l'espace de partitionnement et des vecteurs de reproduction, ou vecteurs de code (code vectors) du décodeur. Un VQ est optimal si la distorsion moyenne E[d(X,Xq)] est minimisée pour l'entrée X. il n'y a pas une méthode directe pour la conception d'un VQ, généralement on utilise des méthodes itératives. Deux conditions sont nécessaires pour l'optimalité du dictionnaire :
Une concerne l'encodeur, elle vise le partitionnement de l'espace d'entrée, c'est la condition du plus proche voisin, l'autre concerne le décodeur, dans ce cas il faut bien choisir le vecteur de code c'est la condition du centroide.

Condition du plus proche voisin

Pour un décodeur dont l'ensemble des vecteurs de code est C, la partition optimale R_i du codeur satisfait la condition :

$$R_i \subset \{x \mid d(x, y_i) \le d(x, y_j); \forall j\} \qquad (1.16)$$

c'est à dire que les régions de partition sont définies par les vecteurs de code $\{y_i\}$ dans C :

$$Q(x) = y_i \text{ seulement si } d(x, y_i) \le d(x, y_j) \forall j. \qquad (1.17)$$

Condition du centroide

Pour une partition de l'encodeur P={Ri| i=1,…,N}, les codes vecteurs optimaux y_i dans C sont les centroides dans chaque partition R_i :

$$y_i = \text{cent}(R_i) = \min_y E[d(x,y) \mid x \in R_i] \qquad (1.18)$$

Dans le cas d'utilisation de l'erreur quadratique comme mesure de distorsion dans la conception du quantificateur, les centroides sont les centres de gravite des partitions.

Il existe deux approches essentielles en quantification vectorielle, qui sont l'approche statistique et l'approche algébrique.

1.9.2.2 Approche statistique

Cette approche est basée sur un algorithme itératif désigné par *K-moyenne* ou *LBG* du nom de ces auteurs Lynde, Buzo et Gray.

Algorithme de Linde-Buzo-Gray (LBG)

Cet algorithme génère une partition d'un signal (séquence d'apprentissage), partant d'un dictionnaire initial composé des vecteurs les plus éloignés possibles. Ces vecteurs doivent bien sûr être représentatifs des vecteurs rencontrés dans les signaux à coder. L'algorithme itératif converge vers un dictionnaire optimal ou localement optimal.

Définition : Un quantificateur Q est dit (globalement) optimal s'il minimise une distorsion D donnée. Un quantificateur Q est localement optimal, si D(Q) est un minimum local.

1. On se donne un dictionnaire initial C^0 de Nc vecteurs Xqi, une mesure de distorsion d, un seuil $e \geq 0$, un compteur d'itération *l*=0, une distorsion moyenne D^{l-1} initialisée à une valeur très importante et une séquence d'apprentissage composée de n vecteurs X_j.

2. A l'aide du dictionnaire $C^l = (Xq_i, i = 1,..., Nc)$, trouver la partition $S^l = (S_i, i = 1,..., Nc)$ de la séquence d'apprentissage minimisant la distorsion, c'est à dire faire :

(a) pour tous les vecteurs X_j de la séquence d'apprentissage (j=1,…,n)

(b) pour tous les vecteurs Xqi du dictionnaire (i=1,…,Nc)

(c) si $d(X_j, Xq_i) \leq d(X_j, Xq_l) \forall l$ alors $X_j \in S_i$.

La recherche de la distorsion minimale définit une région de décision S_i pour chaque vecteur Xq_i du dictionnaire. Ainsi, chaque vecteur X_j de la séquence d'apprentissage inclus dans la région de décision S_i est approché par le vecteur Xq_i associé.

3. Calculer la distorsion moyenne $D^l = D(C^l, S^l) = \frac{1}{n}\sum_{j=1}^{n} \min_{Xq \in C^l} d(X_j, Xq)$ (1.19). Cette expression permet de calculer la moyenne des distorsions minimales entre les n vecteurs X_i de la séquence d'apprentissage et les vecteurs d'approximation Xq correspondants.

4. Si $\frac{D^{l-1} - D^l}{D^l} \leq e$ le dictionnaire C^l est conservé et la procédure s'arrête. Sinon, continuer.

5. Rechercher l'ensemble optimal des vecteurs d'approximation $Xq(S^l)=Xq(S_i)$, $Xq(S_i)$ est le barycentre de l'élément i de la partition donné par

$$Xq(S_i) = \frac{1}{\|S_i\|} \sum_{j:X_j \in S_i} X_j \qquad (1.20)$$

avec $\|S_i\| =$ le nombre de vecteurs d'apprentissage X_j inclus dans la cellule S_i.

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

Choix du dictionnaire initial :

Différentes solutions conduisent au dictionnaire initial (méthodes aléatoires, par exemple, et méthodes par division successives). Cette dernière approche, initialement incluse dans l'algorithme LBG, procède de la manière suivante. Le centroide, moyenne de tous les vecteurs de la séquence d'apprentissage, est calculé. Le vecteur résultant est divisé en deux, en rajoutant et retranchant la même quantité au vecteur.

L'algorithme LBG est ensuite appliqué pour construire le dictionnaire optimal de taille 2. Le processus est itéré pour construire successivement le dictionnaire optimal de taille 2^2,2^3,…, et enfin Nc.

1. on cherche d'abord le dictionnaire initial composé d'un seul vecteur $Xq_1(0)$ minimisant la distorsion moyenne : c'est le barycentre ou centroide de l'ensemble d'apprentissage ;
2. à partir de $Xq_1(0)$ on construit deux vecteurs $Xq_1(1)$ et $Xq_2(1)$ par : $Xq_1(1) = Xq_1(0) - \xi$ et $Xq_2(1) = Xq_1(0) + \xi$ avec ξ une valeur « petite ».
3. Connaissant $Xq_1(1)$ et $Xq_2(1)$, on classe tous les vecteurs de l'ensemble d'apprentissage en deux classes. On calcule le barycentre $Xq_1(2)$ de tous les vecteurs associés à $Xq_1(1)$, et le barycentre $Xq_2(2)$ de tous les vecteurs associés à $Xq_2(1)$.
4. on partage à nouveau ces deux vecteurs en deux ;
5. on arrête l'algorithme lorsque le nombre de vecteurs désiré Nc est atteint.

On peut accepter un calcul relativement lourd à ce niveau, le dictionnaire étant fait en temps différé et une fois, en préalable au codage en ligne.

La technique du VQ statistique permet de bien rendre compte de la distribution du signal à coder, mais la taille du dictionnaire et la dimension du vecteur d'entrée sont sévèrement limitées par des contraintes technologiques en matière de stockage et de complexité de calcul. Plusieurs techniques ont été proposées pour contourner le problème de complexité, comme l'organisation des dictionnaires en arbre, les quantificateurs vectoriels en Treillis (Lattice VQ), la quantification vectorielle à étages (Multi Stage VQ), la quantification vectorielle par division (Split VQ) et les quantificateurs vectoriels en forme de gain (gain shape vector quantizers).

1.9.2.3 Approche algébrique

L'inconvénient majeur de la quantification vectorielle statistique vient de la détermination du représentant le plus proche voisin d'un vecteur à coder qui se fait par une recherche exhaustive dans un dictionnaire stocké en mémoire. Ceci limite l'emploi de dimension importante. Des méthodes de codage accélérées et sous-optimales ont été développées mais conduisant généralement à un accroissement de l'allocation mémoire de stockage. Il serait donc intéressant de construire un dictionnaire virtuel qui nous évite le stockage de la totalité des représentants. Ce dernier doit posséder également une structure permettant un codage rapide des vecteurs. De telles structures existent et exploitent les propriétés des réseaux réguliers de points dans l'espace à plusieurs dimensions.

Les réseaux réguliers de points sont déjà utilisés dans le domaine de codage du canal en codes correcteurs d'erreurs, ils constituent une réponse au problème de l'empilement des sphères où il s'agit de trouver les coordonnées des centres de sphères tangentes, de rayon unité et qui remplissent au mieux l'espace.

1.10 Calcul de distorsion

La distorsion, dans la reproduction d'une source, est la mesure de la fidélité de reproduction de la source à la sortie. Pour la reproduction à haute fidélité, le signal reproduit doit être très proche du signal original avec une faible distorsion.

La mesure de distorsion est l'entité mathématique qui évalue l'approximation. Généralement, c'est une fonction qui attribue aux deux entités x et x_q le nombre non négatif tel que $d(x, x_q) \geq 0$, où x est la donnée originale, x_q est l'approximation, et $d(x, x_q)$ est la quantité de distorsion entre x et x_q.

La distorsion entre des vecteurs est la distorsion moyenne entre leurs éléments :

$$d(X^n, X_q^{\ n}) = \frac{1}{n}\sum_{i=1}^{n}(x_i, x_{q_i}) \qquad (1.21)\ .$$

La distorsion est souvent appelée distance. Il existe plusieurs méthodes de la calculer :

Distance de Hamming : $$d(x, x_q) = \begin{cases} 0, & \text{si } x = x_q \\ 1, & \text{si } x \neq x_q. \end{cases} \quad (1.22)$$

Distance de l'erreur quadratique : $$d(x, x_q) = (x - x_q)^2 \quad (1.23)$$

Distance de Minkow sky : $$d(X^n, X_q^n) = \sum_{i=1}^{n} \left| x_i - x_{q_i} \right| \quad (1.24)$$

Distance Euclidienne : $$d(X^n, X_q^n) = \left[\sum_{i=1}^{n} \left| x_i - x_{q_i} \right|^2 \right]^{1/2} \quad (1.25)$$

Distance de Chebychev : $$d(X^n, X_q^n) = \underset{i}{Max} \left| d(x_i, x_{q_i}) \right| \quad (1.26)$$

1.11 Conclusion

Afin de bien saisir et de pouvoir traiter le codage de la parole, il est indispensable de connaître le signal parole, ses caractéristiques et sa production, c'est pourquoi nous avons commencé ce chapitre par une description du signal de parole. Nous avons ensuite, présenté les critères de performance dans le codage de la parole et les différents types de codeurs de parole, nous avons parlé aussi sur les étapes suivies dans la procédure de codage de la parole. Le calcul de la distorsion est le dernier sujet à traiter dans ce chapitre.

Chapitre 2
La prédiction Linéaire
en Codage de la parole

2.1 Introduction

La prédiction linéaire LP (Linear Prediction) est l'une des méthodes les plus puissantes dans l'analyse du signal de parole pour l'estimation des paramètres essentiels du signal vocal, son succès est dû au fait qu'elle représente une solution linéaire au problème de l'estimation du modèle de production de la parole. Ont peut distinguer deux types de prédiction, la prédiction à court terme et la prédiction à long terme.

La prédiction à court terme cherche à éliminer la redondance entre les échantillons voisins, le filtre utilisé est appelé *filtre d'analyse* LP, il supprime la structure formantique du signal parole (c'est pourquoi il est appelé aussi filtre *d'analyse de formant*) et laisse l'erreur de prédiction de sortie à basse énergie qui est connue sous le nom *résiduel* ou *excitation.* L'inverse du filtre d'analyse est le filtre de synthèse, il modélise le conduit vocal et sa fonction de transfert décrit l'enveloppe spectrale du signal parole.

La prédiction linéaire à long terme est utilisée pour exploiter les corrélations à long terme existantes dans les signaux voisés. Ce filtre est appelé *prédicteur de pitch*, il exploite la périodicité du signal, l'inverse du predicteur de pitch est appelé *filtre de pitch* il modélise la fonction de la glotte et sa fonction de transfert décrit la structure harmonique du signal parole. Le predicteur de pitch n'a aucun effet pour les signaux non voisés puisque l'excitation non voisée est aléatoire et son spectre est monotone. **[19]**.

Dans cette partie on va étudier la prédiction linéaire à court terme.

2.2 L'analyse par prédiction linéaire à court terme

Le principe fondamental de la prédiction linéaire est qu'un échantillon du signal *S*(n) peut être modélisé comme la sortie d'un système Auto Régressif à Moyenne Ajustée (ARMA) avec une entrée u(n) :

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + G\sum_{l=0}^{q} b_l u(n-l), \quad b_0 = 1, \qquad (2.1)$$

Où $\{a_k\}$, $\{b_l\}$, et le gain G sont les paramètres du système. L'équation précédente prédit la sortie courante en utilisant une combinaison linéaire des sorties antérieures et les entrées courantes et antérieures.

Dans le domaine fréquentiel, la fonction de transfert du modèle de prédiction linéaire de la parole est de la forme :

$$H(z) = \frac{B(z)}{A(z)} = \frac{G[1 + \sum_{l=1}^{q} b_l z^{-l}]}{1 - \sum_{k=1}^{p} a_k z^{-k}} \qquad (2.2)$$

H(z) est aussi appelée modèle pôle-zéro dans lequel les racines du dénominateur et du numérateur sont, respectivement, les pôles et les zéros du système.

Si $a_k = 0$ pour $1 \le k \le p$, $H(z)$ devient un modèle tout zéro ou modèle à moyenne ajustée (MA). Si $b_l = 0$ pour $1 \le l \le q$, $H(z)$ se réduit à un modèle tout pôle ou modèle Auto Régressif (AR) :

$$H(z) = \frac{1}{A(z)} \qquad (2.3)$$

Dans l'analyse de la parole, les classes de phonèmes comme les fricatives et les nasales contiennent des vallées spectrales qui correspondent aux zéros dans H(z). Par contre les voyelles contiennent des résonances qui peuvent être modélisées par le modèle tout pôle. Pour des raisons de simplicité, le modèle tout pôle est préféré pour l'analyse par prédiction linéaire de la parole.

Ainsi, le signal prédit est égal à :

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) \qquad (2.4)$$

et l'erreur de prédiction ou résiduel du signal est la sortie $e(n)$:

$$e(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \qquad (2.5)$$

L'ordre *p* du système est choisi de façon que l'estimation de l'enveloppe spectrale soit adéquate. Une façon de procéder est d'allouer une paire de pôles pour chaque formant présent dans le spectre. On ajoute 2 ou 3 pôles pour approximer les zéros dus aux sons non voisés.

Quand la prédiction linéaire est basée sur les échantillons de parole passés s(n), celle-ci, est dite Prédiction Linéaire Adaptative Progressive (Forward), dans laquelle les coefficients de prédiction doivent être transmis au décodeur en tant qu'une information latérale. Si la prédiction linéaire est basée sur les échantillons de parole reconstruits antérieurs s(n), celle-ci, est dite Prédiction Linéaire Adaptative Régressive (Backward). Pour calculer les coefficients du filtre à court-terme $\{a_i\}$ du processus AR, la méthode classique des moindres carrés peut être utilisée. La variance ou l'énergie du signal d'erreur e(n) est minimisée sur une trame de parole. Deux grandes approches sont utilisées pour l'analyse LP (Linear Prediction) à court-terme : la méthode d'autocorrélation et la méthode de covariance **[20]**.

2.3 Estimation des coefficients LP

2.3.1 Méthode d'Autocorrélation

La méthode d'Autocorrélation garantit la stabilité du filtre LP. Une fenêtre d'analyse w(n) de longueur finie est d'abord multipliée par le signal parole S(n) pour obtenir un segment de parole fenêtré $S_w(n)$:

$$S_w(n) = W(n) \cdot S(n) \qquad (2.6)$$

durant cette durée finie, le signal parole est considéré comme stationnaire. Plusieurs fenêtres d'analyse de formes différentes sont proposées. La plus simple est la fenêtre rectangulaire à N échantillons :

$$W(n) = \begin{cases} 1, & 0 \le n \le N-1, \\ 0, & \text{ailleurs.} \end{cases} \qquad (2.7)$$

La fenêtre rectangulaire possède un lobe principal le plus étroit, mais présente des lobes fréquentiels importants. Une fenêtre d'analyse allongée permet de réduire l'effet des composants en dehors de la fenêtre en minimisant l'erreur de prédiction pour les premières et dernières valeurs de S(n) pour la fenêtre courante. La fenêtre de Hamming qui est une fenêtre cosinusoidale modifiée, est souvent employée comme fenêtre d'analyse. Voir **Fig.2.1**.

$$W(n) = \begin{cases} 0.54 - 0.46\cos\left(\dfrac{2\pi n}{N-1}\right), & \text{pour } 0 \le n \le N-1, \\ 0, & \text{ailleurs.} \end{cases} \qquad (2.8)$$

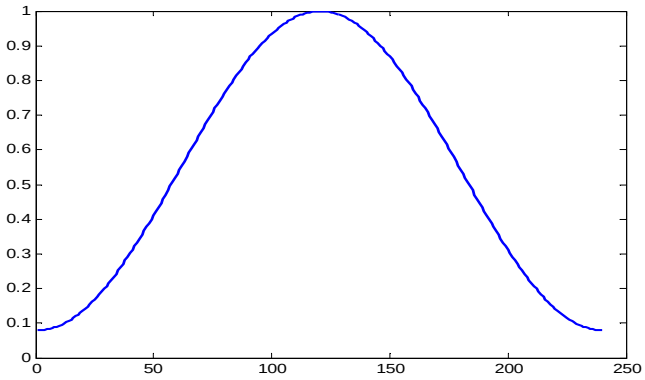


Fig 2.1 Fenêtre de Hamming à N=240 échantillons.

Les bords lisses de la fenêtre permettent un décalage périodique de la fenêtre sans affecter les paramètres spectraux de la parole. La fenêtre de Hamming peut être une fenêtre hybride comme celle utilisée dans le standard G.729 **[18].**

Après la multiplication du signal de parole par la fenêtre d'analyse, l'autocorrélation du segment de parole fenêtré est calculée. La fonction d'autocorrélation du signal fenêtré $S_w(n)$ est donnée par :

$$R(i) = \sum_{n=i}^{N-1} s_w(n)s_w(n-i) \qquad 1 \le i \le p\,, \qquad (2.9)$$

la fonction d'autocorrélation est une fonction paire avec R(i)= R(-i). Pour trouver les coefficients a_k du filtre LPC, l'énergie du résiduel de prédiction sur l'intervalle fini $0 \le n \le N-1$ définie par la fenêtre d'analyse W_n doit être minimisée :

$$E = \sum_{n=-\infty}^{\infty} e^2(n) = \sum_{n=-\infty}^{\infty} (s_w(n) - \sum_{k=1}^{p} a_k s_w(n-k))^2 \qquad (2.10)$$

en annulant les dérivations partielles de l'énergie par rapport aux coefficients du filtre a_k :

$$\frac{\partial E}{\partial a_k} = 0 \qquad 1 \le i \le p \qquad (2.11)$$

on obtient p équations linéaires avec "p" coefficients inconnus a_k :

$$\sum_{k=1}^{p} a_k \sum_{n=-\infty}^{\infty} s_w(n-i)s_w(n-k) = \sum_{n=-\infty}^{\infty} s_w(n-i)s_w(n) \qquad 1 \le i \le p \qquad (2.12)$$

Les équations linéaires peuvent être écrites sous la forme :

$$\sum_{k=1}^{p} R(|i-k|)a_k = R(i) \qquad 1 \le i \le p \qquad (2.13)$$

Dans la forme matricielle, l'ensemble des équations linéaires est représenté par $R \cdot a = v$ qui peut être réécrite comme suit:

$$\begin{pmatrix} R(0) & R(1) & R(2) & \cdots & R(p-1) \\ R(1) & R(0) & R(1) & \cdots & R(p-2) \\ R(2) & R(1) & R(0) & \cdots & R(p-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R(p-1) & R(p-2) & R(p-3) & \cdots & R(0) \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} R(1) \\ R(2) \\ R(3) \\ \vdots \\ R(p) \end{pmatrix} \qquad (2.14)$$

La matrice d'autocorrélation $p \times p$ obtenue est une matrice de Toeplitz. L'algorithme de Levinson-Durbin est utilisé pour trouver les coefficients de prédiction minimisant la moyenne quadratique de l'erreur de prédiction.

2.3.2 Méthode de covariance

La méthode d'autocorrélation et de covariance diffèrent dans l'emplacement de la fenêtre d'analyse. Dans la méthode de covariance, le signal d'erreur est fenêtré au lieu du signal parole de telle façon que l'énergie à minimiser soit :

$$E = \sum_{n=-\infty}^{\infty} {e_w}^2(n) = \sum_{n=-\infty}^{\infty} e^2(n)w^2(n) \qquad (2.15)$$

en annulant les dérivations partielles par rapport aux coefficients du filtre $\delta E / \delta a_k = 0$ pour $1 \le \mathrm{k} \le \mathrm{p}$, on aura "$p$" équations linéaires.

$$\sum_{k=1}^{p} \Phi(i,k)a_k = \Phi(i,0) \qquad 1 \le i \le p \qquad (2.16)$$

où la fonction de covariance $\Phi(i,k)$ est définie par:

$$\Phi(i,k) = \sum_{n=-\infty}^{\infty} w^2(n)s(n-1)s(n-k). \quad (2.17)$$

sous la forme matricielle, les p équations deviennent

$$\Phi a = \Psi, \quad (2.18)$$

où

$$\begin{pmatrix} \phi(1,1) & \phi(1,2) & \phi(1,3) & \cdots & \phi(1,p) \\ \phi(2,1) & \phi(2,2) & \phi(2,3) & \cdots & \phi(2,p) \\ \phi(3,1) & \phi(3,2) & \phi(3,3) & \cdots & \phi(3,p) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi(p,1) & \phi(p,2) & \phi(p,3) & \cdots & \phi(p,p) \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} \Psi(1) \\ \Psi(2) \\ \Psi(3) \\ \vdots \\ \Psi(p) \end{pmatrix} \quad (2.19)$$

et $\Psi(i) = \Phi(i.0)$ pour $1 \le i \le p$. (2.20)

La matrice Φ n'est pas une matrice de Toeplitz, elle est symétrique et définie positive. La matrice de covariance peut être décomposée en une matrice triangulaire supérieure et une autre inférieure :

$$\Phi = LU \quad (2.21)$$

la décomposition de Cholesky est utilisée pour convertir la matrice de covariance en:

$$\Phi = CC^T \quad (2.22)$$

où C= L et C^T= U. le vecteur **a** est trouvé en résolvant d'abord l'équation:

$$Ly = \Psi \quad (2.23)$$

puis :

$$Ua = y \quad (2.24)$$

2.4 Expansion de la bande passante

L'analyse LP peut générer des filtres de synthèse ayant des pics spectraux aigus ; pour y remédier, on emploi une expansion de la bande passante, elle consiste à élargir la bande passante des pics des formants de la réponse fréquentielle. Les racines du filtre tout pôle sont multipliées par des facteurs d'expansion de bande γ, résultant dans le filtre

$$H'(z) = \frac{1}{A'(z)} = \frac{1}{A(\gamma z)} \qquad (2.25)$$

Les coefficients de prédiction pondérés seront alors:

$$a'_k = a_k \gamma^k, 1 \le \mathrm{k} \le \mathrm{p}. \qquad (2.26)$$

le facteur d'expansion γ pour une expansion de f_b Hz est calculé par :

$$\gamma = e^{\frac{-f_b \pi}{f_s}} \qquad (2.27)$$

f_s étant la fréquence d'échantillonnage.

Par exemple, un facteur γ = 0.996 donne une expansion de 10 Hz pour une fréquence d'échantillonnage de 8kHz. Pour l'analyse de la parole une expansion de 10 à 25 Hz est souvent réalisée.

2.5 Représentation des paramètres de la parole

Les coefficients $\{a_k\}$ présentent quelques inconvénients lors de la quantification. En effet, une faible erreur de quantification de l'un de ces paramètres entraîne de fortes variations dans le spectre restitué par l'ensemble du filtre et génère souvent des problèmes d'instabilité au filtre de synthèse. Par conséquent, un nombre de représentations des coefficients LP été considéré pour essayer de trouver la représentation qui minimise ces limitations. Les représentations les plus utilisées sont les coefficients de réflexion, les LAR (log-area ratios) et les LSPs (Line Spectrum Pairs).

2.5.1 Les coefficients de réflexion

Ont peut obtenir les coefficients LP à partir des coefficients de réflexion $\{k_m\}$. Initialement, on calcul l'énergie moyenne E_0 dans la trame de parole tel que E_0=R(0). Ensuite en résout les équations suivantes pour chaque itération m, avec m=1,2,…p.

$$k_m = \frac{1}{E_{m-1}} \left[R(m) - \sum_{k=1}^{m-1} \alpha_{m-1}(k) R(m-k) \right] \qquad (2.28)$$

$$\alpha_k(m) = \alpha_k(m-1) - k_m \alpha_{m-k}(m-1), \; 1 \le \mathrm{k} \le \mathrm{m}\text{-}1 \qquad (2.29)$$

$$E_m = (1 - k_m^2) E_{m-1}. \qquad (2.30)$$

Les coefficients $\alpha_k(m)$ représentent les coefficients de prédiction d'un predicteur linéaire d'ordre m:

$$a_k = \alpha_k(m), \quad 1 \le k \le m. \qquad (2.31)$$

Puisque E_m, une erreur quadratique, n'est jamais négative, $|k_m| \prec 1$. Cette condition sur les coefficients de réflexion garantit aussi la stabilité du filtre de synthèse LP. Les valeurs négatives des coefficients de réflexion sont appelées les corrélations partielles (Partial Correlation) ou coefficients PARCOR. On peut trouver les coefficients de réflexion à partir des coefficients LP $a_k = \alpha_p(k)$, en calculant de manière récursive les deux équations suivantes pour m=p,p-1,…,2 :

$$\alpha_{m-1}(i) = \frac{\alpha_m(i) k_m \alpha_m(m-i)}{1-k_m^2}, \quad 1 \le i \le m - i \qquad (2.32)$$

$$\text{avec} \qquad k_{m-i} = \alpha_{m-i}(m-1) \qquad (2.33)$$

Une propriété importante des coefficients de réflexion, est qu'on peut juger la stabilité du filtre à partir des valeurs k_m, en effet, si $|k_m| < 1$, le filtre de synthèse est stable, tandis que Si $|k_m| \ge 1$ alors le spectre du signal est altéré, mais les sorties instables sont éliminées.

L'inconvénient des coefficients de réflexion est leur grande sensibilité spectrale pour des amplitudes proches de l'unité. Cependant, cet inconvénient peut être contourné par les transformations non linéaires qui élargissent la région au voisinage de la valeur $|k_m| = 1$. la transformation en coefficients LARs, en est une parmi ces solutions.
Les coefficients LAR sont calculés comme suit :

$$g_m = \log\left(\frac{1+k_m}{1-k_m}\right), \quad 1 \le m \le p. \qquad (2.34)$$

pour revenir aux coefficients de réflexions :

$$k_m = \frac{e^{g_m} - 1}{e^{g_m} + 1}, \quad 1 \le m \le p. \qquad (2.35)$$

2.5.2 Les fréquences de raies spectrales (LSF's)

Les paramètres LSF (Line Spectral Frequencies), appelés aussi paramètres LSP (Line Spectral Pair), sont liés à la position des pôles sur l'axe des fréquences et ont la propriété d'être rangés par ordre croissant, cet agencement des paramètres permet de prendre en compte des critères perceptifs et offre une propriété de codage efficace avec un meilleur contrôle de la stabilité. Les fréquences de raies spectrales ont été introduites pour la première fois par Itakura comme une alternative aux coefficients de prédiction linéaire **[10]** . Les LSPs sont les solutions des deux équations suivantes :

$$P(z) = A(z) + z^{-(p+1)} A(z^{-1}) \qquad (2.36)$$

$$Q(z) = A(z) - z^{-(p+1)} A(z^{-1}) \qquad (2.37)$$

avec

$$A(z) = \frac{1}{2}[P(z) + Q(z)] \qquad (2.38)$$

$$\text{et} \quad A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k} \ \text{ pour modèle AR} \qquad (2.39)$$

où les coefficients de réflexion d'ordre p+1 sont mis à 1 pour Q(z) et à -1 pour P(z).

Il a été démontré que si H(z) est stable, où A(z) est à phase minimale, alors les racines de P(z)et Q(z) se trouvent sur le cercle unité et sont alternées entre les deux polynômes lorsque w augmente. Les LSPs correspondent à des positions angulaires. Les racines apparaissent sous forme de paires conjuguées et par conséquent, il existe p LSPs positionnés entre 0 et π. Il a été démontré que si les LSPs, notés par w_i, sont dans un ordre ascendant et uniques, alors le filtre inverse A(z) correspondant est à phase minimale, ce qui garantit la stabilité.

$$0 \prec w_1 \prec w_2 \prec \ldots \prec w_p \prec \pi \ \text{[radians/sec]} \qquad (2.40)$$

le modèle des LSFs correspond au spectre du filtre LP. Les LSFs se regroupent autour des pics spectraux (voir figure **Fig 2.2**). Un changement d'un LSF quelconque ne peut altérer le spectre que sur la partie avoisinant cet LSF. Les LSFs peuvent être calculés par plusieurs méthodes, Soong et Juang calculent les LSFs par application d'une transformation en cosinus discret des coefficients des polynômes

$$G(z) = \begin{cases} \dfrac{P(z)}{1 + z^{-1}}, & p \quad pair, \\ P(z), & p \quad impair. \end{cases} \qquad (2.41)$$

et

$$H(z) = \begin{cases} \dfrac{Q(z)}{1 - z^{-1}}, & p \quad pair, \\ \dfrac{Q(z)}{1 - z^{-2}}, & p \quad impair. \end{cases} \qquad (2.42)$$

Kabal et Ramachandran ont utilisé les polynômes de Chebyshev pour calculer les LSPs. Ils sont disposés sur un cercle unité du plan z et entrelacés entre valeurs paires et impaires.

$$T_m(x) = \cos(mw) \qquad (2.43)$$

où $x = cox(w)$ fait correspondre l'image du demi-cercle du plan z à l'intervalle de valeurs réelles [-1, 1]. Les polynômes $G'(w)$ et $H'(w)$ peuvent être exprimés par

$$G'(x) = 2\sum_{i=0}^{l} g_i T_{l-i}(x), \qquad (2.44)$$

$$H'(x) = 2\sum_{i=0}^{m} h_i T_{m-i}(x), \qquad (2.45)$$

où

$$\begin{array}{l} l = m = p/2 \qquad\qquad\quad \text{pour} \quad p \quad pair \\ l = (p+1)/2 \quad \text{et} \quad m = (p-1)/2 \quad \text{pour} \quad p \quad impair \end{array} \qquad (2.46)$$

les racines de G' et H' sont déterminées itérativement en cherchant les changements de signe dans l'intervalle [-1, 1] . Les LSFs correspondent aux racines du polynôme en utilisant la transformation

$$w = \cos^{-1}(x) \qquad (2.47)$$

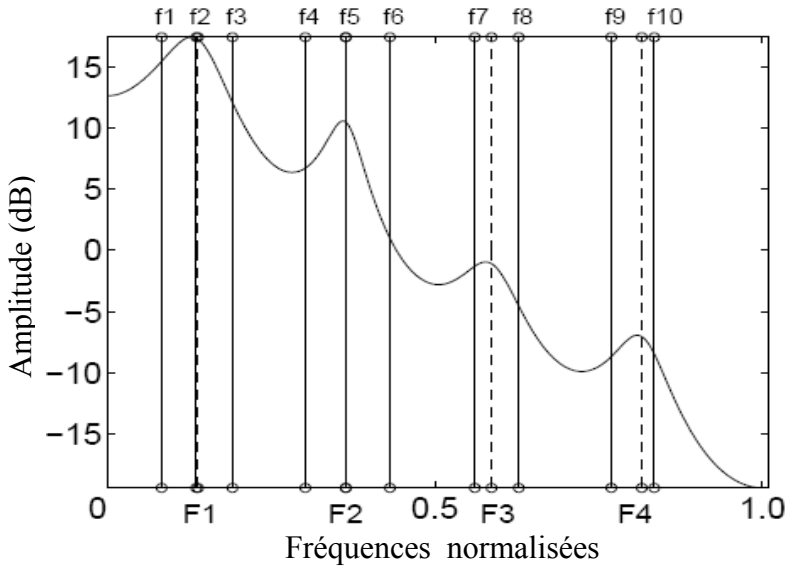


Fig. 2.2 Spectre LP avec les positionnements des LSFs
(f1,f2,…f10 représentent les LSFs et F1,…,F4 représentent les formants).

2.6 Evaluation de la Qualité de la parole

Après avoir effectué la conception d'un algorithme de codage, il est nécessaire de mettre ce dernier sous tests afin de l'évaluer et de voir s'il répond aux normes et aux critères de codage. Les algorithmes de codage de la parole sont évalués selon plusieurs critères (voir paragraphe : Critères de performances dans le chapitre 1) dont les plus importants sont : la qualité du signal, le débit binaire, la complexité de l'algorithme et le retard de communications, nous consacrons cette partie au premier critère qui est la qualité du signal.

Dans les communications numériques, la qualité du signal parole est évaluée selon quatre catégories :

- **Qualité diffusion ou broadcast :** qui se réfère aux larges bandes (typique 50-7000 Hz et 20-20000 Hz pour disques compacts) c'est la plus haute qualité qu'on peut atteindre, elle nécessite des débits au moins de 32 à 64 kbps.
- **Qualité réseau ou toll :** c'est la qualité qui permet d'entendre la parole sur un réseau téléphonique (pour une bande de 200-3200 Hz avec un rapport signal sur bruit de 30 dB et une distorsion moins de 2 à 3 %).
- **Qualité de communications :** elle implique une certaine dégradation de la qualité de la parole, néanmoins, elle présente une qualité naturelle et hautement intelligible. Cette qualité peut être atteinte à des débits supérieurs à 4 kbps.

- **Qualité synthétique :** la parole synthétique est intelligible, néanmoins, elle n'est pas naturelle et perd la reconnaissabilité de locuteur.

Le but actuel dans le codage de la parole est d'atteindre la qualité *toll* pour des débits de 4 kbps. Actuellement, les codeurs opérant en dessous de 4 kbps de débit, fournissent une qualité synthétique. La mesure de qualité est une tache importante mais très difficile. Il y a deux manières pour mesurer la qualité de la parole, on distingue la mesure subjective et la mesure objective.

2.6.1 Mesure subjective de la qualité de la parole

La procédure d'évaluation subjective est achevée par des tests d'écoute de l'ensemble des syllabes, mots ou phrases. Le test est souvent concentré sur les consonnes car elles sont plus difficiles à synthétiser que les voyelles. Dans ces tests la qualité est mesurée par l'intelligibilité qui est définie par un pourcentage de mots ou phonèmes correctement écoutés, et avec une sonorité naturelle. Il existe trois types de mesures subjectives de la qualité :

Test diagnostique de rime (DRT)

Il s'agit d'une mesure d'intelligibilité dont la tache est de reconnaître un ou deux mots possibles parmi un ensemble de paires de rimes, par exemple (meat - heat) .

Mesure diagnostique d'acceptabilité (DAM)

Elle sert pour l'évaluation des systèmes de communications, elle est basée sur l'acceptabilité de la parole par des auditeurs normatifs qualifiés.

Le test MOS

Ce test est largement utilisé pour évaluer la qualité de la parole. Le MOS nécessite 12 à 14 auditeurs (pour le CCITT et TIA les tests nécessitent 32 à 64 auditeur) qui sont entraînés pour évaluer phonétiquement la qualité selon une échelle de cinq (5) niveaux, voir **Tableau 2.1** et figure **Fig 2.3**.

La valeur du MOS est calculée par

$$MOS = \frac{\sum_i N_i i}{N} \qquad (2.48)$$

avec : i=1,…,5

N : nombre d'auditeurs ayant participe au test.

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

Valeur MOS	Qualité de la parole	Niveau de distorsion
5	Excellent	Imperceptible
4	Bon	Juste perceptible mai pas gênant
3	Assez bon	Perceptible légèrement gênant
2	Médiocre	Gênant mais pas désagréable
1	Mauvais	Très gênant et désagréable

Tableau 2.1 Description du test MOS.

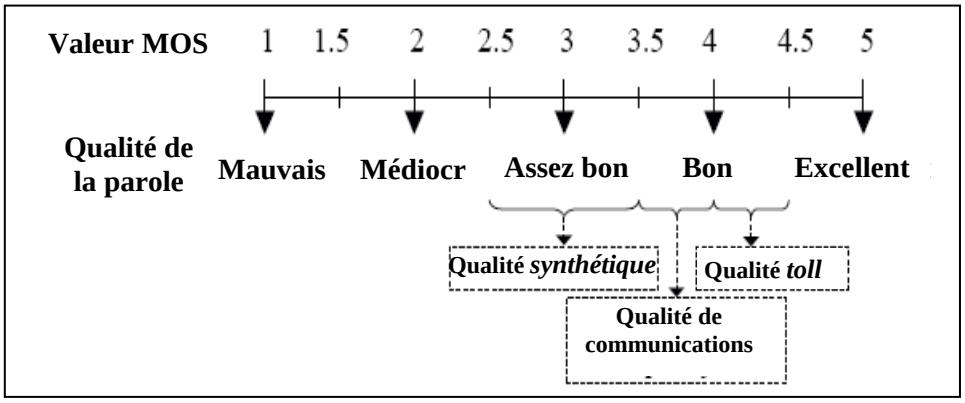


Fig 2.3 Relation entre les valeurs MOS et la qualité de la parole.

2.6.2 Mesure objective de la qualité de la parole

Le système auditif humain est l'évaluateur 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. Malgré que les tests d'écoute subjectifs donnent une bonne évaluation des codeurs de la parole, ils exigent beaucoup de temps et sont inconsistants. Les mesures objectives peuvent donner une évaluation immédiate et efficace de la qualité d'un algorithme de codage. Les mesures objectives de distorsion peuvent être calculées aussi bien dans le domaine temporel (calcul du rapport signal sur bruit) que fréquentiel (mesure de distorsions).

2.6.2.1 Mesures Objectives dans le Domaine Temporel

Les mesures objectives les plus importantes dans le domaine temporel sont les suivantes:

Le rapport signal sur bruit SNR (Signal to Noise Ratio)

C'est la mesure objective de la qualité la plus commune pour l'évaluation des performances des algorithmes de compression. Le SNR est défini comme un rapport de l'énergie moyenne du signal parole sur l'énergie moyenne du signal d'erreur, le SNR est généralement exprimé en décibel *dB* et défini par :

$$\mathrm{SNR} = 10\log_{10}\left(\frac{\textit{Energie}\ \text{moyenne du signal parole}}{\textit{Energie}\ \text{moyenne signal d'erreur}}\right)\mathrm{dB} = 10\log_{10}\frac{\sum_{n=-\infty}^{\infty} s^{2[}[n]}{\sum_{n=-\infty}^{\infty}(s[n]-\hat{s}[n])^2}\,dB \quad (2.49)$$

où ŝ[n] est la version codée du signal parole original s[n]. La mesure SNR n'est par une estimation exacte de la qualité, en effet, le SNR ne donne qu'une seule évaluation pendant toute la durée du signal, on traite le signal parole en tant qu'un seul vecteur, alors qu'en réalité, l'auditeur effectue plusieurs comparaisons pour un signal parole donné. C'est pourquoi on préfère utiliser le SNR segmental.

Le SNR Segmental (SNRseg)

Les variations temporelles de performance peuvent être mieux détectées et évaluées en utilisant un rapport signal sur bruit à court terme (trame par trame), cette mesure s'appelle SNR segmental (SNRseg). Pour chaque trame (typiquement de 15 à 25 msec), on mesure le SNR et la mesure finale sera la moyenne des mesures pour tous les segments du signal. La mesure SNRseg, en dB sur M segments est définie par :

$$\mathrm{SNRseg} = \frac{1}{M}\sum_{j=0}^{M-1} 10\log_{10}\left[\frac{\sum_{n=1}^{N} s^2(n+Nm)}{\sum_{n=1}^{N}(s[n+Nm]-\hat{s}[n+Nm])^2}\right]\mathrm{dB} \quad (2.50)$$

où chaque segment m est de longueur de N échantillons. Pour un signal de parole échantillonné à 8 kHz, la valeur typique de N est entre 100 et 200 échantillons (15-25 msec).

2.6.2.2 Mesures Objectives de la qualité dans le Domaine Fréquentiel

La différence entre l'enveloppe spectre du signal parole original et celle du signal codé, qui peut être traduite par une différence entre les fréquences des formants ou entre leurs largeur, conduit à des sons phonétiquement différents. c'est pourquoi on fait recours à la distorsion spectrale. Une brève description des différentes mesures de distorsion dans le domaine fréquentiel est présentée dans ce qui suit :

Distorsion d'Itakura-Saito

La distorsion d'Itakura-Saito, connue sous le nom de mesure de distance du rapport de vraisemblance, mesure le rapport d'énergie entre le signal résiduel obtenu en utilisant le filtre LP avec les coefficients quantifiés et le signal résiduel obtenu en utilisant le filtre LP avec les coefficients non quantifiés. La distance d'Itakura-Saito et donnée par :

$$d_{IS} = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[e^{V(\omega)} - V(\omega) - 1\right]d\omega \qquad (2.51)$$

avec : $$V(\omega) = \log s(\omega) - \log \hat{s}(\omega) \qquad (2.52)$$

Distorsion spectrale Logarithmique

La mesure de distorsion spectrale Logarithmique est la plus fréquemment utilisée, appelée souvent *distorsion spectrale*, Elle est exprimée par :

$$d_{SD}^{p} = \frac{2}{2\pi}\int_{-\pi}^{\pi}\left|10\log_{10} S(\omega) - 10\log_{10}\hat{S}(\omega)\right|^{p} d\omega \qquad (2.53)$$

où $$S(\omega) = \frac{G}{\left|A(e^{jw})\right|^2} = \frac{G}{[1-\sum_{n=1}^{p} a_n e^{jn\omega}]^2} \qquad (2.54)$$

G est le facteur de gain du filtre LP et $\{a_n\}$ sont les coefficients du filtre.

Lorsque p=2 la distorsion spectrale est appelée RMS (pour Root Mean Square), elle sera donc définie par :

$$d_{SD} = \sqrt{\frac{1}{\omega_u - \omega_l}\int_{\omega_l}^{\omega_u}\left[10\log_{10}\frac{S(\omega)}{\hat{S}(\omega)}\right]^2 d\omega}\ \ dB \qquad (2.55)$$

avec ω_l et ω_u sont des fréquences qui représentent la limite basse et la limite haute de l'intégrale.

Distance euclidienne pondérée

Les LSFs possèdent une relation directe avec la forme de l'enveloppe spectrale. Les formants correspondent aux LSFs voisins (étroitement liés) tandis que les LSFs isolés représentent les vallées. Par conséquent, une distance du carré de l'erreur peut être utilisée pour comparer les vecteurs LSFs originaux et les vecteurs LSFs codés.

Soient deux vecteurs LSFs à m dimensions x et $\hat{x}$, la distance euclidienne est donnée par :

$$d(x,\hat{x}) = (x - \hat{x})^T (x - \hat{x}) = \|x - \hat{x}\|^2 \qquad (2.56)$$

Afin d'obtenir une bonne estimation de la qualité perceptuelle de l'enveloppe spectrale, on préfère utiliser une distance euclidienne pondérée des LSFs :

$$d(x,\hat{x}) = (x - \hat{x})^T W(x - \hat{x}) = \|x - \hat{x}\|^2 \qquad (2.57)$$

où W est une matrice de pondération symétrique et positive de dimensions m x m (m est l'ordre de l'analyse LP). Si W est une matrice diagonale ayant les éléments $w_{ii} > 0$, la distance sera alors :

$$d(x,\hat{x}) = \sum_{i=1}^{m} w_{ii}(x_i - \hat{x}_i)^2 \qquad (2.58)$$

2.6.3 Mesure objective perceptuelle

Les mesures objectives de distorsion sont souvent sensibles aux variations du gain et du délai, en plus elles ne prennent pas en considération les propriétés perceptuelles de l'oreille. D'autre part, les mesures subjectives sont lentes et coûteuses. Les efforts actuels relatifs à la qualité de la parole sont concentrés pour le développement de procédures automatiques d'évaluations et de mesures objectives qui seront capables de prédire la qualité subjective de la parole. On distingue plusieurs mesures objectives perceptuelles, nous pouvons citer : PSQM (pour Perceptual Speech Quality Mesure), NMB, la recommandation ITU-T P.861 PESQ, BSD et EMBSD. Nous détaillons dans ce qui suit les deux dernières mesures.

Distorsions EMBSD et BSD (Enhanced Modified Bark Spectral Distorsion)

La mesure BSD suppose que la qualité de la parole est directement liée à l'intensité (loudness) de la parole qui est définie par une sensation perçue pour une fréquence et à un niveau de pression donné **[1]**. La BSD est définie comme la distance Euclidienne moyenne entre l'intensité estimée $L_T^j(i)$ du signal test T(n) et l'intensité estimée $L_R^j(i)$ du signal de référence

R(n) où i est l'index de la bande critique et j est l'index de la trame, N est le nombre de trames et K représente le nombre des bandes critiques. La BSD est donnée par :

$$BSD = \frac{\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{K}\left[L_R^j(i) - L_T^j(i)\right]^2}{\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{K}\left[L_R^j(i)\right]^2} \qquad (2.59)$$

La MBSD se calcule par :

$$MBSD = \frac{1}{N}\sum_{j=1}^{N}\left[\sum_{i=1}^{K} M(i)\left|L_R^j(i) - L_T^j(i)\right|\right] \qquad (2.60)$$

où M(i) est un indicateur binaire avec $M(i) = \begin{cases} 0 \text{ si distorsion} & \text{imperceptible} \\ 1 \text{ si distorsion} & \text{perceptible} \end{cases}$

l'EMBSD est obtenue à partie de la MBSD par une introduction d'un nouveau modèle de connaissance basé sur les effets de poste masquage et par l'utilisation de 15 valeurs d'intensité. Les vecteurs d'intensité sont normalisés et les fonctions de propagation dans le calcul du seuil de masquage de bruit sont supprimées.

2.7 Conclusion

La prédiction linéaire exploite la redondance dans le signal parole et extrait les coefficients (paramètres LP) qui caractérisent le comportement du signal. La simplicité de son concept, la linéarité dans la résolution des systèmes et ses performances dans le codage de la parole, la rendent la plus admise et la plus largement utilisée dans le codage du signal de parole.

Chapitre 3

Transmission de la voix à travers les réseaux IP (VoIP)

3.1 Introduction

La voix sur IP constitue actuellement l'évolution la plus importante du domaine des télécommunications. Avant 1970, la transmission de la voix s'effectuait de façon analogique sur des réseaux dédiés à la téléphonie. La technologie utilisée était la technologie électromécanique (Crossbar). Dans les années 80, une première évolution majeure a été le passage à la transmission numérique (TDM).La transmission de la voix sur les réseaux informatiques à commutation de paquets IP constitue aujourd'hui une nouvelle évolution majeure comparable aux précédentes.

La Voix sur IP correspond à l'ensemble des technologies permettant la transmission d'une conversation vocale sur un réseau au protocole IP. Les réseaux IP présentent la caractéristique de transporter les données sous forme de paquets. Avec la technologie VoIP, la voix est digitalisée, compressée et envoyée sous forme de paquets sur un réseau fonctionnant avec le protocole IP. Les données reçues sont décompressées et converties en voix audible.

3.2 Avantages de la VoIP

La VoIP offre plusieurs avantages par rapport aux réseaux (PSTN) tels que :

- L'intégration voix et données en un seul réseau.
- Optimisation et efficacité d'exploitation de la bande passante.
- Diminution des tarifs et des coûts des communications.
- Facilité d'administration, de supervision et de maintenance des réseaux IP.
- Augmentation et amélioration des services... etc.

3.3 Architecture TCP/IP

La VoIP utilise comme plateforme l'architecture TCP/IP, cette architecture ne nécessite que quatre couches du modèle OSI au lieu des sept couches voir **Fig 3.1.** Chaque couche est responsable d'une tache dans la communication. Ces couches sont décrites ci-dessous **[21] [22]**.

1. Couche liaison de données : comprend le module de gestion de périphériques dans le système d'exploitation, ainsi que la carte d'interface réseau sur le PC. la couche liaison gère tout le matériel qui est physiquement connecté au câble réseau. Ethernet et Token Ring sont des exemples de la couche liaison.
2. Couche réseau : aussi appelée couche Internet, gère le mouvement et le trafic des paquets de données entre les différents hôtes du réseau. Le protocole Internet (IP) est le protocole utilisé dans la couche réseau dans l'architecture TCP/IP.
3. Couche transport : gère et fournit le flux de données entre les deux hôtes d'extrémité pour la couche application. Il y a deux protocoles de transport dans le modèle TCP/IP, le TCP (Transmission Control Protocol) et l'UDP (User Datagram Protocol).
4. Couche application : gère les détails d'une application particulière. Pour des applications de multimédia telles que VoIP, le RTP (Real Time Protocol) est souvent employé. Deux fonctions de base sont incluses dans l'entête RTP, le nombre de séquences et le marqueur de temps (timestamp). Le nombre de séquences est utilisé pour détecter les paquets perdus ou qui sont en désordre, tandis que le marqueur de temps est utilisé pour lire les trames de la parole dans des intervalles appropriés. RTP est typiquement employé en dessus de l'UDP**[23]**.

Chaque couche dans l'architecture TCP/IP ajoute une entête au paquet final. L'entête minimale de l'UDP et TCP est 12 et 20 octets respectivement. L'entête RTP est au moins 12 octets et l'entête de l'IP v4 est de 20 octets.

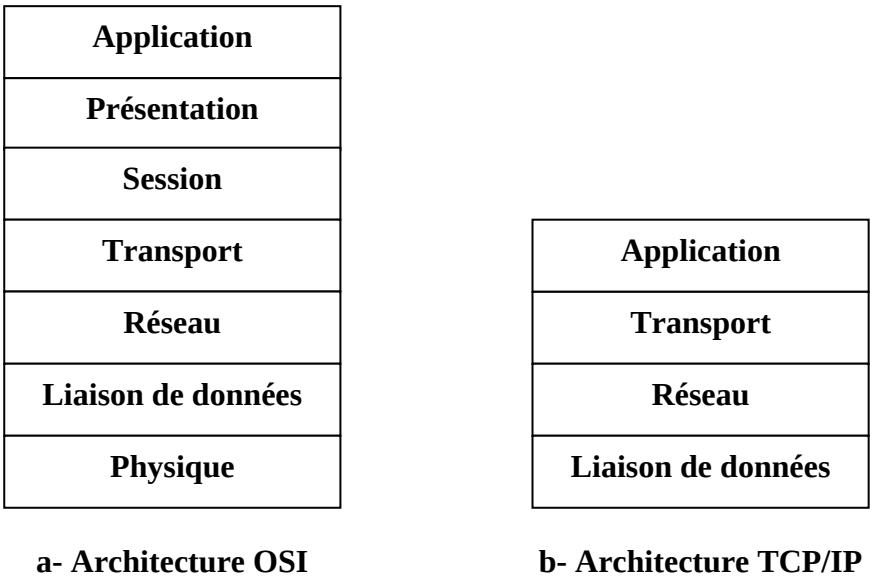


Fig 3.1 Architecture OSI et TCP/IP.

3.4 Systèmes de transmission de la VoIP

La transmission en temps réel de la voix d'un point à un autre dans un système VoIP consiste en plusieurs étapes **[1] [24]**. Au niveau de l'émetteur, le signal de parole analogique est périodiquement numérisé puis codé. Le signal numérique ainsi obtenu est ensuite traité afin d'annuler l'écho créé par la ligne. Un détecteur d'activité vocale (VAD : Voice Activity Detector) est utilisé afin de localiser les périodes de silence. Durant le silence, l'émetteur peut faire deux choses, soit ne pas transmettre les paquets, soit réduire le débit. Le signal parole est compressé et codé sous forme de trames en utilisant des CoDec de parole. Les trames de paroles sont ensuite mises sur des paquets pour le transport sur le réseau. Un paquet (entête) RTP **[25]** est créé par l'ajout de 12 octets à la trame de voix. Le paquet RTP est ensuite encapsulé dans un paquet UDP au niveau de la couche transport, puis dans un paquet IP au niveau de la couche réseau. Le paquet IP est ensuite envoyé sur Internet où il sera routé vers sa destination **[26].**

Comme les paquets peuvent être perdus ou retardés à travers le réseau, plusieurs solutions ont été proposées, l'une d'elles consiste à utiliser, au niveau du récepteur, des buffers dits *buffer de playout* afin de supprimer le retard de la *gigue* et de mémoriser les paquets jusqu'au moment approprié.

Le récepteur extrait la trame de parole du paquet contenant les entêtes IP, UDP et RTP, la décompresse, la décode puis la convertit de nouveau en un signal analogique pour la lire par la suite. Les paquets qui ne sont pas arrivés à destination ou qui arrivent en retard sont considérés comme perdus. Dans ce cas, des algorithmes de dissimulations (PLC) sont employés pour compenser les paquets perdus. La figure **Fig 3.2** illustre le système de transmission VoIP [**27].**

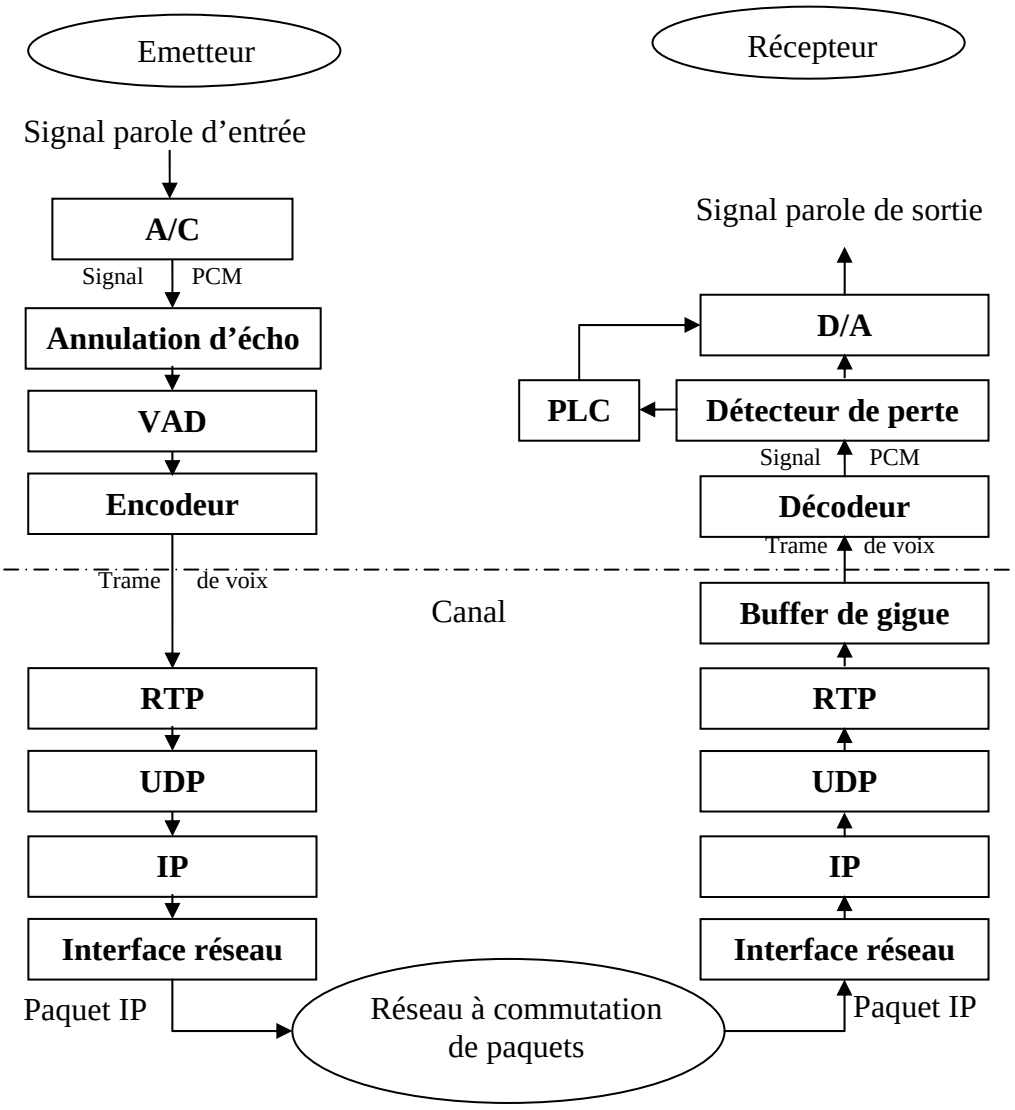


Fig 3.2 Schéma d'un système de transmission de la VoIP.

3.5 Standards et Protocoles dédiés à la VoIP

Plusieurs standards et protocoles ont été réalisés pour gérer et normaliser la VoIP. l'UIT (Union Internationale des Télécommunications) et l'IETF (Internet Engineering Task Force) ont élaboré des familles de standards regroupés sous les appellations génériques H.32x, SIP (Session Initiation Protocol) et MGCP (Media Gateway Control Protocol). Nous détaillons dans ce qui suit chaque protocole à part **[28][29].**

3.5.1 Le standard H.323

Avec H.323, l'UIT a spécifié un environnement complet de protocoles de communications multimédias pour les réseaux IP. L'inter fonctionnement avec les autres réseaux est garanti, car des standards apparentés ont été conçus: H.320 pour le RNIS et H.324 pour le réseau téléphonique analogique. Le H.323 est supporté par la quasi-totalité des constructeurs. Il est, pour cette raison, très largement utilisé comme protocole d'inter fonctionnement **[30].**

Dans un environnement H.323, l'établissement de la communication est effectué au moyen du protocole Q.931, le même que dans le RNIS. Le protocole RAS (Registration, Admission and Status) sert à l'enregistrement des équipements terminaux et au contrôle d'admission à la communication. Le H.245 permet de commander les applications de bout en bout. Les applications de données se servent du T.120; le H.225 gère la signalisation des appels, alors que l'audio et la vidéo disposent de plusieurs types de codecs. Le protocole RSVP (Ressource Reservation Protocol) est le protocole chargé de la gestion de la qualité de service.

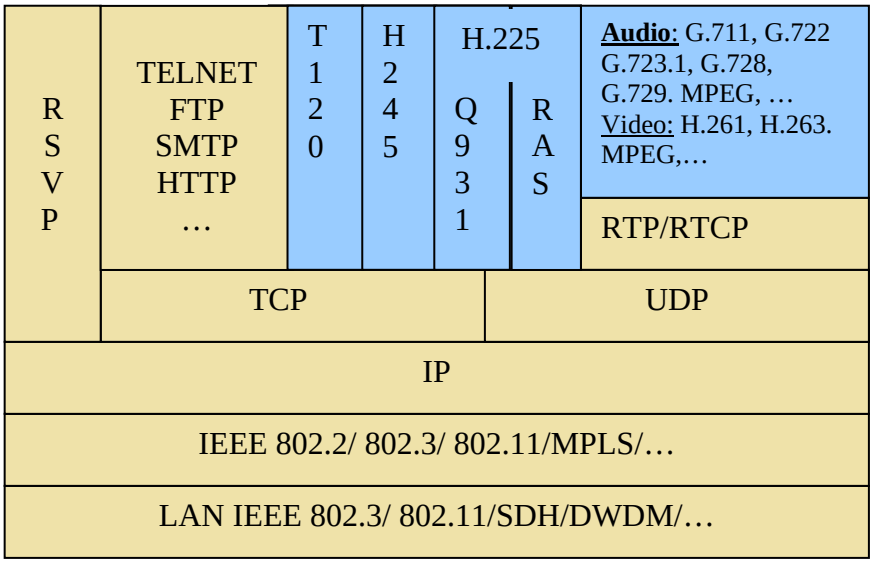


Fig 3.3 Architecture des protocoles selon le standard H323.

3.5.2 Le standard SIP (Session Initiation Protocol)

L'échange des messages de signalisation et de contrôle du protocole SIP est effectué sous la forme de transactions **[31].** Il est apparenté au protocole HTTP. Une transaction est composée d'une requête et d'une réponse. Les requêtes sont toujours émises par un client et les réponses par un serveur. Cette même structure client-serveur va se retrouver dans les terminaux, le serveur d'enregistrement, le proxy et le serveur de re-direction.

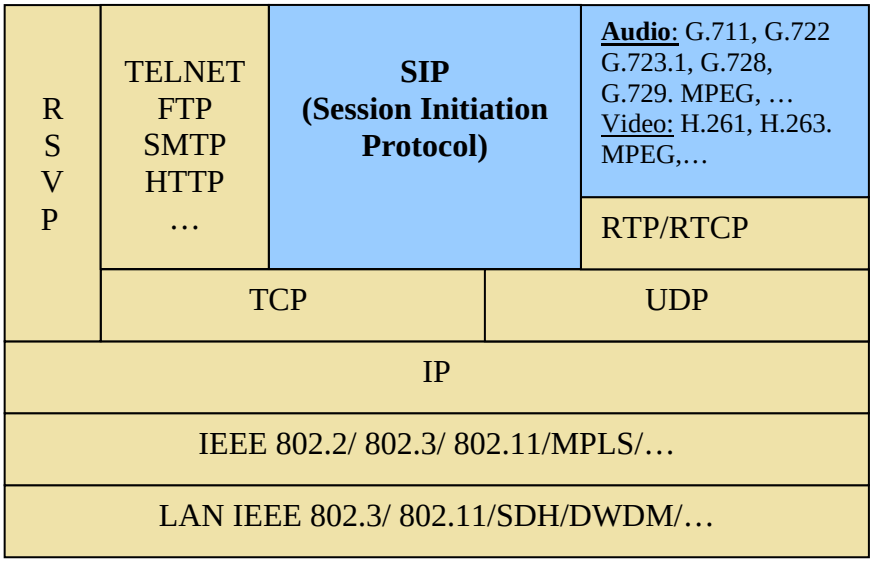


Fig 3.4 Architecture des protocoles selon le standard SIP.

3.5.3 Standard MGCP (Media GAteway Control protocol)

Le protocole MGCP sert à l'échange de messages de signalisation entre un contrôleur de passerelles de médias et des passerelles réparties dans un réseau IP. Pour l'établissement et la libération des connexions, MGCP se sert de signaux et d'événements. La standardisation de MGCP a été stoppée pour faire place à MEGACO/H.248 (Media Gateway Control Protocol), protocole élaboré en collaboration entre l'IETF et l'UIT. Ce nouveau standard n'étant pas dérivé de MGCP, la migration vers MEGACO/H.248 semble très difficile.

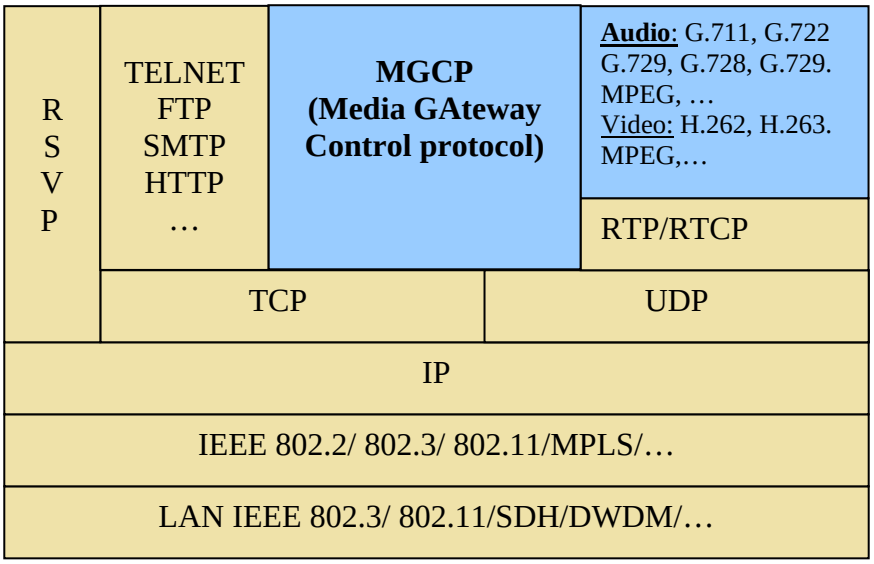


Fig 3.5 Architecture des protocoles selon le standard MGCP.

3.6 La qualité de service dans la VoIP

Internet est un réseau à commutation de paquets qui fournit un service d'acheminement des paquets « au mieux », plus connu sous l'appellation « Best-Effort ». Par cela, nous entendons que le réseau n'est pas capable d'assurer une garantie de service aux paquets, soit en délai, soit en débit, soit en perte de paquets. Tous les types de trafic qui traversent le réseau sont par conséquent traités de la même manière.

La qualité de service (Qos) correspond à l'ensemble des méthodes ou processus qu'une organisation de services met en oeuvre pour maintenir un niveau de qualité précis. Nous pouvons aussi considérer la qualité de service comme étant un ensemble de contraintes que le réseau doit respecter pour offrir un niveau de service approprié à la transmission des données. Ces contraintes sont axées sur le débit, le délai et la perte de paquets.

3.7 Facteurs affectant la qualité de service

3.7.1 Le délai (retard)

La maîtrise du délai de transmission est un élément essentiel pour bénéficier d'un véritable mode conversationnel et minimiser la perception d'écho (similaire aux désagréments causés par les conversations par satellites, désormais largement remplacées par les câbles pour ce type d'usage). Or la durée de traversée d'un réseau IP dépend de nombreux facteurs:

- Le débit de transmission sur chaque lien
- Le nombre d'éléments réseaux traversés
- Le temps de traversée de chaque élément, qui est lui même fonction de la puissance et la charge de ce dernier, du temps de mise en file d'attente des paquets, et du temps d'accès en sortie de l'élément
- Le délai de propagation de l'information, qui est non négligeable si on communique à l'opposé de la terre. Une transmission par fibre optique, à l'opposé de la terre, dure environ 70 ms.
- Le temps de codage et la mise en paquets de la voix

Les chiffres suivants (tirés de la recommandation UIT-T G114) sont donnés à titre indicatif pour préciser les classes de qualité et d'interactivité en fonction du retard de transmission dans une conversation téléphonique. Ces chiffres concernent le délai total de traitement, et pas uniquement le temps de transmission de l'information sur le réseau.

Classe n°	Délai	Commentaire
1	0 à 150 ms	Acceptable pour la plupart des conversations
2	150 à 300 ms	Acceptable pour des communications faiblement interactives
3	300 à 700 ms	Devient pratiquement une communication semi duplex
4	Au-delà de 700 ms	Inutilisable sans une bonne pratique de la conversation semi duplex

Tableau 3.1 : Délais requis pour la VoIP en fonction de la classe d'appartenance

En conclusion, on considère généralement que la limite supérieure « acceptable », pour une communication téléphonique, se situe entre 150 et 200 ms par sens de transmission (en considérant à la fois le traitement de la voix et le délai d'acheminement).

3.7.2 La gigue

La gigue est la variance statistique du délai de transmission. En d'autres termes, elle mesure la variation temporelle entre le moment où deux paquets auraient dû arriver et le moment de leur arrivée effective. Cette irrégularité d'arrivée des paquets est due à de multiples raisons telles que:

- L'encapsulation des paquets IP dans les protocoles supportés
- la charge du réseau à un instant donné
- la variation des chemins empruntés dans le réseau, etc…

Pour compenser la gigue, on utilise généralement des mémoires tampons (buffer de gigue) qui permettent de lisser l'irrégularité des paquets. Malheureusement ces paquets présentent l'inconvénient de rallonger d'autant le temps de traversée global du système. Leur taille doit donc être soigneusement définie, et si possible adaptée de manière dynamique aux conditions du réseau.

La dégradation de la qualité de service due à la présence de gigue, se traduit en fait, par une combinaison des deux facteurs cités précédemment: le délai et la perte de paquets; puisque d'une part on introduit un délai supplémentaire de traitement (buffer de gigue) lorsque l'on décide d'attendre les paquets qui arrivent en retard, et que d'autre part on finit tout de même par perte de certains paquets lorsque ceux-ci ont un retard qui dépasse le délai maximum autorisé par le buffer.

3.7.3 Le CoDec

Le codage consiste à compresser la parole afin de réduire le débit émis, et favoriser ainsi le transfert de données en temps réel. Le CoDec influe directement sur trois facteurs primordiaux dans une communication VoIP, le débit, le délai et le niveau de qualité de la parole.

Le débit est lié au taux de compression de la parole, le délai de traitement dépend généralement de la complexité des algorithmes utilisés, et la qualité de la parole est liée aux techniques de quantification et de prédiction utilisées. Le CoDec ne peut pas satisfaire ces facteurs en même temps, en effet, selon la loi de distorsion de Shannon, la réduction du débit implique automatiquement une dégradation de la qualité de la parole, ainsi, la conception d'un codeur à faible débit nécessite une haute complexité. Prenant à titre d'exemple les codeurs G.711 et G.729 de l'ITU-T, le G.711 est un codeur utilisant la technique PCM, il opère avec un débit

de 64 Kb/s pour une complexité de 0.01 MIPS (Million Instructions Par Second), la qualité MOS est de 4.1. Tandis que le G.729 qui utilise un algorithme plus complexe, c'est le CS-ACELP, il opère avec un débit de 8 Kb/s pour une complexité de 20 MIPS, sa qualité MOS est de 3.92.

Donc le choix du CoDec est un compromis entre le débit, le délai et la qualité de la parole, En conclusion un CoDec optimal est celui qui offre un débit minimum avec une meilleure qualité de voix et ce pour un délai de calcul minimum.

3.8 La perte de paquets

Lorsque les buffers des différents élément réseaux IP sont congestionnés, ils « libèrent » automatiquement de la bande passante en se débarrassant d'une certaine proportion des paquets entrant, en fonction de seuils prédéfinis. Cela permet également d'envoyer un signal implicite aux terminaux TCP qui diminuent d'autant leur débit au vu des acquittements négatifs émis par le destinataire qui ne reçoit plus les paquets. Malheureusement, pour les paquets de la voix, qui sont véhiculés au dessus d'UDP, aucun mécanisme de contrôle de flux ou de retransmission des paquets perdus n'est offert au niveau du transport. D'ou l'importance des protocoles RTP et RTCP qui permettent de déterminer le taux de perte de paquets, et d'agir en conséquence au niveau applicatif.

Si aucun mécanisme performant de récupération des paquets perdus n'est mis en place (cas le plus fréquent dans les équipements actuels), alors la perte de paquets IP se traduit par des ruptures au niveau de la conversation et une impression de hachure de la parole. Cette dégradation est bien sûr accentuée si chaque paquet contient un long temps de parole (plusieurs trames de voix). Par ailleurs, les codeurs à très faible débit sont généralement plus sensibles à la perte d'information, et mettent plus de temps à « reconstruire » un codage fidèle.

Enfin, connaître le pourcentage de perte de paquets sur une liaison n'est pas suffisant pour déterminer la qualité de la voix que l'on peut espérer, mais cela donne une bonne approximation. En effet, un autre facteur essentiel intervient; il s'agit du modèle de répartition de cette perte de paquets, qui peut être soit « régulièrement » répartie, soit répartie de manière corrélée, c'est-à-dire avec des pics de perte lors des phases de congestion, suivies de phases moins dégradées en terme de QoS.

3.9 Mécanismes de masquage des paquets perdus

Afin de réaliser une transmission de la voix en temps réel de haute qualité, un mécanisme de dissimulation de perte de paquets doit être mis en place. Plusieurs algorithmes de masquage des pertes de paquets PLC (Packet Loss Concealment) sont utilisés aussi bien au niveau de l'émetteur qu'au niveau du récepteur voir **Fig 3.6 [32]**.

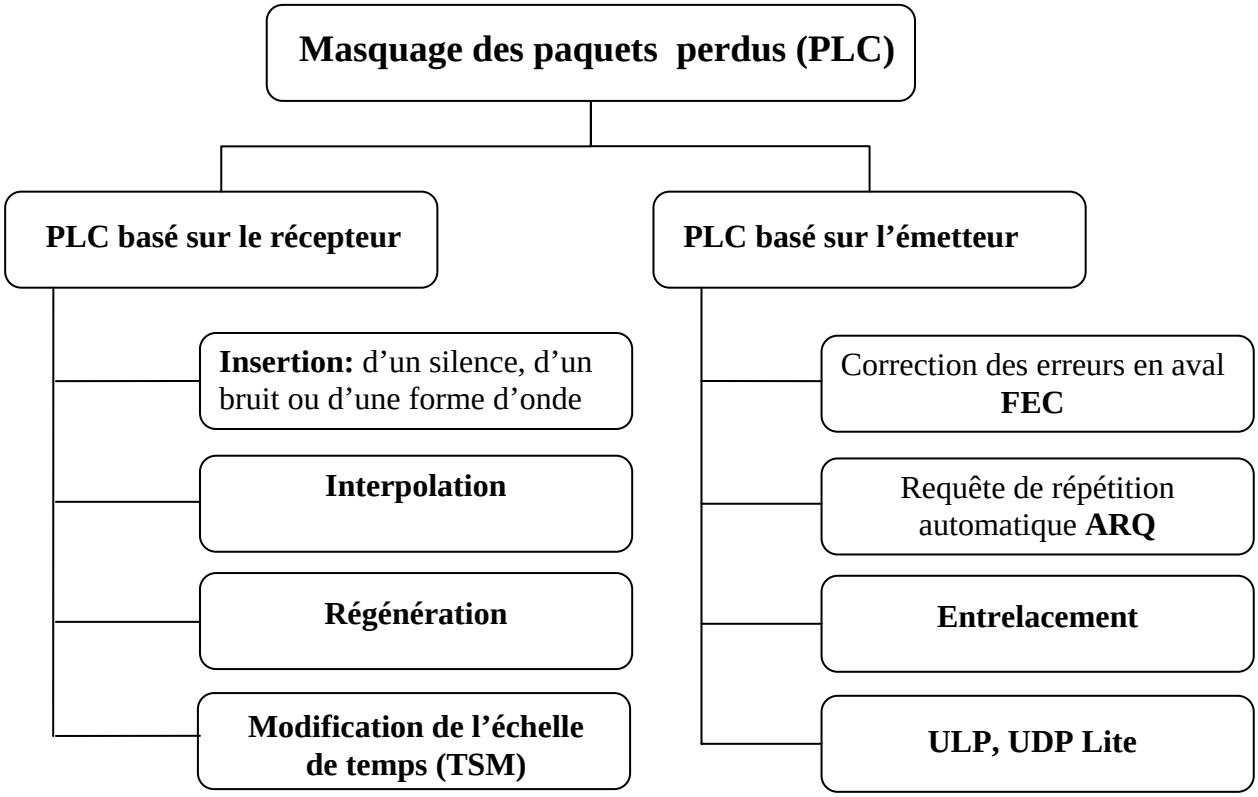


Fig 3.6 Les techniques de masquage des paquets perdus.

3.9.1 Masquage Basé sur l'Emetteur

Il existe plusieurs techniques de masquage basé sur l'émetteur, ces techniques sont généralement plus efficaces mais plus complexes. On peut distinguer deux types de techniques, celles qui ajoutent une redondance de contrôle et celles qui n'ajoutent pas.

Les méthodes qui ajoutent la redondance nécessitent une très large bande passante ou un long retard de bout en bout, parmi ces méthodes, on distingue celles qui envoient des paquets dupliqués, celles qui envoient avec les paquets courant des paquets précédents codés avec un faible débit, celles qui envoient des bits de correction d'erreurs sur les paquets en utilisant la méthode de correction d'erreurs en aval (FEC Forward Error Correction), ou encore celles qui utilisent la requête de répétition automatique (ARQ Automatic Repeat Request).

Les méthodes qui n'ajoutent pas de redondance, utilisent une redondance inhérente dans la trame de voix au niveau de la source. Une méthode typique entrelace les échantillons de la voix dans des paquets distincts et reconstruit les paquets perdus par interpolation en utilisant les échantillons survivants voisins.

3.9.1.1 La Requête de Répétition Automatique (ARQ)

La requête de répétition automatique ARQ (*Automatic Repeat Request*), appelée aussi contrôle en boucle fermée, est une technique de retransmission, dont les stratégies de base sont :

- La détection du paquet perdu qui se fait par le récepteur ou par l'émetteur.
- La stratégie de l'accusé de réception : Le récepteur informe l'encodeur, par un accusé de réception, s'il y a des erreurs, en conséquence, l'encodeur peut retransmettre les paquets perdus.
- La stratégie de rediffusion: elle détermine quelles données doivent être retransmises par l'émetteur.

Malgré sa robustesse contre les pertes brusques, cette technique ne peut pas être utilisée dans les applications en temps réel, telle que la VoIP, à cause du délai considérable et de la large bande passante nécessaires.

3.9.1.2 La correction d'erreurs en Aval (FEC)

Les techniques de correction d'erreurs en Aval FEC (forward error correction), consistent à ajouter des données redondantes au flux binaire transmis à partir desquelles le contenu des paquets perdus peut être récupéré. Il y a deux sortes d'informations redondantes qui peuvent être ajoutées afin d'améliorer le processus de masquage à savoir celles qui sont indépendantes du contenu du flux et celles qui sont basées sur les propriétés de la parole **[33].**

FEC indépendant au media

Dans ce type de FEC, il n'est pas nécessaire de connaître le type de donnes originales (parole ou vidéo). Les données originales auxquelles on ajoute des données de redondance, nommée parité, sont transmises vers le récepteur **[32].**

Les données de redondance sont dérivées des données originale par :

- Soit une opération de *OU exclusif (XOR) :* un seule paquet de parité est généré pour plusieurs paquets originaux.

- Soit on l'encode par le code *Reed-Solomon :* dans ce cas, multiples parités indépendantes peuvent être calculées pour le même ensemble de paquets. Le code *Reed-Solomon* permet d'obtenir une protection optimale contre les pertes, mais en contrepartie, il nécessite une grande complexité de traitement. Malgré que la méthode du XOR fournisse une protection sous optimale, elle est préférable dans les implémentations pratiques puisque on peut calculer plusieurs paquets de parité avec un faible coût de traitement par rapport à la méthode de *Reed-Solomon*.

Le FEC transmet k paquets originaux (D) et h paquets redondants de parité (P), la figure **Fig 3.7** montre un exemple pour k=3 (D1,D2,D3)et h=2. Le FEC génère deux paquets redondants (P1,P2) à partir des paquets de données. Si un paquet de données (D3) et un paquet de parité (P1) par exemple, sont perdus, le récepteur peur reconstituer le paquet de données (D3) par l'utilisation des paquets reçus avec succès D1,D2 et P2.

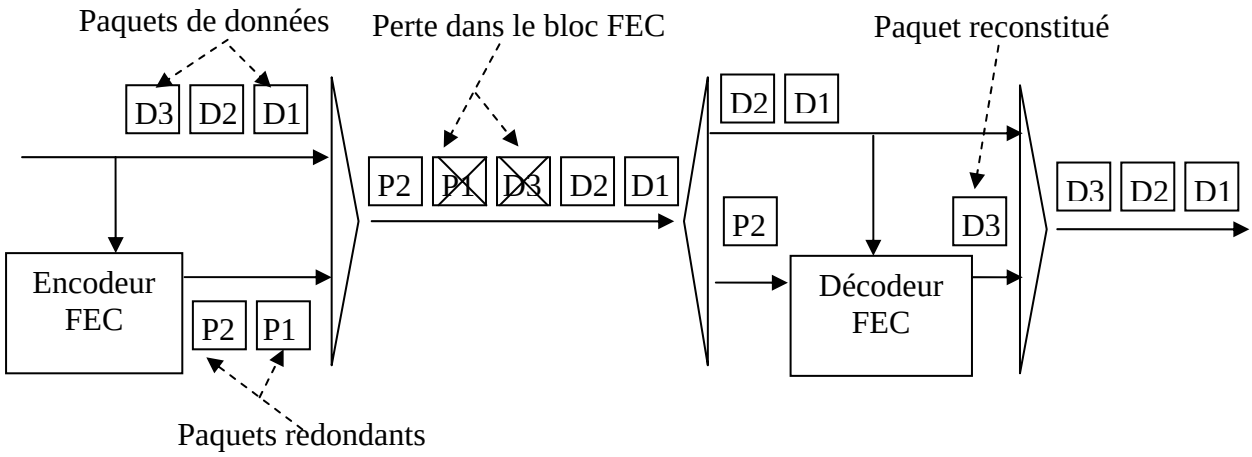


Fig 3.7 Exemple du FEC indépendant au media.

FEC spécifique au média

Si un paquet original de données est perdu, les paquets redondants de données, qui sont reliés au paquet perdu, sont utilisés pour reconstituer le paquet perdu voir **Fig 3.8**.

La première donnée de la parole transmise est définie sous le nom du codage primaire tandis que les transmissions redondantes sont définies comme un codage secondaire, généralement, les paquets secondaires sont produits par des codages à bas débit par rapport au codage primaire, qui signifie une qualité moins que le primaire.

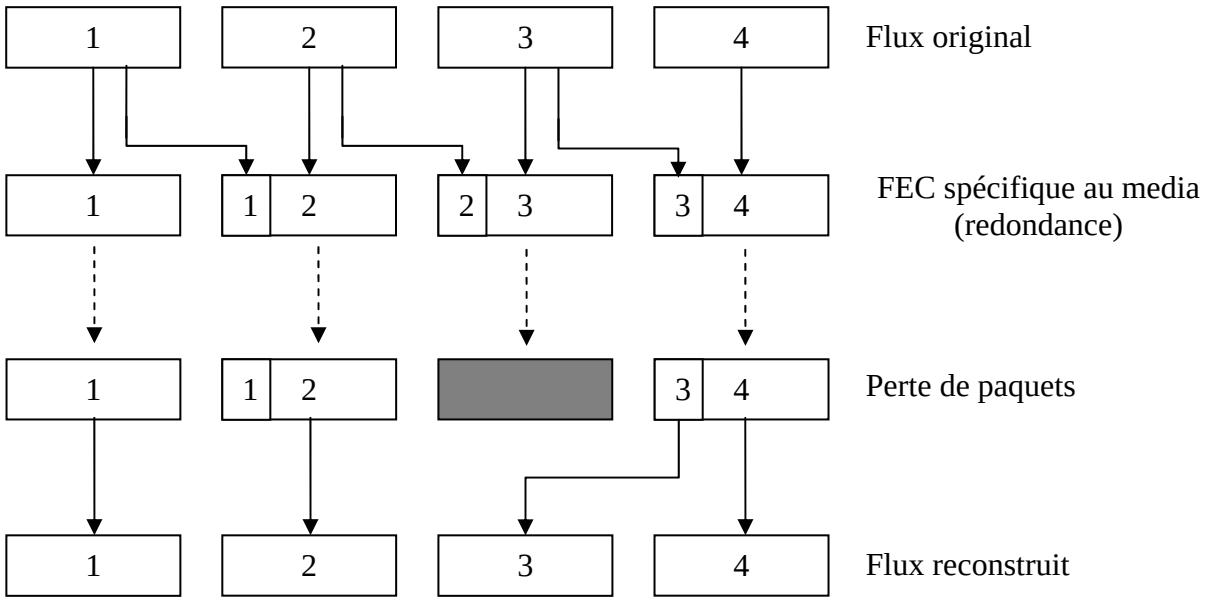


Fig 3.8 FEC spécifique au média.

3.9.1.3 Entrelacement

L'entrelacement devient une technique de masquage utile lorsque le délai de bout en bout aura une importance secondaire. Lorsque la taille des unités de données (trames de parole) est faible par rapport à la taille du paquet, l'ordre séquentiel des unités est changé par rapport à celui produit par le codeur, l'émetteur entrelace les unités de données et par conséquent, il change l'ordre de séquencement, voir figure **Fig 3.9**.

Au niveau du récepteur, les unités de données sont rassemblées, réordonnées puis livrées au décodeur. Le but de cette technique est de distribuer l'effet de perte de paquets sur des petits intervalles séparés au lieu de perdre un grand intervalle. L'effet de perte est diminué pour les raisons suivantes :

- Les petits intervalles de lacunes (gap) correspondent aux intervalles de parole plus courts par rapport à la longueur d'un phonème. Et puisque l'homme est capable d'interpoler mentalement les petites lacunes, donc l'intelligibilité de la parole est préservée.

-De plus, si le récepteur est muni d'un mécanisme de dissimulation de perte (par exemple, les lacunes sont remplacées en utilisant une interpolation des trames reçues adjacentes), alors une performance supérieure est obtenue si l'interpolation se fait sur des petits intervalles au lieu des grands intervalles.

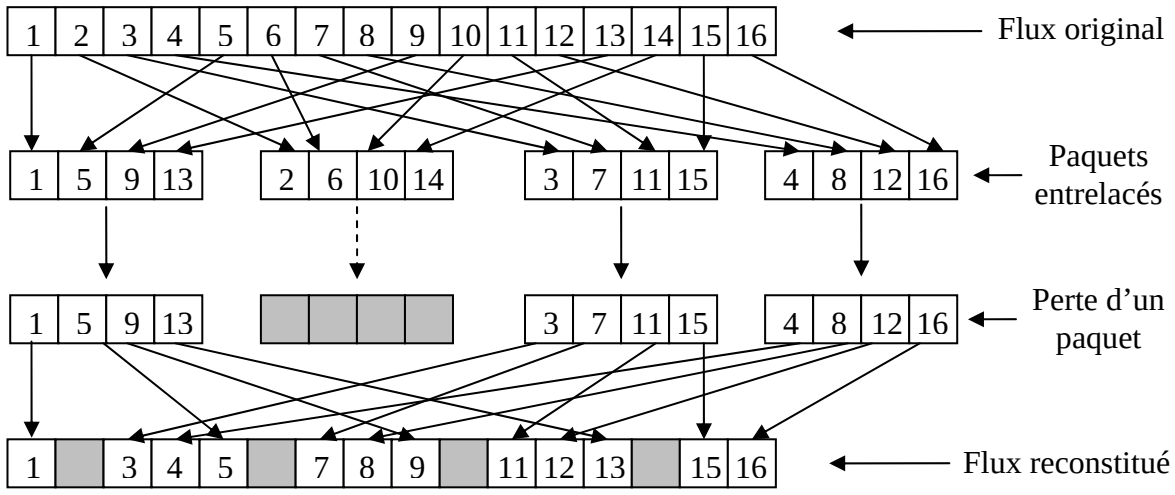


Fig 3.9 Exemple d'entrelacement.

3.9.1.4 La Protection à Niveau Inégal

Lorsque les sous-divisions constituant la donnée n'ont pas la même importance (Les données de la parole, en particulier), une technique dite protection à niveau inégal *ULP* (*Uneven Level Protection*) peut être appliquée. Cette technique attribue plus de protection aux données les plus importantes. Par exemple, si la parole est codée par le codage CELP, alors, le pitch et les paramètres du filtre de prédiction ont une grande importance par rapport à l'excitation. Une erreur sur les paramètres du filtre de prédiction peut réduire considérablement la qualité et conduit à un système instable. Par contre, l'erreur sur l'excitation n'influe pas sur la qualité perceptuel. Ces propriétés conduisent à l'usage d'une protection inégale pour des données n'ayant pas les mêmes importances. Les unités de données sont arrangées dans un paquet de type *RTP* par ordre d'importance décroissant. Plus de protection est appliquée aux débuts des unités, c'est-à-dire aux données les plus importantes.

3.9.2 Masquage Basé sur le récepteur

Les techniques de masquage basé sur le récepteur consistent à produire des remplacements semblables aux paquets originaux perdus. Il existe trois catégories de méthodes de dissimulation : l'insertion, l'interpolation et la régénération.

3.9.2.1 L'insertion

Elle consiste à remplacer les paquets perdus soit par un silence, un bruit ou par répétition de la dernière bonne trame reçue, cette méthode est simple à implémenter, elle est efficace pour des paquets de longueurs courtes (<4ms) et de faibles taux de perte (< 2 %). Ses performances se dégradent rapidement lorsque la taille des paquets augmente.

Insertion de silence

C'est la possibilité la plus simple, elle consiste à remplacer le segment de parole perdu par des échantillons de valeurs mules (0). Cette méthode donne une mauvaise qualité de parole même pour des faibles taux de perte.

Insertion de bruit

Dans cette technique, on augmente légèrement la complexité par rapport à l'insertion de silence en générant un bruit pour remplacer le segment de la parole perdu. Cette technique exploite le phénomène de «restauration phonémique», dans ce phénomène, le système d'audition humain interpole mieux les segments de parole perdus remplacés par un bruit que des segments nuls.

3.9.2.2 Répétition de paquets

La répétition de la dernière bonne trame reçue est la méthode la plus simple pour approximer le signal perdu. Il est simplement nécessaire de mémoriser une copy de la dernière trame. La figure **Fig 3.10** montre le signal original S(n), $\tilde{S}(n)$ est le signal avec perte et $\hat{S}(n)$ est le signal reconstitué en utilisant la méthode de répétition, n étant le nombre d'échantillons. Puisque l'intervalle de packetization L n'est pas choisi en fonction de la période du pitch P, le signal reconstitué présente des discontinuités. Cette méthode améliore légèrement la qualité de la parole par rapport à l'insertion de silence ou de bruit.

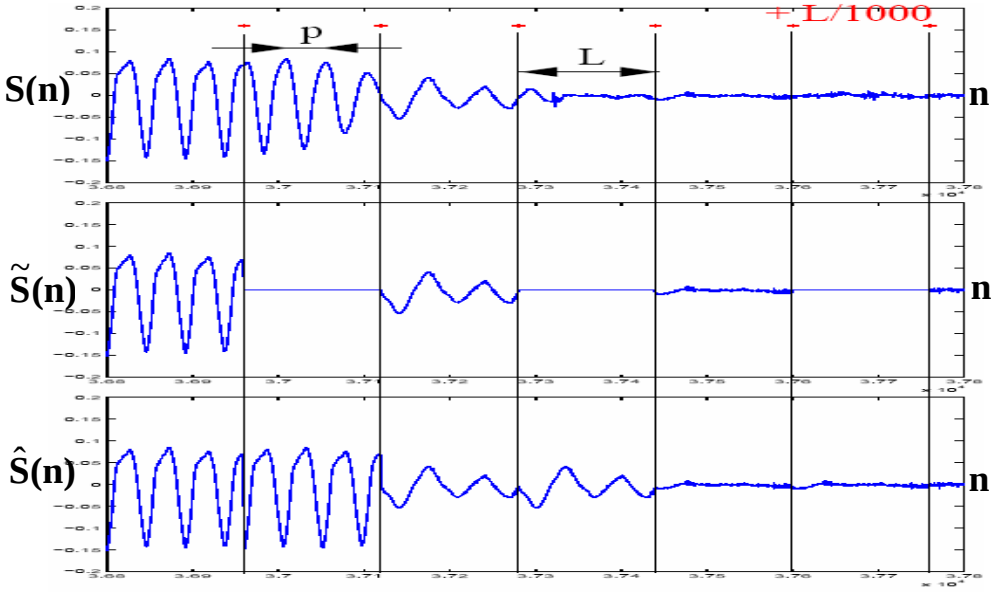


Fig 3.10 Masquage par répétition de paquets.

3.9.2.3 L'alignement temporel (Pattern matching)

Dans cette technique, le récepteur utilise un modèle de signal, constitué d'un segment d'échantillons correctement reçu juste avant le segment perdu. Une recherche d'un segment ayant une similarité avec le segment modèle est ensuite effectuée dans une fenêtre d'échantillons situant avant l'intervalle du modèle. Enfin, l'amplitude du segment ayant une similarité maximale avec le segment modèle est ajustée pour avoir un lissage dans le signal de sortie.

3.9.2.4 L'interpolation

Elle consiste à interpoler quelques paramètres des bonnes trames antérieures et futures afin de trouver un remplacement pour la trame perdue. L'avantage des méthodes d'interpolation par rapport aux méthodes d'insertion, est qu'elles prennent en compte le changement des caractéristiques du signal. Par conséquent, les performances sont meilleures.

3.9.2.5 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 récupération.

3.9.2.6 Modification de l'échelle de temps

La modification de l'échelle de temps (Time Scale Modification TSM) est introduite par H. Sanneck en 1995, elle consiste à allonger un ou plusieurs segments du signal situant avant et/ou après la zone de perte afin de récupérer le segment perdu. Dans la figure **Fig 3.11,** Le segment 2 est perdu, les segments 3,4 et 5 sont allongés, pour assurer une transition lisse entre le segment 1 et 3, il faut utiliser la technique de recouvrement-addition qui consiste à multiplier, dans l'intervalle d'intersection, le segment 1 et 3 par une fenêtre de pondération puis effectuer leurs somme. La modification de l'échelle de temps se fait sans altération de la période du pitch, ni l'intelligibilité.

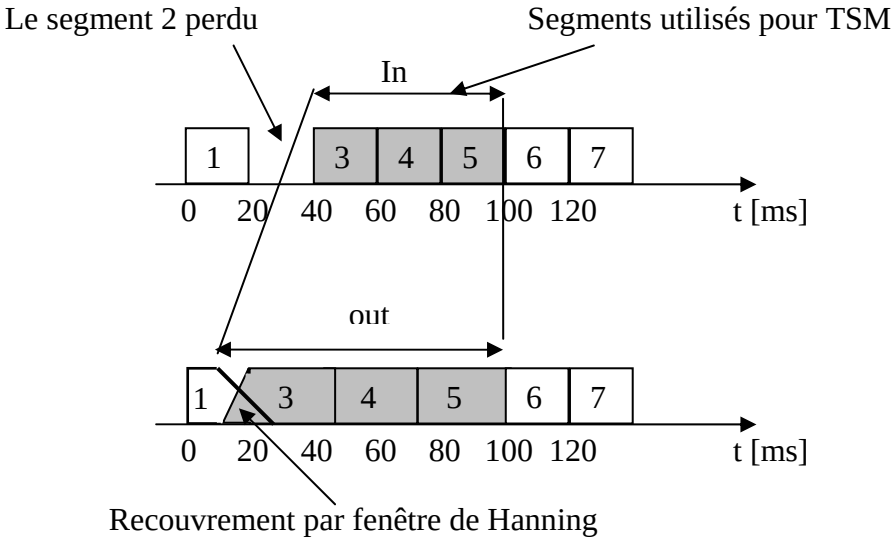


Fig 3.11 Masquage de perte de paquets basé sur la modification de l'échelle de temps.

3.10 Conclusion

Dans ce chapitre nous avons présenté un aperçu sur la transmission de la voix sur le réseau IP. Nous avons commencé par une présentation des architectures TCP/IP et OSI, ensuite, nous avons présenté les différents systèmes et protocoles utilisés dans la VoIP. Nous avons aussi abordé les problèmes et les facteurs affectant la qualité de service, enfin, nous avons terminé ce chapitre par une présentation des techniques utilisées dans la récupération des paquets perdus dans le réseau IP.

Chapitre 4

Codeur de la norme G.729

4.1 Introduction

Le standard G.729, défini par l'ITU-T pour le système de codage à 8 kb/s est un compromis entre l'algorithme Algebraic Code Excited Linear Prediction coder (ACELP) présenté par l'université de Sherbrooke au Canada en association avec France Telecom et l'algorithme Conjugate Structure Code Excited Linear Prediction coder, proposé par Nippon Telephone et Telegraph (NTT) au Japon **[34][18][35].**

Le codeur G.729 est destiné pour des applications de téléphonie sans fil et des applications sous réseaux telles que Internet (VoIP). Il fournit une qualité téléphonique (toll quality) de voix, il donne des résultats meilleurs, même dans des conditions extrêmes telles que le bruit ou les erreurs de transmission.

Dans le codage CELP, utilisé par le G.729, L'approche de recherche des paramètres du filtre et de l'excitation est appelée *analyse par synthèse*. Pour chaque trame, l'encodeur recherche à travers l'espace des paramètres, en effectuant une opération de décodage pour chaque boucle de recherche. La sortie du décodeur (le signal synthétisé) est comparée avec le signal de parole original. Les paramètres qui apportent une petite erreur sont alors choisis et envoyés au décodeur. Dans ce modèle, nous avons analysé le signal en synthétisant à plusieurs reprises la sortie du décodeur, d'où le nom analyse par la synthèse.

4.2 Pourquoi le G.729

Le délai, le débit binaire et la qualité sont des attributs qui favorisent l'utilisation du G.729 en offrant des avantages aux utilisateurs, tels que la qualité de service, l'interopérabilité et une largeur de bande accrue. La qualité de service est meilleure car ce codeur offre une qualité téléphonique de la voix pour un minimum de délai de traitement.

4.3 Fiche technique du codeur G.729

Caractéristiques	Valeurs
Organisme de normalisation	ITU
Type de codeur	CS-ACELP
Date d'apparition	1995
Débit binaire	8 kb/s
Qualité de la voix	Téléphonique (Toll) avec MOS=3.92
Masquage de pertes de paquets	3%
Complexité (MIPS)	22
RAM (Kmots de 16 bits)	2.6
Taille de la trame	10 ms
Délai du Codec	25 ms

Tableau 4.1 Caractéristiques techniques du codeur G.729

4.4 Description générale du CoDec G.729

Le codeur de prédiction CS-ACELP est fondé sur le modèle de codage prédictif linéaire à excitation par code (CELP) (*code-excited linear-prediction*). Le codeur opère sur des trames vocales de 10 ms correspondant à 80 échantillons à raison de 8000 échantillons par seconde. Pour chaque trame de 10 ms, le signal vocal est analysé pour en extraire les paramètres du modèle de prédiction CELP (coefficients du filtre de prédiction linéaire, index et gains de répertoire codé adaptatif et de répertoire codé fixe). Ces paramètres sont codés et transmis. L'affectation des positions binaires aux paramètres de codage est représentée dans le Tableau 4.2. Dans le décodeur, ces paramètres servent à récupérer les paramètres d'excitation et du filtre de synthèse.

On reconstitue la parole par filtrage du flux d'excitation par le filtre de synthèse à court terme, comme indiqué sur la Figure **Fig 4.1**. Ce filtre fait appel à une prédiction linéaire (LP) (*linear prediction*) du 10 ème ordre. Le filtre à long terme ou de synthèse tonale, est mis en oeuvre par la méthode dite du répertoire codé adaptatif. Le signal vocal, après avoir été reconstitué par calcul, est encore amélioré par un postfiltre fondamental.

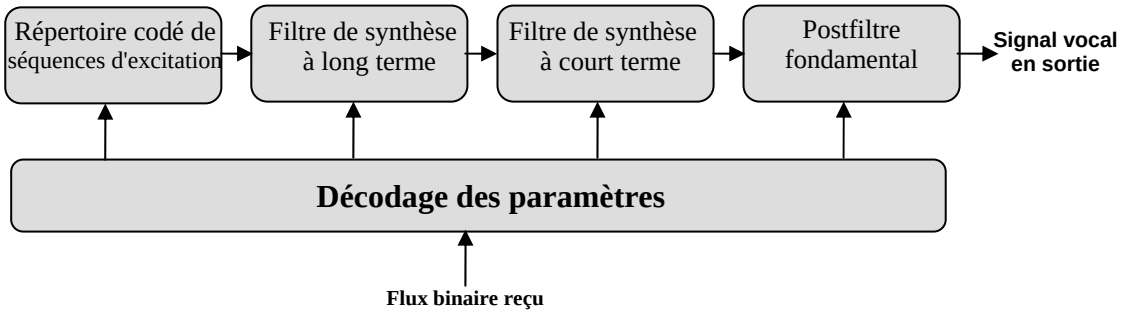


Fig 4.1 Schéma fonctionnel du modèle théorique de la synthèse par algorithme CELP.

4.5 Le codeur

Le principe du codage est schématisé sur la figure **Fig 4.2**. Le signal d'entrée subit un filtrage passe-haut et une normalisation dans le bloc de prétraitement. La sortie de ce dernier sera utilisée comme entrée pour toutes les analyses suivantes. L'analyse prédictive linéaire est effectuée toutes les trames de 10 ms afin de calculer les coefficients de filtrage prédictif linéaire. Ceux-ci sont convertis en paires de lignes spectrales (LSP) (*line spectrum pairs*) et numérisés sur 18 éléments binaires *(L0, L1, L2, L3)* par quantification vectorielle (VQ) (*vector quantization*) prédictive en deux étapes.

Le signal d'excitation est choisi au moyen d'une procédure de recherche par analyse et synthèse dans laquelle l'erreur entre le signal vocal original et le signal vocal reconstitué est minimisée en fonction d'une mesure de distorsion pondérée par la perception. A cette fin, le signal d'erreur passe par un filtre de pondération perceptive dont les coefficients sont déduits du filtre de prédiction linéaire avant quantification. Les poids de la pondération perceptive sont rendus adaptatifs afin d'améliorer la qualité des signaux d'entrée ayant une réponse en fréquence uniforme.

Les paramètres d'excitation (par répertoire codé fixe et par répertoire codé adaptatif) sont déterminés à chaque sous-trame de 5 ms (soit 40 échantillons). Les coefficients du filtre de prédiction linéaire, quantifiés et non quantifiés, sont utilisés pour la deuxième sous-trame, alors que la première utilise une interpolation des coefficients du filtre de prédiction linéaire (aussi bien quantifiés que non quantifiés).

Le délai tonal en boucle ouverte est estimé toutes les trames de 10 ms, sur la base du signal vocal issu du pondérateur perceptif. Les opérations suivantes sont reprises pour chaque sous-trame. Le signal cible $x(n)$ est calculé par filtrage de l'énergie résiduelle du codage prédictif linéaire dans le filtre de synthèse pondérée $W(z)/\hat{A}(z)$. Les états initiaux de ces filtres sont mis à jour par filtrage de l'erreur mesurée entre l'énergie résiduelle du codage prédictif linéaire et l'excitation. Cela équivaut au procédé courant consistant à soustraire – du signal vocal pondéré – la réponse à entrée nulle du filtre de synthèse pondérée. La réponse impulsionnelle $h(n)$ du filtre de synthèse pondérée est calculée. Une analyse tonale en boucle fermée est ensuite effectuée (afin de déterminer le délai et le gain par répertoire codé adaptatif) au moyen du signal cible $x(n)$ et de la réponse impulsionnelle $h(n)$, par recherche autour de la valeur du délai tonal en boucle ouverte. On utilise un délai tonal fractionnaire, de résolution 1/3. Ce délai tonal est codé sur 8 éléments binaires dans la première sous-trame et codé différentiellement sur 5 éléments binaires dans la deuxième sous-trame. Le signal cible $x(n)$ est mis à jour par soustraction de la contribution (filtrée) du répertoire codé adaptatif et ce nouveau signal cible, $x'(n)$, est utilisé lors de l'exploration du répertoire codé fixe afin de déterminer l'excitation optimale. On fait appel à un répertoire algébrique de mots de 17 éléments binaires pour l'excitation par répertoire codé fixe.

Les gains des contributions par répertoire codé adaptatif et par répertoire codé fixe sont quantifiés vectoriellement sur 7 éléments binaires (avec application au gain par répertoire codé fixe d'une prédiction par analyse à moyenne mobile). Finalement, les mémoires des filtres sont mises à jour au moyen du signal d'excitation ainsi déterminé.

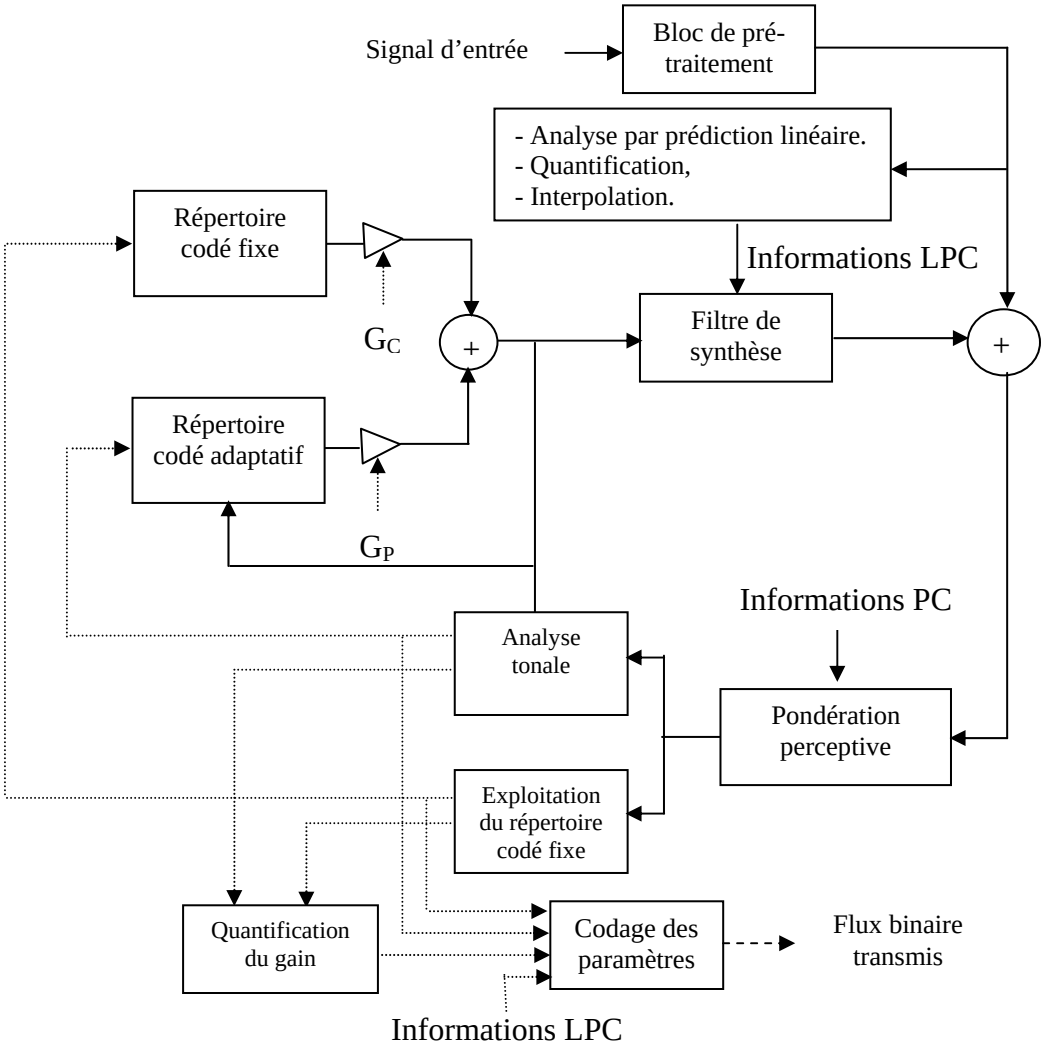


Fig 4.2 Principe du codeur *CS-ACELP* ***G.729.***

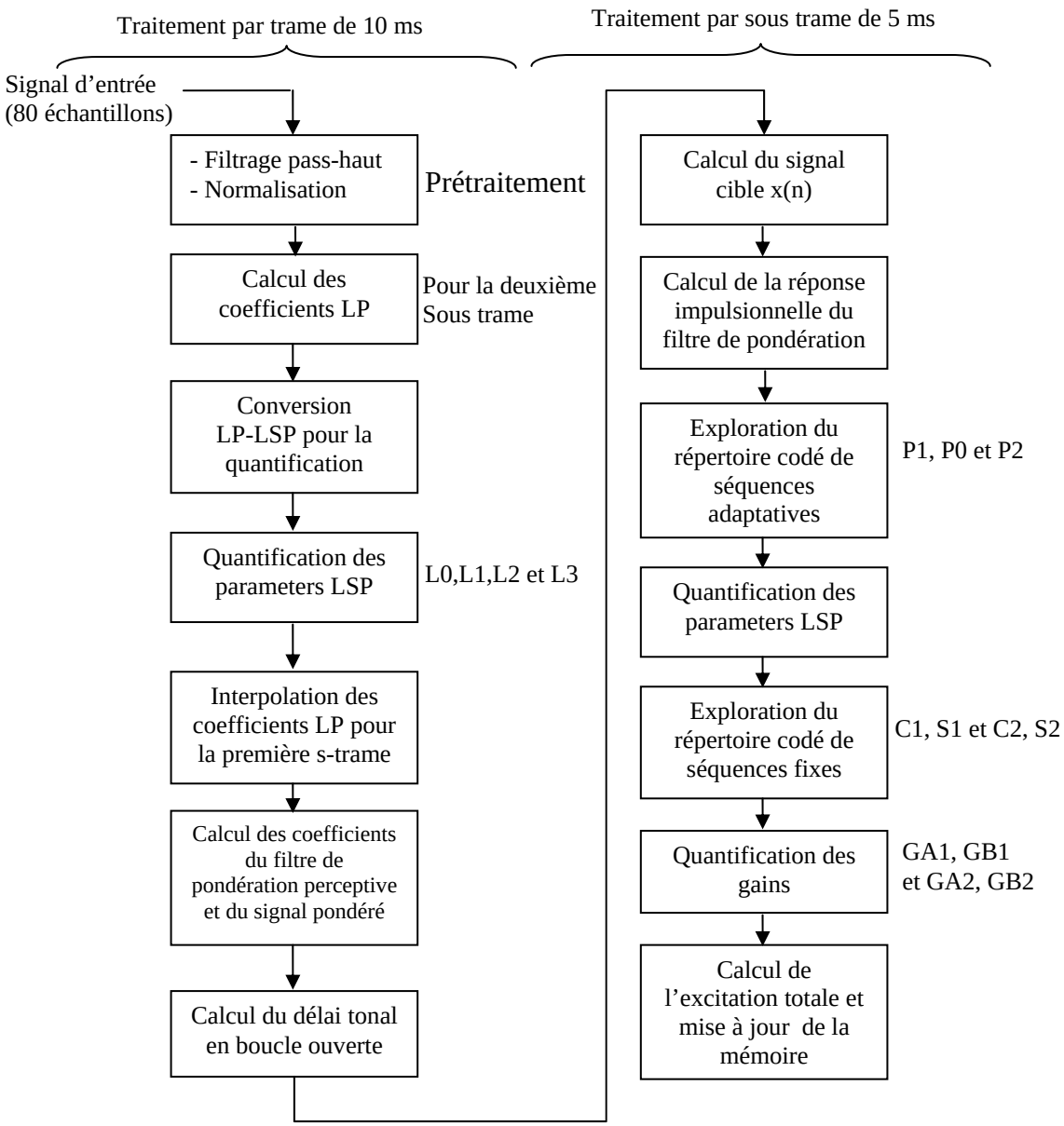


Fig 4.3 Algorithme de codage du G.729 **[35]**.

Tableau 4.2 Affectation des bits dans l'algorithme de codage CS-ACELP à 8 kbit/s (Trames de 10 ms)

Paramètres	Mot de code	Sous-trame1	Sous-trame2	Total par trame
Paires de raies spectrales	L_0, L_1, L_2, L_3			18
Délai du répertoire codé adaptatif	P1, P2	8	5	13
Parité du délai tonal	P0	1		1
Index de répertoire codé fixe	C1, C2	13	13	26
Signe du répertoire codé fixe	*S*1, *S*2	4	4	8
Gains du répertoire (étape 1)	*GA*1, *GA*2	3	3	6
Gains du répertoire (étape 2)	*GB*1, *GB*2	4	4	8
Total				80

4.6 Le décodeur

Le principe du décodeur est représenté sur la Figure **Fig 4.3**. 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 2 délais tonaux fractionnaires, les 2 vecteurs de répertoire codé fixe et les deux séries de gains par répertoire codé adaptatif et par répertoire codé fixe. Les coefficients en paires LSP sont interpolés et reconvertis en coefficients de filtre de prédiction linéaire pour chaque sous trame de 5 ms, qui passe par les étapes suivantes:

L'excitation est construite par combinaison des codes vectoriels adaptatifs et fixes, normalisés par leurs gains respectifs.

Le signal vocal est reconstitué par filtrage de l'énergie d'excitation dans le filtre de synthèse du codage prédictif linéaire.

Le signal vocal reconstitué est envoyé dans un bloc de post-traitement, qui comprend un postfiltre 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.

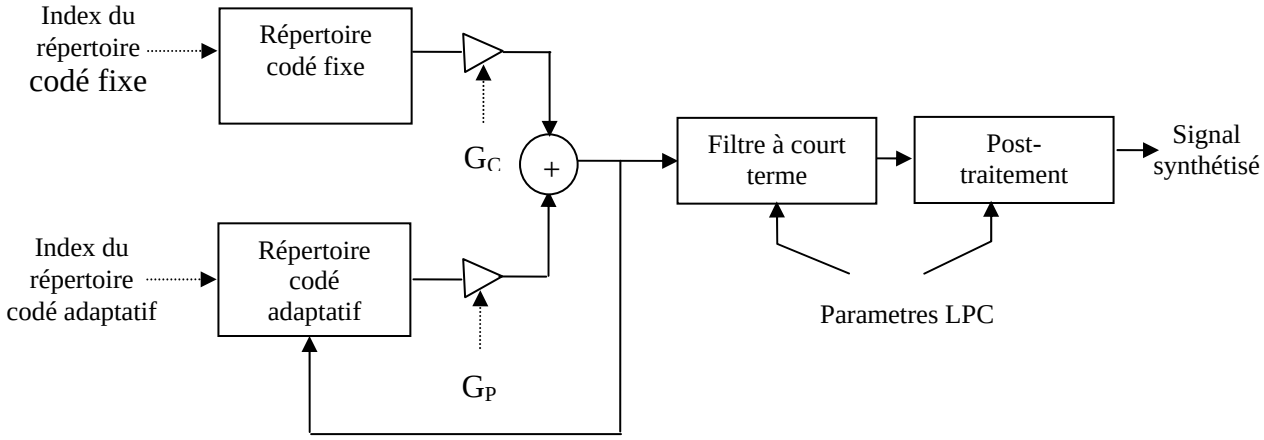


Fig 4.4 Principe du décodeur CS-ACELP G.729.

4.7 Dissimulation des trames effacées

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'effacement de trames dans le flux binaire. 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 **[36].**

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. 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 répertoire codé adaptatif est utilisé et la contribution du répertoire codé 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 répertoire codé adaptatif est mise à zéro, la contribution du répertoire codé fixe est construite par sélection aléatoire d'un index de répertoire et d'un index de signe.

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

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

2) Affaiblissement de gains du répertoire codé adaptatif et du répertoire codé fixe.

3) Affaiblissement de l'énergie mémorisée par le prédicteur de gain.

4) Production de l'excitation de remplacement.

4.7.1 Répétition des paramètres du filtre de synthèse

Le filtre de synthèse pour une trame effacée utilise les paramètres de prédiction linéaire de la dernière trame non effacée. Le registre du prédicteur à moyenne mobile des coefficients LSF contient les valeurs des mots de code $\hat{l}_i$. 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 $\hat{\omega}_i$ et du registre de prédicteur selon la formule suivante:

$$\hat{l}_i = \left[\hat{\omega}_i^{(m)} - \sum_{k=1}^{4} \hat{p}_{i,k} \hat{l}_i^{(m-k)} \right] / \left(1 - \sum_{k=1}^{4} \hat{p}_{i,k} \right) \qquad i = 1,...,10 \quad (4.1)$$

où les coefficients du prédicteur à moyenne mobile, $\hat{p}_{i,k}$ sont ceux de la dernière bonne trame reçue.

4.7.2 Affaiblissement de gains du répertoire codé adaptatif et du répertoire codé fixe

Le gain du répertoire codé fixe est fondé sur une version affaiblie du précédent gain du répertoire codé fixe. Il est donné par:

$$g_c^{(m)} = 0.98 g_c^{(m-1)} \quad (4.2)$$

où *m* est l'index de la sous-trame. Le gain de répertoire codé adaptatif est fondé sur une version affaiblie du précédent gain de répertoire codé adaptatif. Il est donné par:

$$g_p^{(m)} = 0.9 g_p^{(m-1)} \text{ avec la limite } g_p^m < 0.9 \quad (4.3)$$

4.7.3 Affaiblissement de l'énergie mémorisée par le prédicteur de gain

Le prédicteur de gain utilise l'énergie des codes vectoriels de répertoire 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 répertoire codé.

La valeur de $\hat{U}(m)$ pour la sous-trame actuelle m est alignée sur la valeur moyenne quantifiée de l'erreur sur la prédiction de gain, affaiblie de 4 dB, comme suit:

$$\hat{U}^{(m)} = \left(0.25 \sum_{i=1}^{4} \hat{U}^{(m-1)} - 4.0 \right) \quad \text{avec la limite} \quad \hat{U}^{(m)} \geq -14 \quad (4.4)$$

4.7.4 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 répertoire codé adaptatif est utilisé et la contribution du répertoire codé 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. Le gain de répertoire codé adaptatif est fondé sur une valeur affaiblie selon l'équation (4.3).

Si la dernière trame reconstituée avait été classifiée comme étant apériodique, la trame actuelle est considérée également comme apériodique et la contribution du répertoire codé adaptatif est mise à zéro. La contribution du répertoire codé fixe est construite par sélection aléatoire d'un index de répertoire et d'un index de signe. Le générateur de séquences aléatoires utilise la fonction suivante:

$$germe = 31821\ germe + 13849 \qquad (4.5)$$

avec la valeur initiale 21845 comme germe. L'index de répertoire fixe est extrait des 13 éléments binaires de poids faible du nombre aléatoire suivant. Le signe de répertoire fixe est extrait des 4 éléments binaires de poids faible du nombre aléatoire suivant. Le gain de répertoire fixe est affaibli conformément à l'équation (4.2).

4.8 Conclusion

Dans ce chapitre nous avons présenté une description générale du codeur G.729, c'est l'un des codeurs les plus utilisés dans les systèmes VoIP. Le G.729 appartient à la famille des codeurs hybrides CELP, il offre une qualité *téléphonique (toll quality)* de parole à un moyen débit de 8 Kb/s.

Chapitre 5

Modification de l’échelle temporelle du signal parole

5.1 Introduction

La modification de l’échelle temporelle (Time Scale Modification, TSM) se rapporte au traitement de compression ou de dilatation de l’échelle temporelle d’un signal de parole. Un signal dont l’échelle temporelle est compressée aura une durée courte, tandis que celui ayant une échelle temporelle dilatée aura une durée longue. Une simple accélération du signal parole produit des périodes fondamentales plus courtes, ce qui implique une augmentation de la fréquence, qui se traduit par des sons aigus non intelligibles. La modification de l’échelle temporelle maintient les propriétés du signal original telles que la période fondamental, l’identification du locuteur et l’intelligibilité.

5.2 Domaines d’application de la TSM

Les domaines d’application de la modification de l’échelle temporelle sont très variés on peut citer **[37] p36** :

- *Apprentissage d’une langue étrangère*

L’apprentissage d’une langue étrangère peut être redue facile par l’écoute d’un locuteur dont le débit s’adapte aux progrès de compréhension de l’étudiant. Les élèves peuvent plus facilement imiter les gestes articulatoires de la production de parole lorsque la voix est ralentie. De plus un enregistrement de leurs propres voix permet de corriger leurs erreurs d’articulation.

D'autre part, des études ont montré qu'écouter deux fois à une vitesse double un matériau sonore d'apprentissage était plus efficace que l'écouter une seule fois à vitesse normale.

- *Lecture pour les aveugles*

Des appareils de compression temporelle de la voix ont déjà été utilisés dans des programmes d'éducation. D'autre part, la compréhension d'une phrase prononcée plus rapidement qu'on ne peut le réaliser physiquement est possible; ainsi comme des taux d'accélération jusqu'à 2 fournissent toujours une bonne intelligibilité et une bonne compréhension, les textes peuvent être lus beaucoup plus rapidement que ne le fait la lecture en braille. Les bibliothèques sonores peuvent donc utiliser ce type d'outil.

- *Apprentissage à la lecture rapide*

L'apprentissage de la lecture rapide peut être amélioré si le sujet lit un texte en même temps qu'il écoute la voix accélérée.

- *Reconnaissance de la parole*

Des systèmes de reconnaissance de la parole possèdent une phase de pré-traitement qui consiste à normaliser temporellement les mots.

- *Répondeurs téléphoniques, dictaphones et serveurs vocaux*

L'accélération ou le ralentissement de la voix est utile dans les systèmes de répondeurs téléphoniques (accélération pour la recherche rapide des messages, ralentissement pour améliorer l'intelligibilité, les serveurs d'informations vocales et également les dictaphones (synchronisation du débit vocal à la vitesse de frappe).

- *Réduction de largeur de bande et compression de données*

Par compression de l'échelle temporelle, il est possible de réduire la largeur de bande du signal.

- *Amélioration du rapport signal à bruit*

La contraction temporelle suivie de l'expansion temporelle peut être utilisée comme un filtre de corrélation pour améliorer le rapport signal à bruit dans le cadre de signaux vocaux bruités.

- *Production audio et vidéo*

Une fonction de recherche rapide est parfois nécessaire dans les appareils de lecture audio et vidéo. Elle permet de repérer le passage désiré à haute vitesse sans modification des fréquences. Cette application requiert des taux de dilatation élevés, mais se satisfait généralement d'une qualité médiocre.

- *Synchronisation audio/vidéo*

Pour ajuster la durée d'un enregistrement de manière à ce qu'il corresponde exactement à la durée de la séquence sans modifier son contenu spectral. Cela évite par exemple d'avoir à ré-enregistrer la voix d'un acteur pour faire correspondre l'image des mouvements de lèvres au son.

- *Masquage des paquets perdus dans le domaine de la VoIP:* (c'est l'objet de notre travail, nous verrons par la suite, en détail, toute la procédure de masquage de perte de paquets) Dans le cas de perte de paquets, les segments de parole avoisinants le paquet perdu sont allongés dans le but de masquer la zone de perte.

5.3 Dualité temps/fréquence

Le problème dans la modification de l'échelle temporelle d'un signal parole $x_a(t)$ de durée $\Delta(t)$ est lié à la distorsion dans le domaine fréquentiel (**Fig 5.1**) **[37] p21**. La dualité entre la modification de l'échelle temporelle et la modification de l'échelle fréquentielle devient claire en considérant le signal $y_a(t)$ qui correspond au signal original $x_a(t)$ joué à une vitesse α fois la vitesse d'enregistrement. Ainsi, une période originale $\Delta(t)$ est écoutée en $\Delta(t)/\alpha$ et $y_a(t) = x_a(\alpha t)$. D'après la définition de la transformée de Fourier des signaux analogiques, un changement d'échelle uniforme dans un domaine, correspond à un changement inverse dans le domaine transformé.

$$y_a(t) = x_a(\alpha t) \longleftrightarrow Y_a(\Omega) = \frac{1}{|\alpha|} X_a\left(\frac{\Omega}{\alpha}\right) \qquad (5.1)$$

Les transformations précédemment définies ne sont pas les solutions recherchées. Ainsi, il faut pouvoir:

- Produire un signal modifié (dilaté ou compressé) temporellement et conservant le support fréquentiel (conservant le spectre), (appelée aussi *dilatation-p*), ou
- Produire un signal modifié en fréquence et conservant le support temporel (conservant la durée) (appelée aussi *transposition-p*)**[37]** p45.

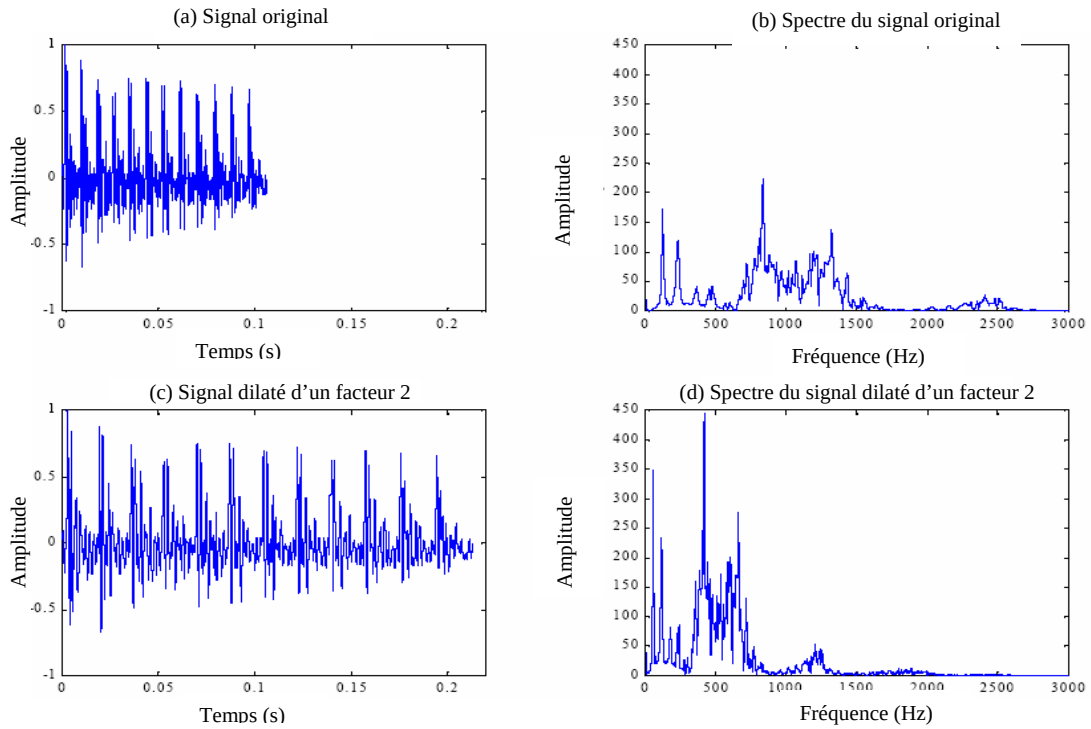


Fig 5.1 Illustration de la dualité entre le domaine fréquentiel et le domaine temporel. (a) signal original (b) spectre du signal original,

Un signal dilaté dans le domaine temporel (c) il est compressé dans le domaine fréquentiel

5.4 Fonction de modification de l'échelle temporelle

La modification de l'échelle temporelle peut être spécifiée par définition de $n \longrightarrow n' = \tau(n)$ entre l'échelle temporelle originale et l'échelle temporelle modifiée. Cette fonction est une sorte de "loi de correspondance" entre l'échelle temporelle originale et l'échelle temporelle modifiée. Cette fonction indique que l'évènement sonore, produit à l'instant n dans le signal original, devra être entendu à l'instant n' dans le signal modifié. Pour un facteur de modification (dilatation ou compression) $\alpha(t)$, où $\alpha(t) \succ 0$ est donné, la fonction de modification de l'échelle temporelle sera donnée par :

$$n \longrightarrow n' = \tau(n) = \frac{1}{T}\int_0^{nT} \alpha(u)du, \qquad (5.2)$$

où T est la période d'échantillonnage.

Pour $\alpha(t) \succ 1$, la modification de l'échelle temporelle correspond à une dilatation (ralentissement) du signal original, et lorsque $0 \prec \alpha(t) \prec 1$, la modification de l'échelle temporelle correspond à une compression (accélération) du signal original.

5.5 Techniques utilisées dans la modification de l'échelle temporelle

Les techniques de modification de l'échelle temporelle se classent habituellement selon 2 types : méthodes paramétriques et non-paramétriques (**Fig 5.2**). Cette distinction est faite selon que la représentation repose sur un modèle ou non.

La représentation paramétrique repose sur une utilisation explicite d'un modèle de signal (sinusoïdal par exemple), d'où sont tirés des paramètres. Il s'agit d'estimer au mieux l'ensemble de ces paramètres potentiellement variables au cours du temps, symbolisés par le vecteur $\vec{P}(t)$, afin que le signal synthétisé à partir des paramètres d'analyse non modifiés soit le plus "proche" possible du signal original (au sens de la perception sonore).

La représentation non-paramétrique ne repose pas sur l'utilisation explicite d'un modèle. Elle consiste en un ensemble de données, qui sont issues de la décomposition sur une famille de vecteurs bien choisis. La transformation mathématique associée porte généralement le nom de "transformée".

De manière générale et dans le cas d'un signal arbitraire, les représentations paramétriques perdent de l'information, alors que ce n'est pas le cas des représentations non paramétriques. Dans notre étude on s'intéresse uniquement aux méthodes non-paramétriques.

Les méthodes non-parametriques peuvent être classifiées, selon le domaine de traitement, en trois classes, nous distinguons les *méthodes temporelles* les *méthodes fréquentielles* et les méthodes *temps-fréquence.*

Nous nous concentrons dans cet exposé sur les méthodes temporelles.

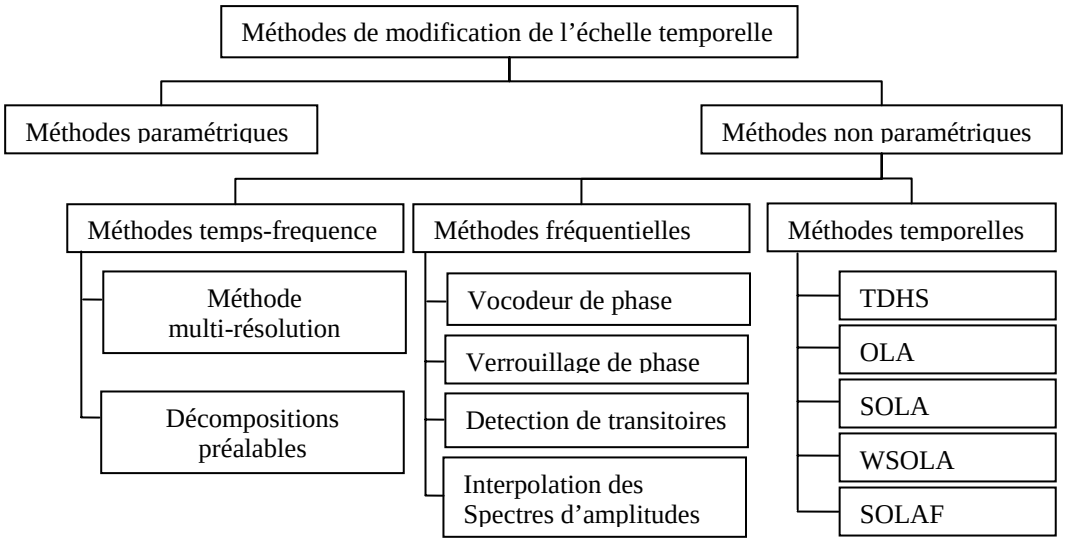


Fig 5.2 Les différentes techniques utilisées dans la modification de l'échelle temporelle.

5.6 Transformée de Fourier à Court Terme TFCT (en anglais STFT)

La transformée de Fourier d'un signal x(n) est donnée par

$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x(n)\, e^{-j\omega n}\,, \qquad (5.3)$$

est la représentation fréquentielle la plus employée. Si on considère la parole comme étant un signal quasi-stationnaire, nous pouvons alors appliquer une analyse à court terme avec la transformation de Fourier pour obtenir une transformation de Fourier à court terme (STFT) comme une représentation temps-frequence **[38] p48 et [39] p99 [40].**

La transformée de Fourier à court terme (STFT) d'un signal *x(n)* est définie par la segmentation du signal en utilisant une fenêtre *w(n)*

$$x_w(n,m) = w(n)\,x(n+m)\,, \qquad (5.4)$$

Prenons ensuite sa transformée de Fourier

$$X(\omega,m) = \sum_{n=-\infty}^{\infty} x(n+m)w(n)e^{-j\omega n}\,, \qquad (5.5)$$

5.7 Méthode de synthèse par recouvrement-adition (Overlap-Add)

Considérons un signal x(n), et soit $X(\omega,m)$ sa TFCT. Si ce signal est modifié pour obtenir une modification temporelle (time scaling), un autre signal $\hat{Y}(\omega,n)$ est donc produit, sa TFCT inverse, si elle existe, elle est différente de *x(n)*. La formule de synthèse qui fournit une valeur correcte de $\hat{Y}(\omega,n)$ avec une TFCT inverse valide, repose sur la technique Recouvrement-addition (ovelap-addition OLA) introduite par Griffin et Lim. Dans cette méthode, on construit y(n) de telle façon que sa TFCT $Y(\omega,n)$, soit au maximum près de $\hat{Y}(\omega,n)$ dans le sens des moindres carrés, c.-à-d., tel que l'erreur quadratique totale

$$E = \sum_k \frac{1}{2\pi} \int_{-\pi}^{+\pi} \left|\hat{Y}(\omega,k) - Y(\omega,k)\right|^2 d\omega\,, \quad (5.6)$$

est minimisée sur tous les signaux y(n) (la somme est sur tous les instants k pour lesquels $\hat{Y}(\omega,n)$ est définie).

Le théorème de Parceval permet de réécrire l'équation (5.6) comme suit

$$E = \sum_{k} \sum_{m=-\infty}^{+\infty} (\hat{y}_w(m,k) - y(m+k)w(m))^2 , \qquad (5.7)$$

où $\hat{y}_w(m,k)$ est la transformée de Fourier inverse de $\hat{Y}(\omega,k)$. Le signal *y(n)* qui minimise '*E*' est obtenu par

$$\frac{\partial E}{\partial y(n)} = -2\sum_{k} (\hat{y}_w(n-k,k) - y(n)w(n-k))w(n-k) = 0 , \qquad (5.8)$$

D'où

$$y(n) = \frac{\sum_{k} w(n-k)\hat{y}_w(n-k,k)}{\sum_{k} w^2(n-k)}, \qquad (5.9)$$

où

$$\hat{y}_w(n-k,k) = \frac{1}{2\pi}\int_{-\pi}^{+\pi} \hat{Y}(\omega,k)e^{j\omega(n-k)}d\omega , \qquad (5.10)$$

est la transformée de Fourier inverse de $\hat{Y}(\omega,k)$ décalée par k échantillons. La formule de synthèse OLA reconstruit le signal original, si $X(\omega,m)$ est une TFCT valide, si non, elle reconstruit le signal dont la TFCT est au maximum proche de $X(\omega,m)$ dans le sens des moindres carrés. En outre, le dénominateur de l'équation (5.9) n'est utile que pour compenser la pondération non uniforme des échantillons lors de la procédure de fenêtrage. L'opération de synthèse peut être simplifiée si la fonction de fenêtrage et les instants de synthèse k sont choisis tels que

$$\sum_{k} w^2(n-k) = 1 , \qquad (5.11)$$

Ils existent plusieurs possibilités qui satisfont cette condition de simplification, la plus employée est la fenêtre de Hanning avec une superposition de 50% entre les segments successifs.

5.8 Techniques de modification de l'échelle temporelle

5.8.1 Recouvrement-Addition (OLA)

Avec la synthèse OLA, Il est possible de réaliser des modifications de l'échelle temporelle, en n'utilisant que des opérations dans le domaine temporel. En effet, si on adopte la stratégie d'analyse à court terme pour construire $X(\omega,m)$ et si on utilise le critère de recouvrement-addition pour synthétiser le signal *y(n)* à partir de la représentation modifiée $\hat{Y}(\omega,m) = M_{xy}[X(\omega,m)]$, on obtiendra toujours des algorithmes de modification qui peuvent être opérés dans le domaine temporel si l'opérateur de modification $M_{xy}[.]$ ne dépend que de l'index de temps m (où m est l'instant d'analyse $t_a(u)$ et l'opérateur $M_{xy}[.]$ est associé avec la fonction de modification temporelle (time warping) qui égale à $\tau^{-1}(m)$) tel que:

$$\hat{Y}(\omega,m) = [X(\omega, M_{xy}[m])] \qquad (5.12)$$ modification,

$$\hat{y}(n,m) = [x_w(n, M_{xy}[m])] \qquad (5.13)$$ transformée de Fourier inverse,

$$y(n) = \frac{\sum_m w(n-m)x_w(n-m, M_{xy}[m])}{\sum_m w^2(n-m)} \qquad (5.14)$$ synthèse OLA .

il est claire de l'équation (5.14) que la modification est obtenue par un coupage des segments $x_w(n, M_{xy}[m])$ à partir du signal d'entrée en utilisant des fenêtres, puis les repositionner tout au long de l'axe temporel avant de construire le signal de sortie par recouvrement-addition des segments pondérés. Cependant, la **Fig 5.3 (b)** montre que, si on applique les formules citées ci-dessus pour réaliser une modification temporelle (time warping) $\tau(m)$ du signal, la périodicité du signal modifié est changée par rapport au signal original **Fig 5.3 (a)**. Ainsi, des mauvais résultats sont généralement obtenus en utilisant $\hat{Y}(\omega,m) = X(\omega,\tau^{-1}(m))$.

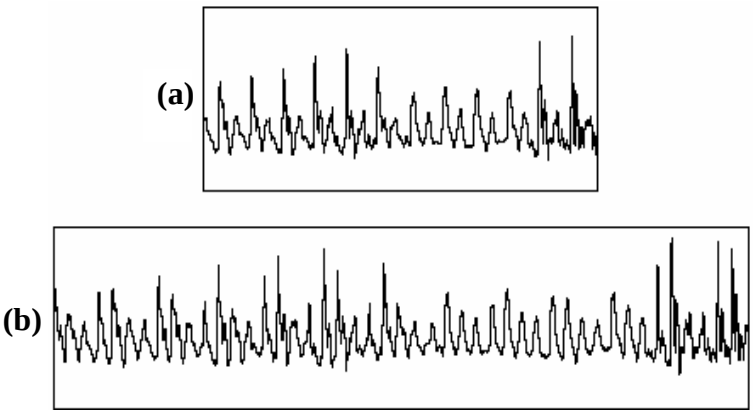


Fig 5.3 Echec de la méthode OLA basée sur TFCT de reproduire la structure quasi-périodique du signal original (a) à sa sortie (b)

5.8.2 Recouvrement-addition synchronisé (SOLA)

L'algorithme de Recouvrement-addition synchronisé (SOLA) a été développé par Roucos et Wilgus, cet algorithme est beaucoup plus efficace en terme de puissance de calcul, et semble donner des résultats au moins aussi bons que ceux de Griffin et Lim pour la parole **[38]**.

L'algorithme SOLA divise la procédure de modification de l'échelle temporelle en deux étapes, l'analyse et la synthèse, l'étape d'analyse consiste à fenêtrer le signal d'entrée chaque ' Sa' (shift analysis) échantillons (**Fig 5.4**). L'étape de synthèse consiste au recouvrement-addition des fenêtres (L_w est la longueur de la fenêtre, qui est fixe et multiple de la période du fondamentale) issues de l'étape d'analyse. Chaque nouvelle fenêtre est alignée pour maximiser la corrélation avec la somme des fenêtres précédentes avant d'être ajoutée. Ceci réduit les discontinuités résultantes des différents intervalles inter-trames utilisés pendant l'analyse et la synthèse. Le signal modifié résultant est exempt des clicks et de pops. La figure **Fig 5.4** illustre un exemple de dilatation temporelle d'un signal en utilisant l'algorithme SOLA.

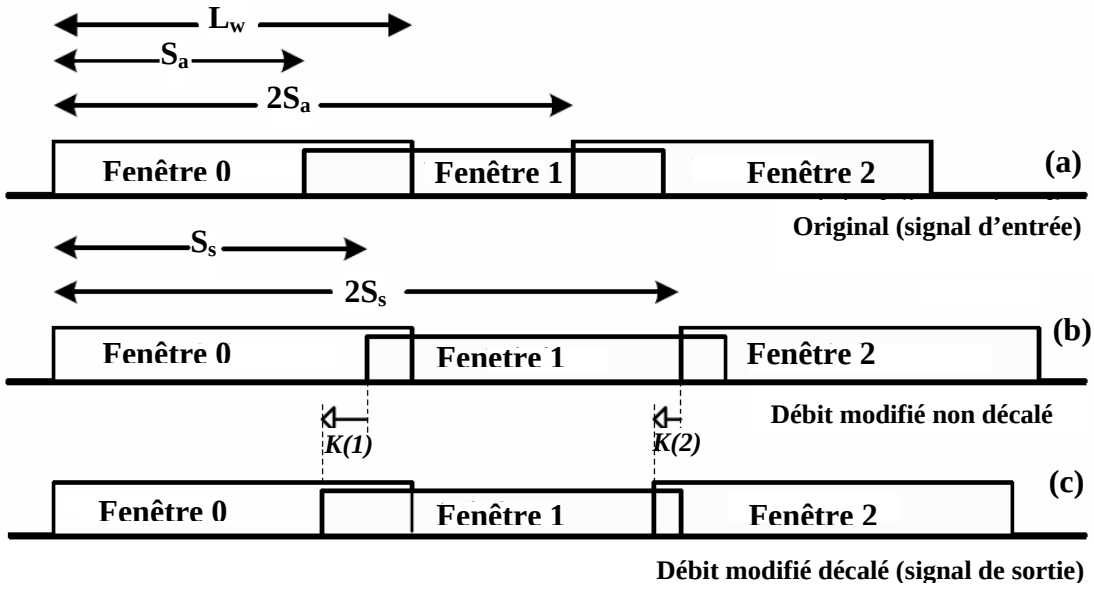


Fig 5.4 Modification de l'échelle temporelle (dilatation) par la méthode SOLA.

Dans la méthode SOLA, les fenêtres sont additionnées d'une manière synchrone avec la période locale. Le signal modifié *y(n)* obtenu par Recouvrement-addition synchronisé des segments fenêtrés $x_w(n) = w(n)x(n)$ (où *x(n)* est le signal d'entrée et *w(n)* est la fonction fenêtre), est donne par :

1- initialisation des signaux $y_w(n)$ et $r(n)$:

$$\left.\begin{aligned} y_w(n) &= x_w(n) \\ r(n) &= w(n) \end{aligned}\right\}, \text{ pour } n = 0...L_w - 1 \qquad (5.15)$$

2- mise à jour des signaux $y_w(n)$ et $r(n)$ pour chaque nouvelle trame du signal d'entrée, $x_w(n)$, comme suit :

$$y_w(mS_s - k(m) + j) = \begin{cases} y_w(mS_s - k(m) + j) + x_w(mS_a + j) & \text{pour } 0 \le j \le L_m - 1 \\ x_w(mS_a + j) & \text{pour } L_m \le j \le L_w - 1 \end{cases} \qquad (5.16)$$

où L_m est le nombre de points de recouvrement entre la nouvelle fenêtre $x_w(mS_a + j)$ et la séquence existante $y_w(mS_s - k(m) + j)$ pour la trame courante m.

$$r(mS_s - k(m) + j) = \begin{cases} r(mS_s - k(m) + j) + w(mS_a + j) & \text{pour } 0 \le j \le L_m - 1 \\ w(mS_a + j) & \text{pour } L_m \le j \le L_w - 1 \end{cases} \qquad (5.17)$$

$$k(m) = \max R_{xy}^m(k),$$

$$R_{xy}^m(k) = \frac{\sum_{j=0}^{L_m-1} y_w(mS_s - k + j)x_w(mS_a + j)}{\sqrt{\left[\sum_{j=0}^{L_m-1} y_w^2(mS_s - k + j)\right]\left[\sum_{j=0}^{L_m-1} x_w^2(mS_a + j)\right]}} \qquad (5.18)$$

3- normalisation de $y_w(n)$ par *r(n)* pour obtenir la sortie finale y(n) :

$$y(j) = \frac{y_w(j)}{r(j)}, \qquad \text{pour tout les j.} \qquad (5.19)$$

Conformément aux équations ci-dessus, *k(m)*>0 correspond à un décalage vers l'arrière le long de l'axe temporel de la trame de rang m, qui maximise l'intercorrélation normalisée $R_{xy}^m(k)$ entre la fenêtre de rang m et le signal décalé à débit modifié composé des fenêtres 0,…,(m-1). L_w est le nombre de points de données dans chaque trame $x_w(mS_a + j)$.

La maximisation de l'intercorrelation permet d'ajouter et de normaliser le segment temporel courant dans la région la plus similaire du signal reconstruit. L'opération de décalage garantie une préservation de la périodicité d'amplitude dans le signal à débit modifié. Ce signal est appelé signal décalé à débit modifié **Fig 5.4** (c), pour le distinguer du signal non décalé à débit modifie **Fig 5.4** (b), qui est obtenu par une simple opération de recouvrement-addition. Comme on l'a déjà vu au paragraphe précédent, la synthèse OLA basée sur la TFCT sous-échantillonnée $\hat{Y}(\omega, kS) = X(\omega, \tau^{-1}(kS))$ produit un signal

$$y_1(n) = \frac{\sum_k w^2(n - kS)x(n - kS + \tau^{-1}(kS))}{\sum_k w^2(n - kS)} \qquad (5.20)$$

fortement distordu (**Fig 5.3**), *'S'* est un facteur de sous-échantillonnage, il est introduit pour réduire la quantité d'information à traiter.

Afin d'éviter des discontinuités des périodes du fondamental ou des sauts de phase au niveau des raccordement entre les segments, Roucos et Wilgus ont proposé de réaligner chaque segment d'entrée avec la partie déjà formée du signal de sortie avant d'effectuer l'opération de

recouvrement-addition. Ainsi, l'algorithme SOLA synthétise le signal à échelle temporelle modifiée

$$y(n) = \frac{\sum_k v(n - kS + \Delta_k)x(n - kS + \tau^{-1}(kS) + \Delta_k)}{\sum_k v(n - kS + \Delta_k)} \qquad (5.21)$$

de gauche à droite avec une fonction de fenêtrage *v(n)* et un facteur de décalage $\Delta_k \in [-\Delta_{\max} ... \Delta_{\max}]$ choisi d'une manière à maximiser l'intercorrelation entre le segment temporel courant $v(n - kS + \Delta_k)x(n - \tau^{-1}(kS) - kS + \Delta_k)$ et la portion du signal de sortie déjà constituée

$$y(n, k-1) = \frac{\sum_{l=-\infty}^{k-1} v(n - lS + \Delta_l)x(n + \tau^{-1}(lS) - lS + \Delta_l)}{\sum_{l=-\infty}^{k-1} v(n - lS + \Delta_l)} \qquad (5.22)$$

L'algorithme SOLA est efficace, en terme de temps de calcul, il ne nécessite pas d'itérations et opère dans le domaine temporel. Le traitement dans le domaine temporel implique que la modification de la TFCT n'affecte que l'axe temporel $\hat{Y}(\omega, kS - \Delta_k) = X(\omega, \tau^{-1}(kS))$.

Le paramètre de décalage Δ_k implique une tolérance sur la fonction de modification temporelle : afin d'assurer un recouvrement-addition synchronisé des segments, la fonction de modification temporelle $\tau(n)$ désirée ne doit pas être réalisée exactement. Une déviation dans l'ordre de la période du fondamental est permise.

5.8.3 L'ajout en chevauchement des similarités des ondes WSOLA

En 1993, Verhelst et Roelands proposent une méthode qu'ils nomment WSOLA (Waveform Similarity OLA) **[2][41][42]**. Cette méthode s'inspire fortement des méthodes OLA et SOLA, et sa théorie est également développée dans un contexte temps-fréquence.

Un signal synthétisé par la méthode WSOLA, possède, à l'exception du changement de la durée, les mêmes propriétés acoustiques que le signal original. Ceci est réalisé par la recherche des régions similaires du signal original x(n), puis le recouvrement de ces régions à l'aide de la fonction de modification temporelle $\tau(n)$. Nous obtiendrons ainsi un signal à échelle temporelle modifiée y(n). On peut exprimer ça mathématiquement par l'expression suivante:

$$\forall m : y(n + \tau(m))w(n) = x(n + m)w(n) \qquad (5.23)$$

où *w(n)* est la fonction de fenêtrage et (=) dénote le sens de 'similarité maximale'. $\tau(m)$ est la fonction de modification temporelle, elle fournit la modification temporelle du signal. Le signal synthétisé de sortie, qui présente une similarité locale maximale suivant la définition précédente est donné par:

$$y(n) = \frac{\sum_k w^2(n - S_k)\, x(n + \tau^{-1}(S_k) - S_k + \Delta_k)}{\sum_k w^2(n - S_k)} \qquad (5.24)$$

S_k représente les positions consécutives de la fonction fenêtre (instant de synthèse ou marques d'écriture), $\tau^{-1}(S_k) + \Delta_k$ représente l'instant d'analyse ou marque de lecture , les facteurs de décalage Δ_k, où $\Delta_k \in [-\Delta_{\max} ... \Delta_{\max}]$, sont calculés d'une manière à maintenir une similarité maximale avec le signal original (naturel) au niveau des régions de jointure. Autrement dit, la méthode WSOLA assure une continuité suffisante, au niveau des jointures du signal synthétisé en exigeant un maximum de similarité avec la continuité naturelle existante dans le signal d'entrée.

En se basant sur cette idée, plusieurs variétés d'implémentations pratiques peuvent être réalisées. L'opération de la version de base de la méthode WSOLA est illustrée dans la figure (**Fig 5.5**).

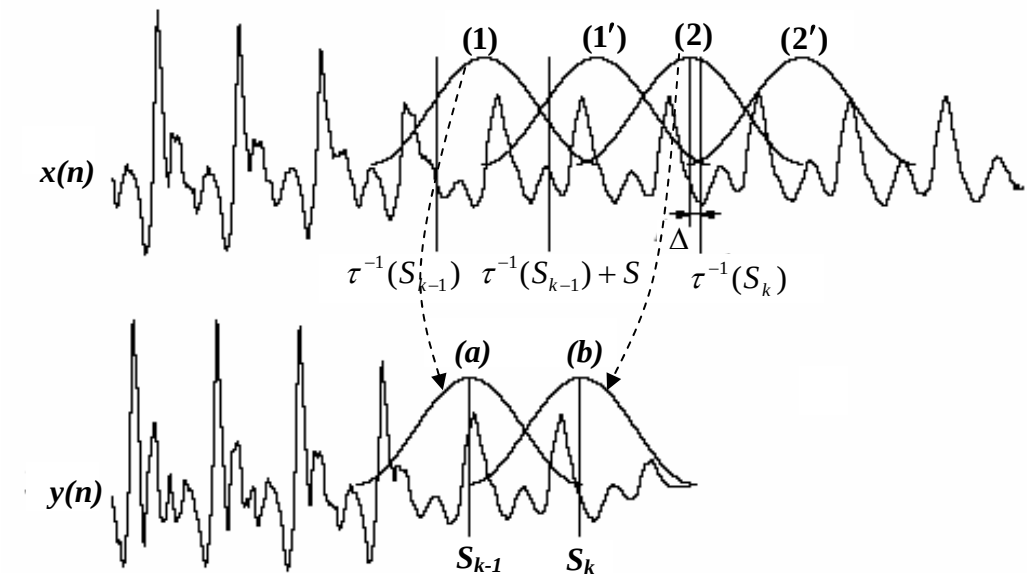


Fig 5.5 Illustration de l'algorithme WSOLA

(1') et (2'): Segments modèles, (1) et (2): Segments élus, (a) et (b) segments synthétisés.

En choisissant un espacement régulier entre les instants de synthèse $S_k = kS$ et une fenêtre symétrique tel que $\sum_k (w^2(n-kS) = 1$ (5.25)

L'équation de synthèse (5.24) se simplifie alors à :

$$y(n) = \sum_k w^2(n-kS)x(n+\tau^{-1}(kS)-kS+\Delta_k) \qquad (5.26)$$

Soit α le facteur de modification (dilatation /compression), avec $\alpha \succ 0$. La fonction de modification temporelle inverse sera donnée par

$$\tau^{-1}(KS) = kS/\alpha . \qquad (5.27)$$

En remplaçant l'équation (5.27) dans (5.26), on obtient :

$$y(n) = \sum_k w^2(n-kS)x(n+kS/\alpha-kS+\Delta_k) \qquad (5.28)$$

De l'équation (5.28), on déduit que :

Si $\alpha \prec 1$, le signal de sortie *y(n)* subit une compression (accélération).

Si $\alpha \succ 1$, le signal de sortie *y(n)* subit une dilatation (ralentissement).

5.9 L'algorithme WSOLA

En se référant à la figure **Fig 5.5**, et en effectuant un traitement de gauche vers la droite, on suppose que le segment (1) soit le dernier segment (élu), ayant été coupé du signal d'entrée et ajouté au signal de sortie à l'instant $S_{k-1} = (k-1)S$, c.à.d que le segment (synthétisé de sortie) (a) = segment (d'entrée) (1). WSOLA a maintenant besoin de rechercher un segment (b) qui sera recouvert-additionné avec (a) d'une manière synchronisée et que l'on coupe du signal d'entrée autour de l'instant $\tau^{-1}(kS)$.

Soit (1') le segment qui recouvre naturellement le segment (1) dans le signal original *x(n)*, (1') est appelé aussi modèle (*template*), WSOLA sélectionne le segment (b) d'une manière à ce qu'il ressemble le plus que possible au segment (1') et qu'il soit situé autour de l'instant $\tau^{-1}(kS)+\Delta_k$ où $\Delta_k \in [-\Delta_{\max} ... \Delta_{\max}]$.

Soit (2) le segment élu, c.à.d le segment similaire à (1'), sa position correspond au maximum de l'intercorrelation (ou au minimum de la fonction AMDF) entre la séquence d'échantillons du segment (1') et le signal d'entrée. Après recouvrement-addition du segment (b) avec (a), WSOLA entame le prochain segment, cette fois ci le segment (2') qui recouvre le segment (2), joue le rôle de modèle, même rôle que (1') dans l'étape précédente.

La figure **Fig 5.6** illustre avec plus de détails, comment déterminer la position du segment élu *m*, en cherchant la valeur $\delta = \Delta_m$ appartenant à la région de tolérance $[-\Delta_{\max} ... \Delta_{\max}]$ autour de $\tau^{-1}(mS)$ et qui maximise la mesure de similarité choisie $c(m,\delta)$ entre la portion du signal qu'on désir avoir avec lui une continuité naturelle, et le segment *m*-1 choisi précédemment.

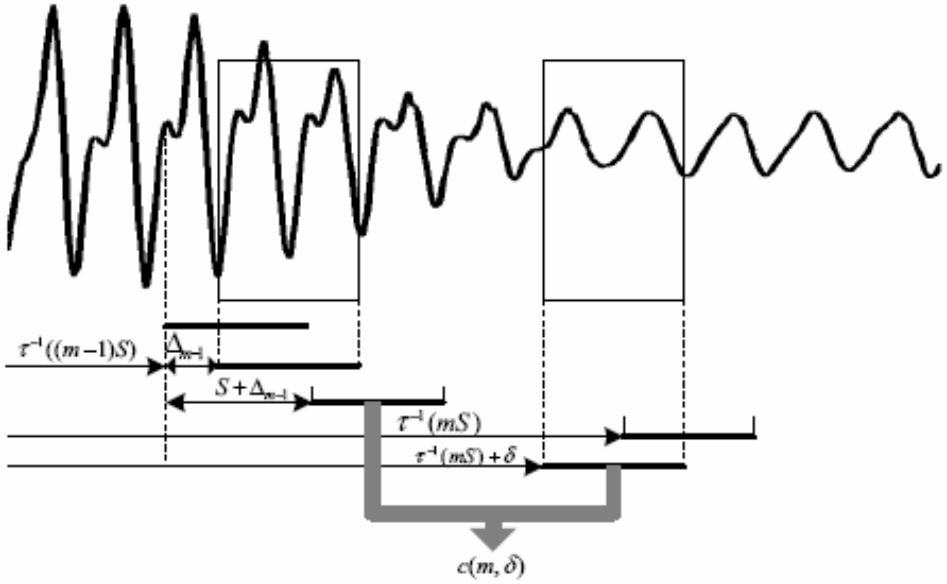


Fig 5.6 Illustration de la segmentation d'un signal basée sur la similarité dans la méthode WSOLA

5.10 Fonctions de mesure de similarité

Plusieurs fonctions peuvent être utilisées pour le calcul de la similarité, les plus employées sont :

Fonction d'intercorrelation

$$c_c(m,\delta) = \sum_{n=0}^{N-1} x(n+\tau^{-1}((m-1)S)+\Delta_{m-1}+S) \times x(n+\tau^{-1}(mS)+\delta), \qquad (5.29)$$

Fonction d'intercorrelation normalisée

$$c_n(m,\delta) = \frac{c_c(m,\delta)}{\left(\sum_{n=0}^{N-1} x^2(n+\tau^{-1}(mS)+\delta)\right)^{1/2}} \qquad (5.30)$$

Fonction AMDF

$$c_A(m,\delta)=\sum_{n=0}^{N-1}\left|x(n+\tau^{-1}((m-1)S)+\Delta_{m-1}+S)-x(n+\tau^{-1}(mS)+\delta)\right|, \qquad (5.31)$$

où N représente la longueur de la fenêtre,

5.11 Comparaison de WSOLA avec les méthodes SOLA et TD-PSOLA

	TD-PSOLA	**SOLA**	**WSOLA**
Méthode de synchronisation	Dépend de pitch	Similarité en sortie	Similarité en entrée
Longueur de la fenêtre effective	Adaptée selon le pitch	Fixe (> 4.pitch)	Fixe
Normalisation Dénominateur = constant	Non	Non	Oui
Algorithme et Efficacité de calcul	Faible	Haute	Très haute
Robustesse	Faible	Haute	Haute
Qualité de parole	Haute	Haute	Haute
Modification de pitch	oui	Non	Non

Tableau 5.1 Comparaison de WSOLA avec les méthodes SOLA et TD-PSOLA.

5.12 Conclusion

La modification de l'échelle temporelle des signaux audio est une technique à usage multiple, allant d'une simple compression ou dilatation des signaux jusqu'au codage, nous avons présenté dans ce chapitre les différentes techniques de modification de l'échelle temporelle. Nous avons traité, avec plus détail, la méthode WSOLA qui est considérée comme la plus performante de ces méthodes et qui constitue l'objet de notre recherche.

Chapitre 6

Résultats et Evaluations

6.1 Introduction

Dans ce chapitre nous présentons l'implémentation de la méthode proposée, elle consiste en l'amélioration du mécanisme de récupération des trames perdues du codeur G.729. Nous effectuons une modification de l'échelle temporelle (Time-Scale Modification TSM) en utilisant la technique de l'ajout en chevauchement des similarités des ondes (WSOLA) pour reconstruire le signal d'excitation des trames perdues. Les différents résultats obtenus sont illustrés et commentés.

6.2 Principe de la méthode proposée

Contrairement au mécanisme de dissimulation des trames perdues du G.729 qui utilise les paramètres modifiés (LSP, pitch, gain) de la dernière bonne trame reçue. La technique proposée, mis à part les LSPs, ne tient pas compte de ces paramètres (pitch, gain). Il s'agit de dilater les trois dernières trames reçues pour récupérer la trame courante perdue **Fig 6.1**.

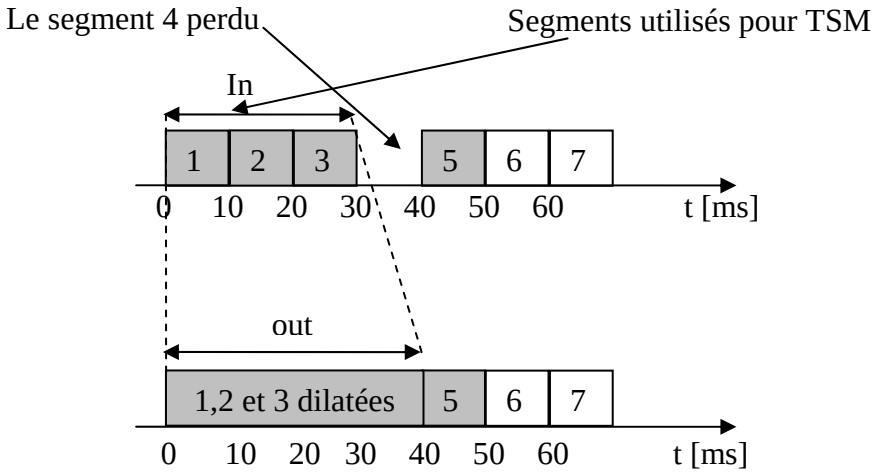


Fig 6.1 Dilatation des trois trames 1,2 et 3 pour récupérer la trame perdue 4.

6.3 Description de l'algorithme WSOLA

L'algorithme traite des trames de 10 ms soit 80 échantillons.

Pour récupérer une trame perdue, on fait une dilatation des trois dernières trames situant juste avant la perte. Pour ce faire, un certain nombre de buffers doit être mis en place, ces buffers sont les suivants:

History buffer : Contient les 4 dernières trames soit 40 ms ou 320 échantillons.

TSM buffer : Contient les 3 dernières trames du *history buffer*, soit 240 échantillons.

TSM out buffer : contient le signal dilaté, sa taille est variable, elle est toujours supérieure ou égale à 200 échantillons.

Past overlap buffer : ce buffer contient les 20 premiers échantillons du signal reconstruit *TSM out buffer*.

Out buffer : contient la trame (80 échantillons) de sortie qui remplace la trame perdue.

Une *fenêtre de Hanning* de 240 échantillons, utilisée pour le recouvrement addition.

Une *fenêtre triangulaire de Bartlett* utilisée pour le recouvrement addition de 40 échantillons.

Les figures **Fig 6.2** et **Fig 6.3** décrivent l'Algorithme de cette méthode en détail.

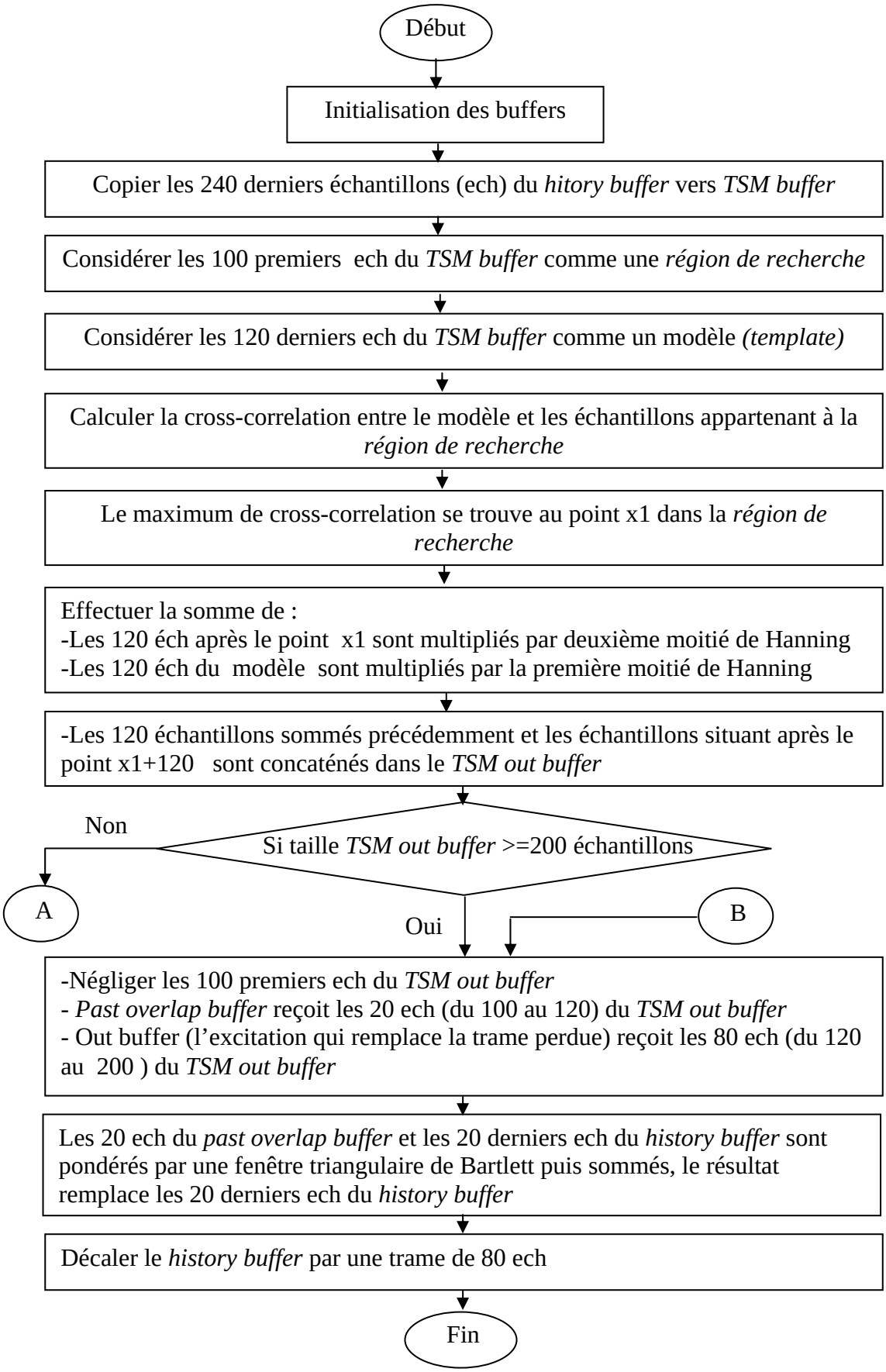


Fig 6.2 L'Algorithme WSOLA pour la récupération d'une trame perdue.

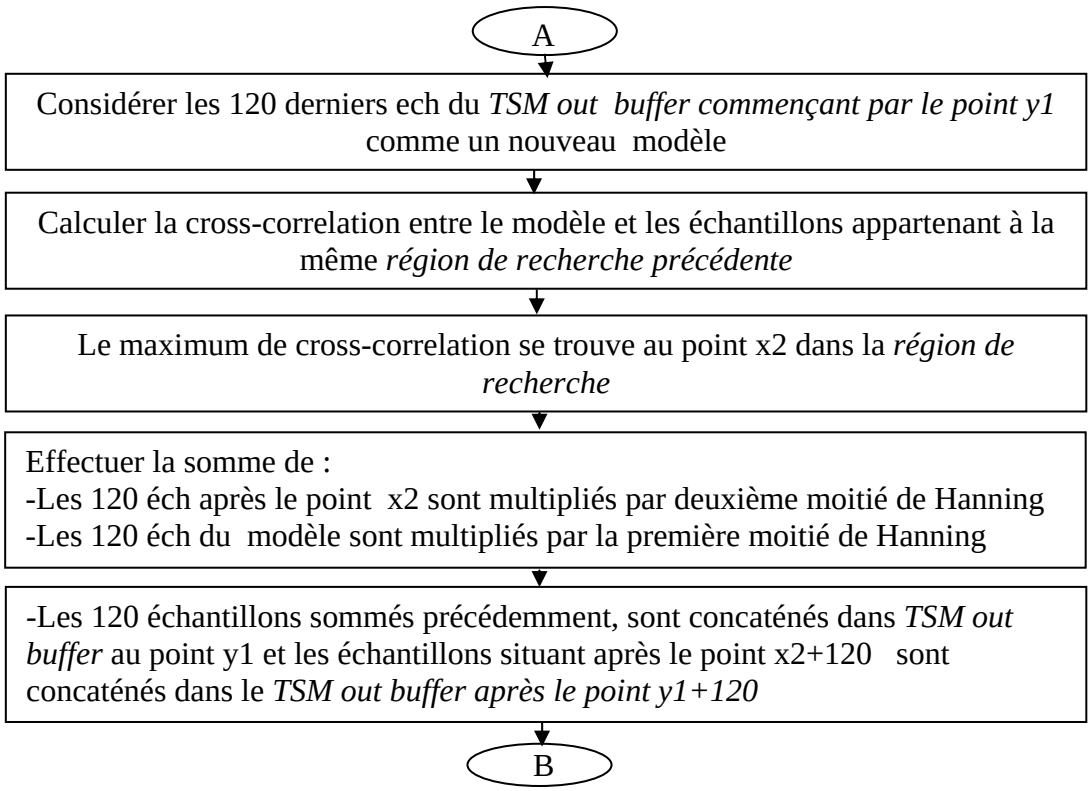


Fig 6.2 (suite) L'Algorithme WSOLA pour la récupération d'une trame perdue.

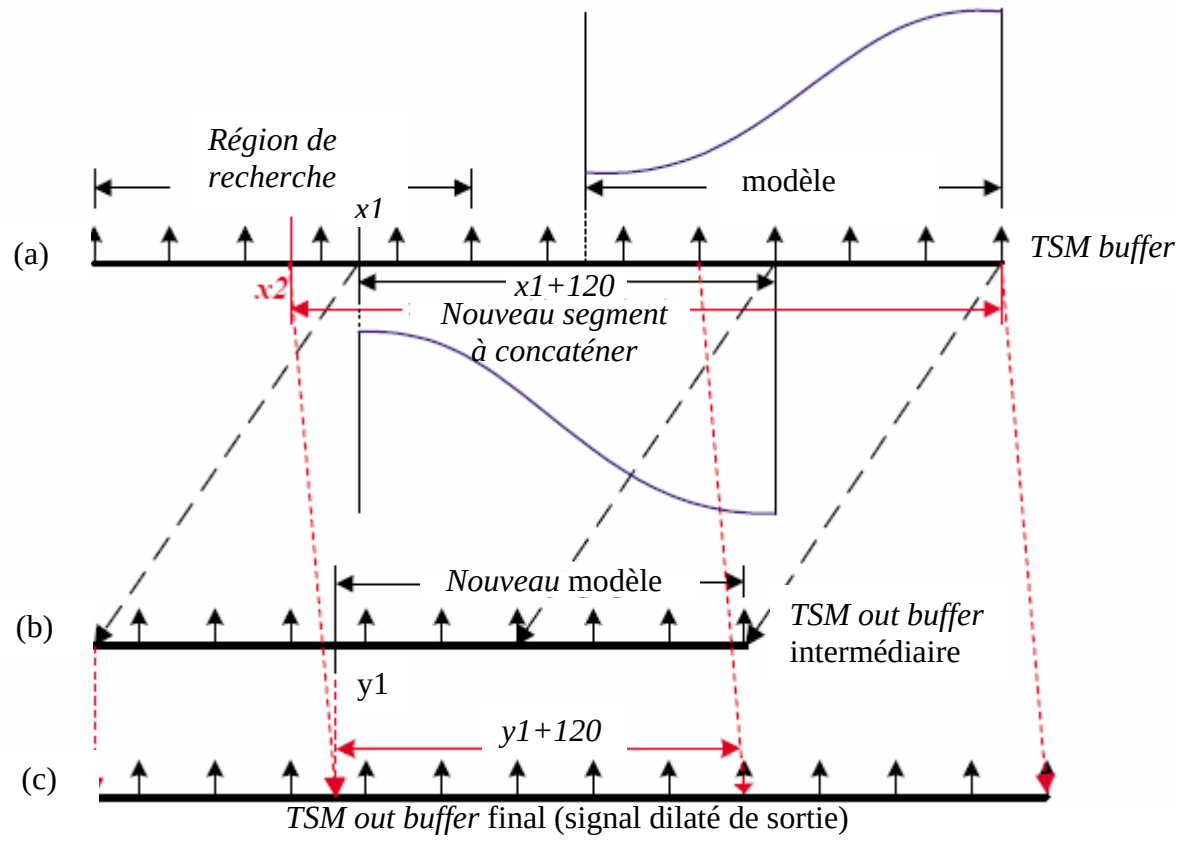


Fig 6.3 L'Algorithme de dilatation WSOLA, (a) *TSM buffer* (signal d'entrée), (b) *TSM buffer* obtenu après la première itération, le nombre des échantillons est inférieur à 200 (c) TSM out buffer, c'est le signal dilaté de sortie.

6.4 Intégration de WSOLA dans le codeur G.729

L'algorithme WSOLA à été intégré dans le G.729 pour générer l'excitation de la trame perdue, si une trame est déclarée perdue, le module WSOLA est activé tandis que le mécanisme de dissimulation des pertes noyé dans le G.729 est désactivé, à l'exception des LSP qui sont décodés pour synthétiser l'excitation. La figure **Fig 6.4** illustre le principe.

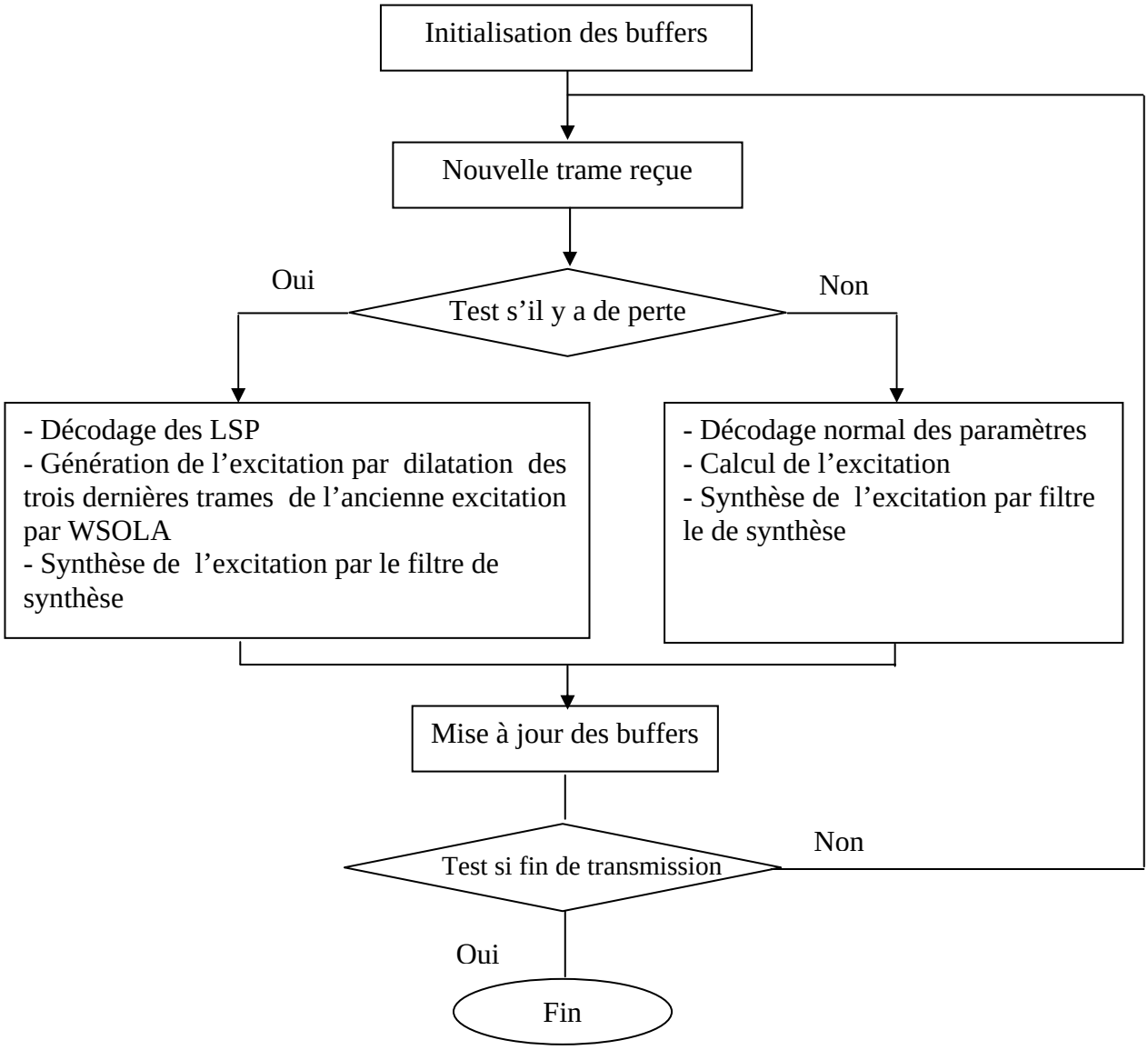


Fig 6.4 Algorithme de masquage basé sur la méthode WSOLA, proposé pour le CoDec G.729.

6.5 Outils de programmation, de tests et de simulations

Programmation : Les programmes ont été développés dans l'environnement *C++*.

Affichage des graphes: pour l'affichage des graphes, nous avons utilisé *MATLAB 6.5*.

Les tests Subjectifs : pour les tests subjectifs nous avons utilisé le PESQ **[43],** c'est une application qui évalue la qualité de la parole en donnant des valeurs MOS variant selon la qualité de -0.5 à 4.5.

Les tests objectifs : l'outil utilisé est l'*EMBSD* c'est une application qui mesure la distorsion spectrale entre deux signaux.

Visualisation et comparaison des signaux : *Cool Edit Pro 2.00*, c'est un environnement multitrack complet pour l'enregistrement, l'édition et le mixage des signaux audio pour WINDOWS. Il permet de visualiser, comparer et de convertir des signaux audio.

Simulation des pertes dans le réseau IP : L'une des modélisations de perte de paquets la plus répandue est le modèle de Gilbert qui représente un modèle de Markov à deux états, voir figure **Fig 6.5**. Dans ce modèle, l'état *L* représente une perte de paquets et l'état *S* représente une transmission avec succès. Soit *p* la probabilité de transition de l'état *S* vers l'état *L* et *q* la probabilité de transition de l'état *L* vers l'état *S*. En posant (1- *q*) > *p* il est alors possible de créer un canal perturbé et instable. Le taux d'erreur (de perte) sera la probabilité inconditionnelle de l'état *L* qui peut être exprimé par l'équation **[27]:**

$$\Pr(L) = \frac{p}{p+q} \quad (6.1).$$

Si *p*=*q*, le modèle de Gilbert se convertit en modèle de Bernoulli.

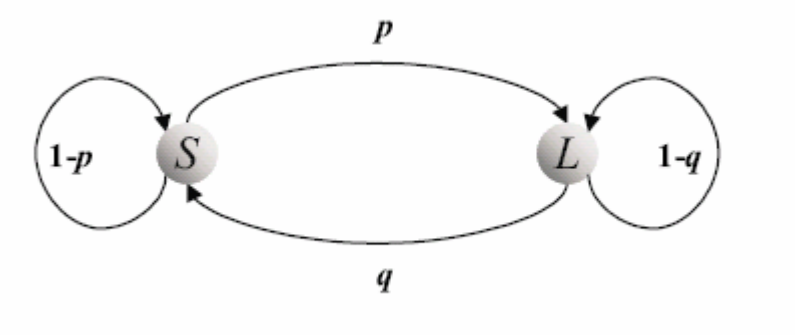


Fig 6.5 Modélisation de perte de paquets par le modèle de Gilbert.

Dans la simulation, en variant les valeurs de *p* et de *q*, nous avons effectué une série de tests dans lesquels on a changé le taux de perte de 0% jusqu'à 50%.

6.6 Description des signaux de parole utilisés dans les tests

Les signaux de test ont été pris à partir de la base de données TIMIT**[44]**, cette base à été préparée par le « National Institute of Standards and Technology » (NIST) et sponsorisée par « Defense Advanced Research Projects Agency - Information Science and Technology Office » (DARPA-ISTO) . Les instituts et laboratoires de recherche qui ont contribué au design du corpus de texte sont : Massachusetts Institute of Technology (MIT), Stanford Research Institute (SRI) et Texas Instruments (TI).

Cette base contient un total de 6300 phrases, 10 phrases prononcées par chacun des 630 orateurs des 8 régions du dialecte des États-Unis. La fréquence d'échantillonnage du signal parole des fichiers est de 8 kHz. Les phrases sont prononcées par des hommes et des femmes.

Les fichiers TIMIT utilisent la structure NIST SPHERE Header Structure, dans cette structure le fichier audio possède une entête contenant toutes les informations concernant le fichier, que se soit le signal audio ou son environnement, en effet, nous pouvons identifier à partir de l'entête, entre autres, la version de la base de données, le locuteur, la région du dialecte..etc. Les fichiers TIMIT ont un débit de 16000 bit/s, chaque échantillon est codé sur 16 bits.

Avant d'effectuer les tests, nous avons supprimé l'entête, pour ne garder que les échantillons de la parole, puis convertir les fichiers TIMIT de 16 Kb/s vers 8Kb/s.

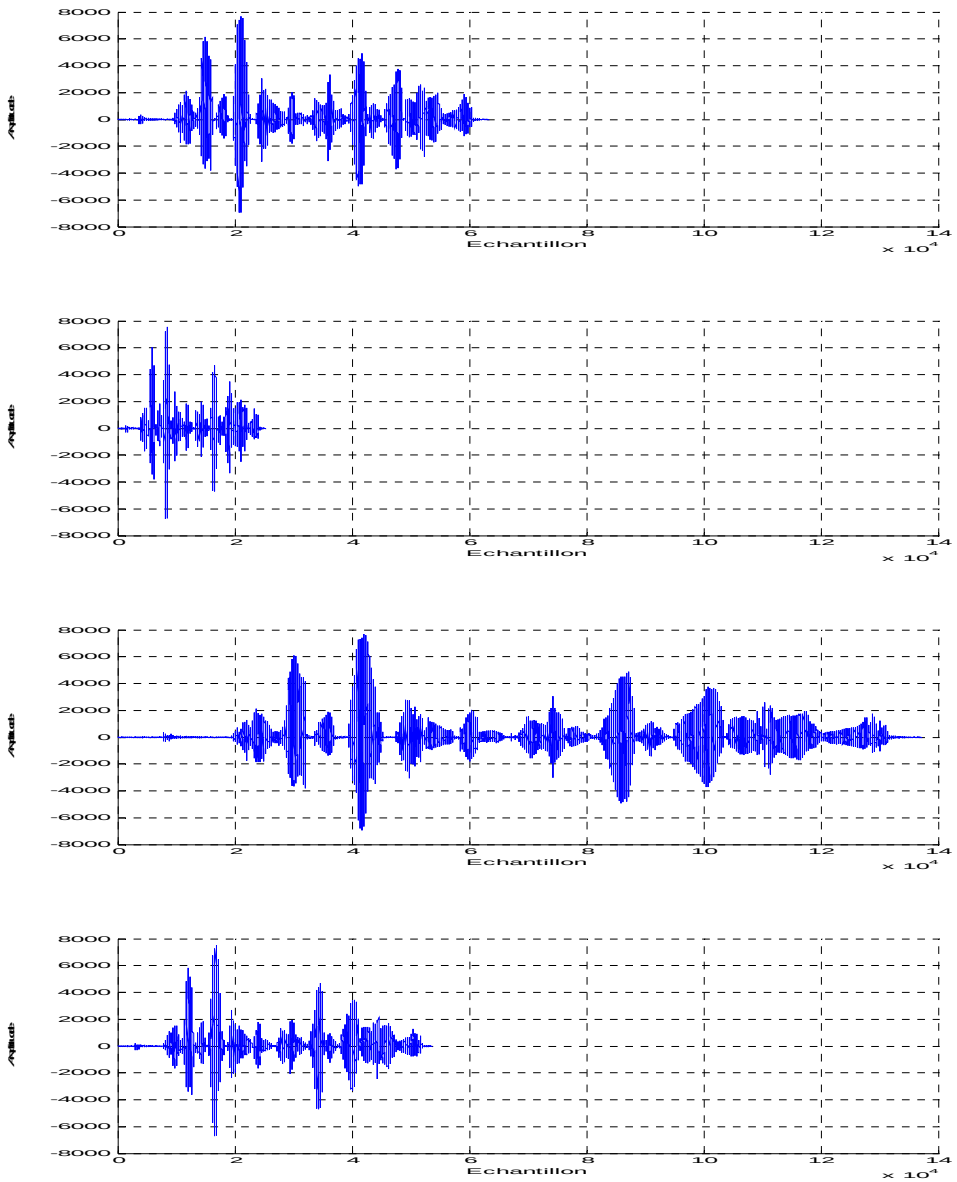
Afin de pouvoir varier les tests, nous avons choisi deux voix de sexe, de région et de contenu différents, ces signaux sont le *SA1* et *SX38,*

SA1 : représente une voix féminine « She had your dark suit in greasy wash water all year » enregistrée en 1986 et parlée par la dénommée *AKS0* né en 1957 appartenant à la première région dialecte (dr1).

SX38 : représente une voix masculine « Young people participate in athletic activities.» enregistrée en 1986 et parlée par le dénommé *SVS0* né en 1959 appartenant à la sixième région dialecte (dr6).

6.7 Résultats intermédiaires des fichiers compressés et dilatés par la méthode WSOLA

Le fichier utilisé est SA1 pris à partir da la base de données TIMIT.



(a)

(b)

(c)

(d)

Fig 6.6 Comparaison entre un signal original et des signaux compressés et dilatés par la méthode WSOLA. (a) signal original SA1, (b) signal compressé, (c) signal dilaté , (d) signal reconstruit après compression.

6.8 Résultats concernant la dissimulation des trames perdues

6.8.1 Procédure de simulation et de test de la méthode proposée

L'organigramme montre les différentes étapes et procédures pour tester la méthode proposée.

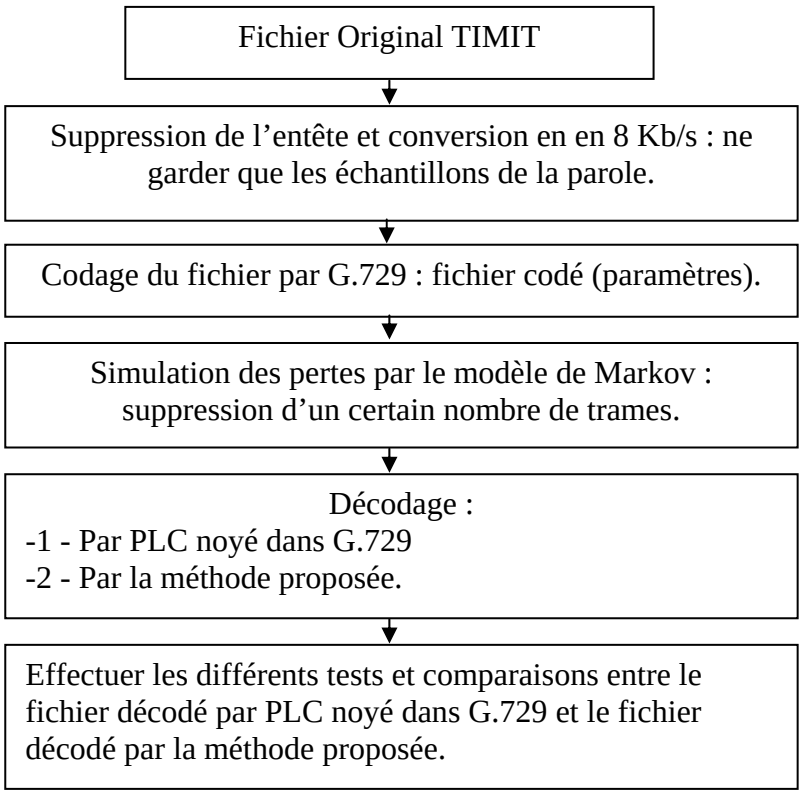


Fig 6.7 Les différentes étapes et procédures pour tester la méthode proposée.

6.8.2 Résultats obtenus en utilisant le fichier de test SA1.wav

PESQ

Taux de perte de paquets (FER) en %	PLC G.729	PLC proposé (WSOLA)
0	3.92	3.92
5.78	2.823	3.028
17.59	2.352	2.477
50	0.958	1.606

Tableau 6.1 Comparaison entre le PESQ du signal SA1 décodé par PLC noyé dans G.729 et le même signal décodé par la méthode proposée pour différents taux de perte

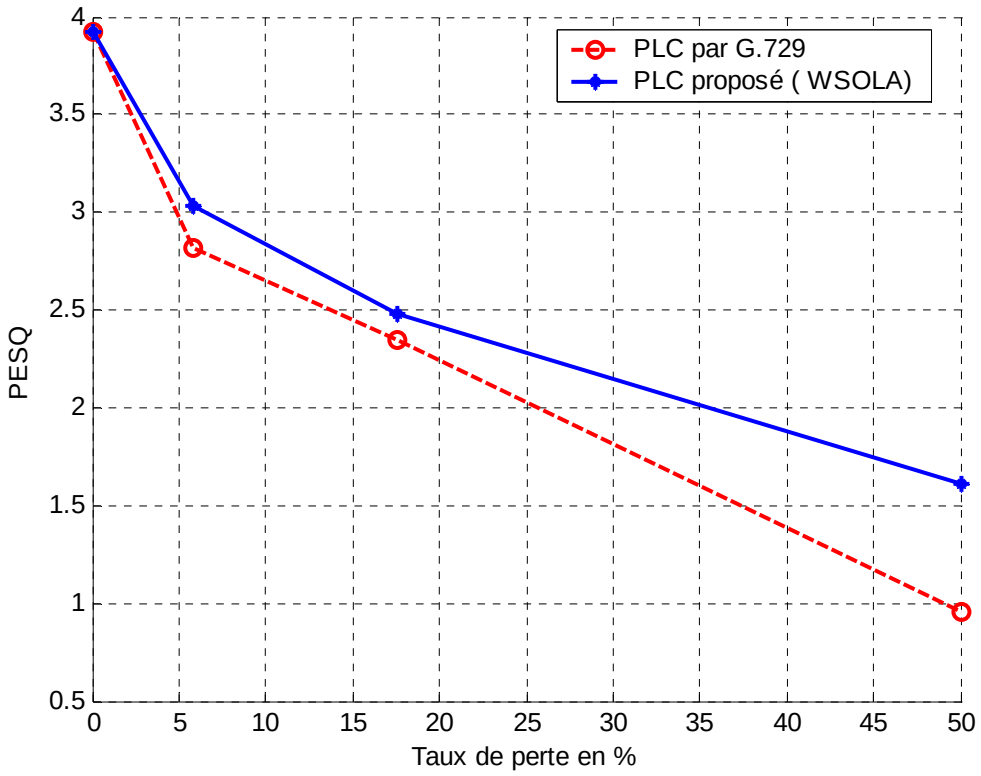


Fig 6.8 Comparaison entre le PESQ du signal SA1 décodé par PLC noyé dans G.729 (en pointillé) et le même signal décodé par la méthode proposée en (continu) pour différents taux de perte

EMBSD

Taux de perte de paquets (FER) en %	PLC G.729	PLC proposé (WSOLA)
0	0.859	0.859
5.78	1.247	1.074
17.59	1.947	1.612
50	2.660	2.535

Tableau 6.2 Comparaison entre l'EMBSD du signal SA1 décodé par PLC noyé dans G.729 et le même signal décodé par la méthode proposée pour différents taux de perte.

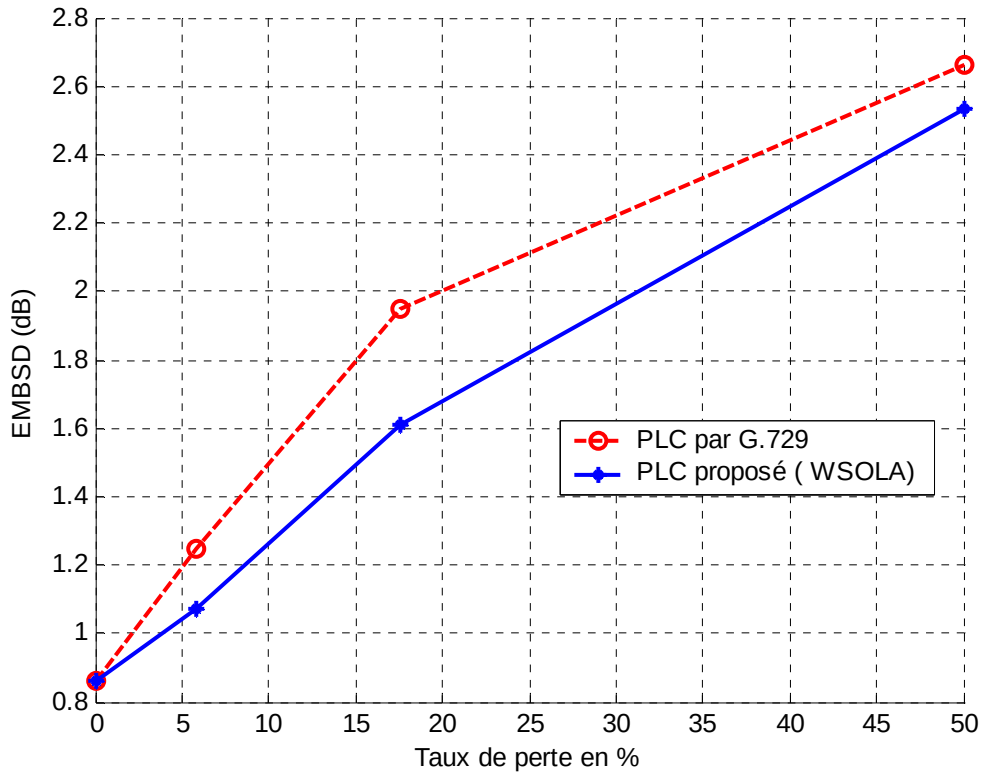


Fig 6.9 Comparaison entre l'EMBSD du signal SA1 décodé par PLC noyé dans G.729 (en pointillé) et le même signal décodé par la méthode proposée en (continu) pour différents taux de perte

6.8.3 Résultats obtenus en utilisant le fichier de test SX38.wav

PESQ

Taux de perte de paquets (FER) en %	PLC G.729	PLC proposé (WSOLA)
0	3.92	3.92
1.42	3.897	3.877
2.84	3.875	3.817
5.11	3.674	3.662
8.52	3.308	3.166
11.08	3.122	2.976
15.90	2.840	2.847
21.02	2.564	2.647
30.39	2.062	2.182
40.05	1.563	1.996
50.00	1.122	1.500

Tableau 6.3 Comparaison entre le PESQ du signal SX38 décodé par PLC noyé dans G.729 et le même signal décodé par la méthode proposée pour différents taux de perte

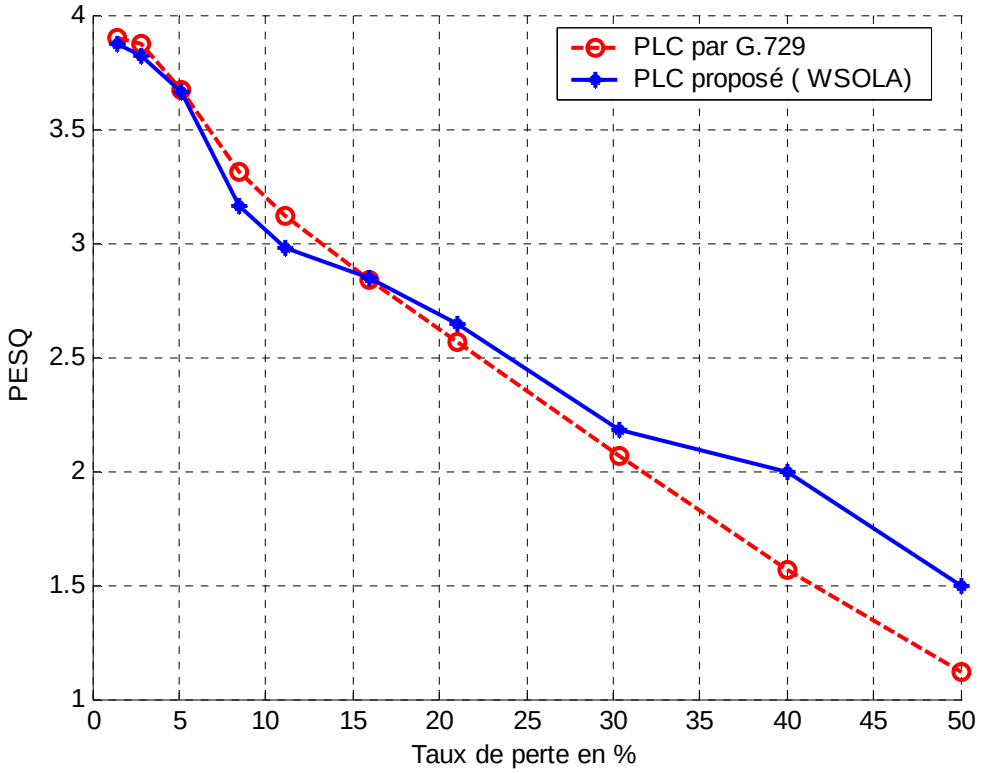


Fig 6.10 Comparaison entre le PESQ du signal SX38 décodé par PLC noyé dans G.729 (en pointillé) et le même signal décodé par la méthode proposée en (continu) pour différents taux de perte.

EMBSD

Taux de perte de paquets (FER) en %	PLC G.729	PLC proposé (WSOLA)
1.42	2.800	2.804
2.84	2.874	2.818
5.11	2.888	2.875
8.52	3.147	3.320
11.08	3.452	3.936
15.90	4.607	4.934
21.02	5.172	5.372
30.39	5.345	5.406
40.05	8.096	7.060
50.00	10.141	8.449

Tableau 6.4 Comparaison entre l'EMBSD du signal SX38 décodé par PLC noyé dans G.729 et le même signal décodé par la méthode proposée pour différents taux de perte

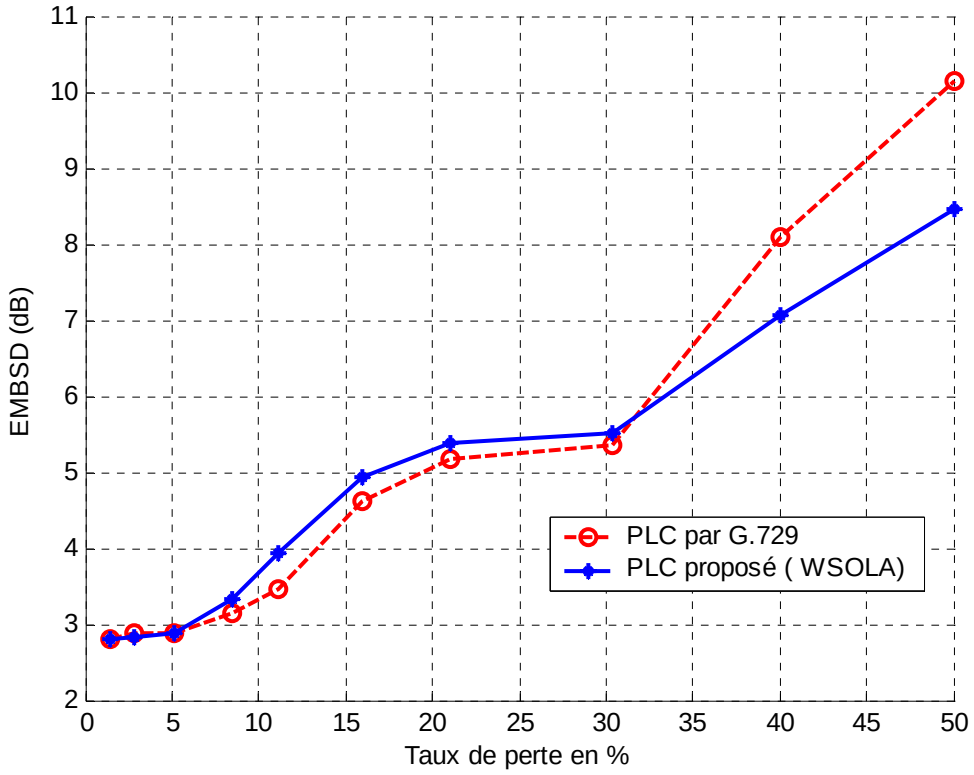
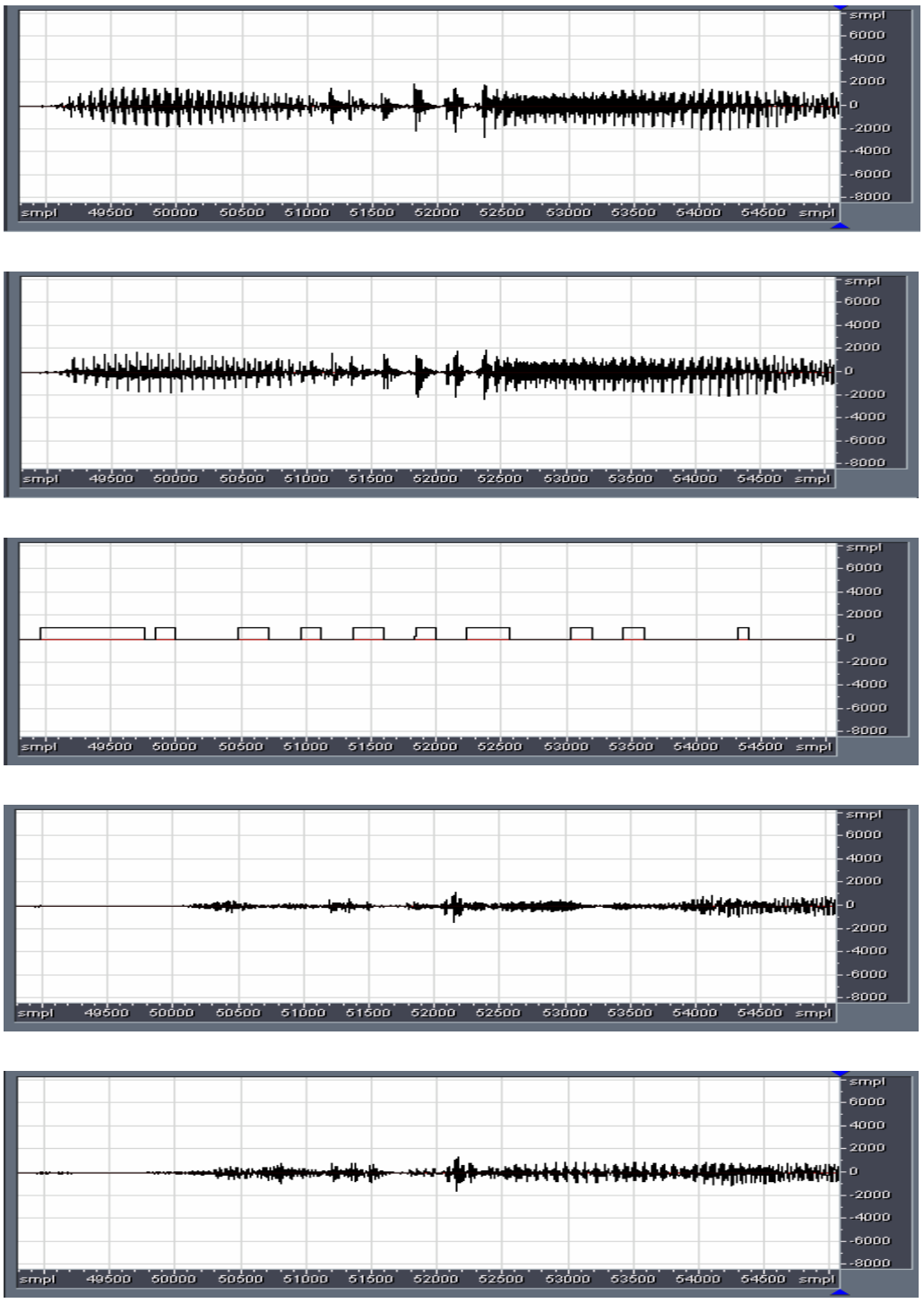


Fig 6.11 Comparaison entre l'EMBSD du signal SX38 décodé par PLC noyé dans G.729 (en pointillé) et le même signal décodé par la méthode proposée en (continu) pour différents taux de perte

6.9 Comparaison entre les formes d'ondes

Segment entre la trame 610 (ech 48800) jusqu'au trame 688 (ech 55072)



(a)

(b)

(c)

(d)

(e)

Fig 6.12 Comparaison entre des segments [de la trame 610 à 688] de formes d'ondes de (a) signal original (b) signal décodé sans perte (c) indicateur de perte (d) signal décodé par PLC noyé dans G.729 (e) signal décodé par la méthode proposée (FER 17.59%).

Segment entre la trame 650 (ech 52000) jusqu'au trame 728 (ech 58272)

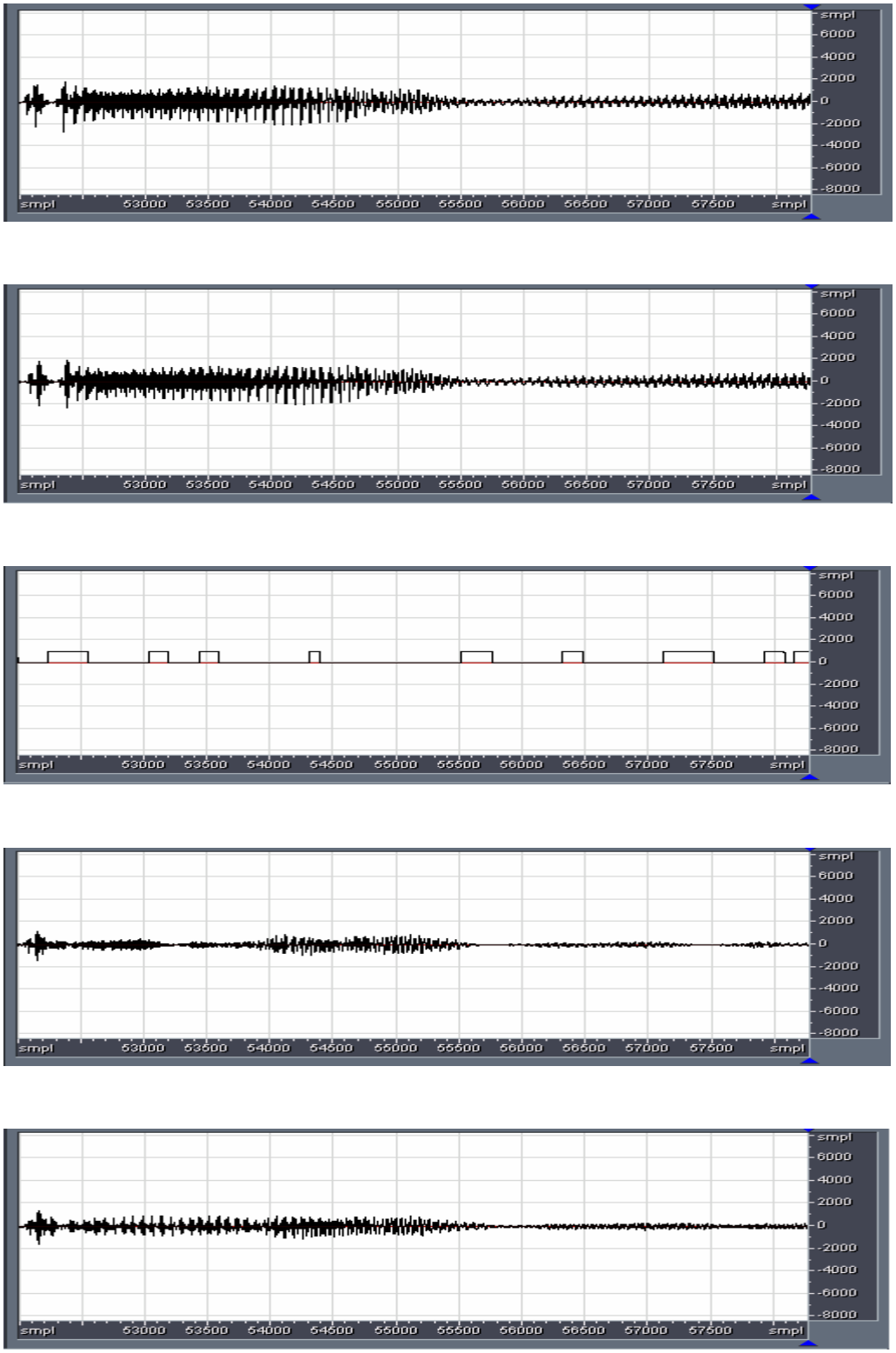


Fig 6.13 Comparaison entre des segments [de la trame 650 à 728] de formes d'ondes de (a) signal original (b) signal décodé sans perte (c) indicateur de perte (d) signal décodé par PLC noyé dans G.729 (e) signal décodé par la méthode proposée (FER 17.59%).

6.10 Conclusion et Interprétation des résultats

Les figures Fig 6.8 et Fig 6.9 et les tableaux 6.1 et 6.2 sont relatifs au signal SA1 et Les figures Fig 6.10 et Fig 6.11 et les tableaux 6.3 et 6.4 sont relatifs au signal SX38.

D'après le tableau 6.1 et la figure Fig 6.8 on remarque que la Qualité du signal reconstitué par la technique de dissimulation des trames perdues proposée est nettement meilleur par rapport au mécanisme de dissimulation noyé dans le G.729. Cette amélioration croît proportionnellement avec le taux de perte, pour un taux de perte de 50 % on a une amélioration de PESQ de 0.648, c'est une amélioration importante dans le codage de la parole.

D'après le tableau 6.2 et la figure Fig 6.9 qui représentent la distorsion spectrale EMBSD en fonction du taux de perte, on conclut que la distorsion du signal reconstitué par la technique de dissimulation des trames perdues proposée est nettement inférieur par rapport au mécanisme de dissimulation noyé dans le G.729.

Pour le signal SX38, nous avons simulé les pertes par dix taux, variant de 0% à 50%, ceci permet de bien voir l'évolution de la qualité et de la distorsion en fonction du taux de perte. D'après les figures Fig 6.10 et Fig 6.11 et les tableaux 6.3 et 6.4 on remarque que la qualité de la parole et la distorsion tendent toujours vers l'amélioration de la technique de dissimulation des trames perdues proposée par rapport au mécanisme de dissimulation noyé dans le G.729.

Les figures Fig 6.12 et 6.13 montrent que la méthode proposée réduit la distorsion dans la forme du signal, ainsi, on remarque que le signal reconstitué par la méthode proposée est plus similaire au signal original que le signal reconstitué par le mécanisme de dissimulation noyé dans le décodeur G.729.

Conclusion

Le standard G.729 utilise le codage prédictif. Dans ce type de codage, la perte des paquets cause une perte de synchronisation entre le codeur et le décodeur. Donc, les erreurs ne se produisent pas seulement dans les trames perdues, mais se propagent aussi dans les trames suivantes, c'est pourquoi la qualité de la parole se dégrade considérablement en fonction du taux de perte.

Dans ce travail nous avons proposé une amélioration du CoDec G.729 de l'UIT par implémentation d'une nouvelle technique de dissimulation des trames perdues basée sur le récepteur. Cette technique consiste en la modification de l'échelle du temps (TSM Time Scale Modification) (changer la durée du signal parole sans changer la période du fondamental « pitch») du signal de parole basée sur la méthode WSOLA (Waveforme Similarity Ovelapp-Add) afin de compléter les paquets perdus.

Les résultats obtenus montrent que cette méthode est très efficace, et apporte une amélioration importante de la qualité de la parole, spécialement lorsque le taux de perte augmente, ce qui signifie que l'application de cette technique devient recommandée dans les réseaux perturbés et instables.

La technique proposée ne dépend pas du type de codage utilisé, elle peut, par conséquent, être utilisée dans n'importe quel type de codage et n'importe quel codeur.

Nous pouvons aussi, en plus de l'amélioration concernant la perte des paquets, utiliser cette technique pour réduire le débit du codeur jusqu'à la moitié. L'idée consiste à compresser le signal original, par la technique WSOLA, avant de le coder au niveau de l'émetteur, la taille du signal à envoyer devient alors la moitié du signal original. Au niveau du récepteur, on décompresse le signal puis on le décode.

Avant de terminer cette conclusion, on peut dire que ce travail nous a apporté d'immenses intérêts tant sur le plan théorique que expérimental. En effet, De plus qu'il nous a permis de nous intégrer dans le domaine de la recherche, nous avons été amené, par ce travail, à nous instruire dans un domaine clé dans les télécommunications et les applications multimédias, c'est le codage et le traitement de la parole et par conséquent le traitement numérique du signal.

Nous avons, aussi, pu nous familiariser avec un domaine d'actualité, c'est la transmission de la voix à travers le réseau IP, nous avons étudié les systèmes VoIP, leurs architectures, les protocoles utilisés et les problèmes rencontrés, notamment la perte de paquets et comment remédier à ce problème en utilisant la technique de changement de l'échelle temporelle du signal parole au niveau du récepteur. Enfin, en étudiant le standard G.729, nous avons pu saisir l'analyse par prédiction linéaire, les techniques de quantification ainsi que les techniques de filtrage.

Bibliographie

[1] SANNECK Henning, " Packet Loss Recovery and control for Voice Transmission over the Internet", Thèse de doctorat, Berlin, 2000.

[2] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," in *Proc. ICASSP 93*, Apr. 1993, vol. II, pp. 554–557.

[3] L. Vandendorpe, " Cours de Télécommunications" Université de Louvain, 2002.

[4] André DIDIER, "Acoustique audiométrie" Encyclopédie UNIVERSALIS 5, 1999.

[5] Gilles Léothaud, "Théorie de la phonation" Cours de DEUG,DMU3D1B, 2005.

[6] Gilles Léothaud, "Acoustique musicale, Psychoacoustique" Cours de DEUG,DMU2D1B, 2005.

[7] DUTOIT Thierry, "Introduction au traitement automatique de la parole", Notes de cours - DEC2. Faculté Polytechnique de Mons, TCTS Lab, 2000.

[8] James H. Y. Loo, "Intraframe and Interframe Coding of Speech Spectral Parameters", Thèse de Master, McGill University,1996.

[9] Eddie L. T. Choy, " Waveform Interpolation Speech Coder at 4 kb/s", Thèse de Master, McGill University, 1998.

[10] Sara Grassi**,** "Optimized implementation of speech processing algorithms ", Thèse de doctorat, Université de Neuchatel, 1998.

[11] Fatiha MERAZKA, " Techniques de codage de la parole : applications aux LSPs et aux systèmes VoIP", Thèse de Doctorat d'Etat, ENP Alger, 2005.

[12] C. E. SHANNON, " A Mathematical Theory of Communication", Bell System Technical Journal, 1948.

[13] Mohammad M. A. Khan, "Coding of Excitation Signals In a Waveform Interpolation Speech Coder", Thèse de Master, McGill University, July 2001.

[14] HERNÁNDEZ, José, "Algorithmes d'acquisition, compression et restitution de la parole à vitesse variable, étude et mise en place", Projet de fin d'études de Licence d'Informatique, ENSEA (Paris), Avril 1995.

[15] ITU-T Recommendation G.711, “Pulse code modulation (PCM) of voice frequencies”, Nov. 1988.

[16] ITU-T Recommendation G.726, “40, 32, 24, 16 kbit/s adaptive differential pulse code modulation (ADPCM) ”, 1990.

[17] ITU-T Recommendation G.722, “ 7 khz audio-coding within 64 kbit/s ”, 1990.

[18] ITU-T Recommendation G.723.1, “Dual Rate Speech Coder for Multimedia CommunicationsTransmitting at 5.3 and 6.3 kbit/s,” 1996.

[18] ITU-T Recommendation G.729 - "Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP)", March 1996.

[19] Costantinos Papacostantinou, “Improved Pitch Modelling for Low Bit-Rate Speech Coders”, Thèse de Master, McGill University, August 1997.

[20] Patrice COLLEN, “ Techniques d'enrichissement de spectre des signaux audionumériques ” Thèse de doctorat, ENST France, 2002.

[21] Guy Pujolle, “Les réseaux”. Editions Eyrolles, Paris, 2000.

[22] Christian Hoene, Iacopo Carreras, Adam Wolisz, “Voice Over IP : improving the quality over wireless LAN by adopting a booster mechanisme – an experimental approach”, article, berlin, 2001.

[23] Aziz Shallwani, “An Adaptive Playout Algorithm with Delay Spike Detection for Real Time VoIP”, Thèse de Master, McGill University, 2001.

[24] Spirent Communications, “ Voice over IP (VoIP)”, Article, Spirent Communications, 2001.

[25] Colm Elliott, “Stream Synchronization for Voice over IP, Conference Bridges”, Thèse de Master, McGill University, November 2004.

[26] Choon Shim, Liehue Xie, Bryan Zhang and C.J. Sloane, “ How Delay and Packet Loss Impact Voice Quality in VoIP ”, Article de Qovia, 2003.

[27] Jonas Lindblom, “Packetized Speech Transmission-Combatting the Packet Loss Problem”, Thesis for the degree of Licentiate of Engineering, Göteborg 2001.

[28] François TOUTAIN, “ Téléphonie Internet”. École Nationale Supérieure des Télécommunications de Bretagne, Techniques de l’Ingénieur, traité réseaux, IP2960.

[29] Jean Chiappini, “Performances de la VoIP sur réseaux wireless”, PFE en réseaux et services, Ecole d’Ingénieurs du Canton de Vaud, 2002.

[30] Antoine Delley, “Voix sur IP-Architectures”, Article, HES-SO / EIA-Fribourg, 2002.

[31] Antoine Delley, “ Réseaux IP -Voix et multimédia sur IP”, Journal ICTnet, Fribourg, 2002.

[32] Moo Young Kim and Renat Vafin, “Packet-loss recovery techniques for VoIP”, Article, Dept. of Speech, Music, and Hearing, Royal Institute of Technology (KTH), 2002.

[33] Yvan Calas,"Performances des codes correcteurs d'erreur au niveau applicatif dans les réseaux", Thèse de doctorat, Université de Montpellier, 2003.

[34] ITU-T Recommendations on CD-ROM – March 2000.

[35] Lim Hong Swee, " Implementation of G.729 on the TMS320C54x", Texas Instruments Singapore, Application Report SPRA656 - March 2000.

[36] Jonathan D. Rosenberg, "G.729 Error Recovery for Internet Telephony", Lucent Technologies, Bell Laboratories& Columbia University

[37] Grégory PALLONE, "Dilatation et transposition sous contraintes perceptives des signaux audio : application au transfert cinema-video." Université d'Aix-Marseille II, Thèse de doctorat, 2003.

[38] Ejaz Mahfuz, " Packet Loss Concealment for Voice Transmission over IP Networks", Thèse de Master, McGill University, 2001.

[39] Jean Laroche, "Traitement des Signaux Audio-Fréquences", Support de cours, Département Signal, Groupe Acoustique TELECOM, Paris, Février 1995.

[40] Daniel W**.** Griffin and Jae S. Lim, " Signal Estimation from Modified Short-Time Fourier Transform ", IEEE Transactions on acoustics, speech, and signal processing, vol. assp-32, no. 2, april 1984.

[41] Mike Demol, Werner Verhelst, Kris Struyve, Piet Verhoeve, "Efficient non-uniform time-scaling of speech with WSOLA", Article, Vrije Universiteit Brussel, Belgium,2005.

[42] Werner Verhelst, "Overlap-Add Methods for Time-Scaling of Speech.", Speech Commun. 30 (2000) 207–221.

[43] ITU-T Recommendation P.862, "Perceptual evaluation of speech quality (PESQ), an objective method of end-to-end speech quality assessment of narrowband telephone networks and speech codecs," May. 2000.

[44] The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT) Training and Test Data and Speech Header Software, NIST Speech Disc CD1-1.1, October 1990.
http://www.ltam.lu/Tutoriel_Ansi_C.
http://www.itu.int/publications/bookstore.html.