9/03

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

Ecole Nationale Polytechnique

Département d'Electronique

Mémoire de Projet de Fin d'Etudes en vue de l'Obtention du Diplôme d'Ingénieur d'Etat en Electronique

Thème

Implémentation d'une méthode interpolative de masquage de perte au standard « G.729 »

Proposé et Dirigé par :
M^cllc F.MERAZKA

Etudié par :
Mr SMATTI M^cd Lamine
Mr BENALIA Youcef

Soutenu le : 06 juillet 2003
Devant le jury composé de :

Président : Mr S.AIT CHEIKH	Chargé de cours ENP
Examinateur : Mr D.BERKANI	Professeur ENP
Encadreur : M^cllc F.MERAZKA	Chargée de Recherche ENP

Promotion : 2002/2003

Dédicace

Je dédie ce modeste travail :

- A mes très chers parents qui m'ont toujours soutenu.
- A mes frères et sœurs : Aiman, Imane, Saddik, Soundous, Monsif et le petit Hatem.
- A toute ma famille.
- A tous mes amis d'Ouled Djellal.
- A tous mes amis et collègues de Bouraoui « les Bouraouistes ».
- A tous qui me sont chers.

Samine

Je dédie ce travail :

- A mes très chers parents.
- A mes chers frères et sœurs.
- A toute ma famille.
- A tous qui ont de près ou de loin assisté à ma formation.
- A tous mes camarades et collègues.

Youcef

Remerciement

نحمد الله تعالى الذي وفقنا لإتمام هذا العمل

Nous exprimons notre profonde gratitude à notre promotrice, M^elle MERAZKA pour ses conseils judicieux, son suivi attentif et sa confiance qui nous a été très précieuse, de plus nous tenons à la remercier pour nous avoir assurer l'encadrement et la documentation nécessaire pour l'élaboration de ce mémoire.

Nous adressons, aussi, nos vifs remerciements à Mr BERKANI pour ses conseils qui nous a été très utiles et très précieux, et d'avoir accepté d'examiner et de juger notre travail.

Nous tenons aussi à remercier Mr AIT CHEIKH qui nous a fait l'honneur de présider le jury de ce mémoire.

Nos remerciements à l'ensemble des enseignants et les membres du personnel de notre Département d'électronique pour leurs aides et leur disponibilité tout au long de notre cursus à l'ENP.

Aussi nos remerciements s'adresse à Mr LEMDELDEL pour son aide et ses conseils judicieux.

Enfin, nous remercions toute personne ayant contribué de prés ou de loin à l'élaboration de ce travail.

ملخص

تعالج هذه المذكرة استعمال مشفر الكلام "G.729" على شبكة الانترنت من اجل تطبيقات الهاتف، و بالأخص يتطرق هذا العمل إلى مشاكل ضياع و استدراك حزيمات المعلومات.
هذا المشفر يكمم الوسائط "LSF " اهتماما بشدة التدفق و التعقيد.
كون أن التشفير ما بين الحزيمات المستعمل من طرف "G.729"، بامكانه إحداث انتشار الخطأ عندما يحصل ضياع في حزيمات معلومات الكلام ، قمنا بدراسة فعالية التشفير ما بين الحزيمات في حالة ضياع، وقارناها بطريقة التشفير داخل الحزيمات مع استعمال إخفاء استكمالي بدل طريقة الإخفاء التكراري المتبناة من طرف "G.729".
النتائج التي تحصلنا عليها اضهرت انه من اجل 2.5% كتدفق زائد، فان التشفير داخل الحزيمات ومع الإخفاء الاستكمالي جد مقاوم مع تحسينات في النوعية ب0.7 dB في التشويه.
المفاتيح : G.729، الإخفاء، VoIP، ما بين الحزيمات ، داخل الحزيمات، LSF ، SVQ، محو الحزيمات

Abstract

This work discusses the use of the ITU G.729 CS-ACELP speech coder on the Internet for telephony applications. In particular, the memo explores issues of error resiliency and recovery. Since this codec quantizes the LSF parameters with the primary concerns of bit-rate and complexity, the inter-frame coding of LSF's used by G.729 can cause error propagation when frame erasures occur. We investigate the erasure performance of interframe LSF coding and compare it with an intraframe coding method with an interpolative concealment. Our results show that with only 2.5% extra bit-rate, intra-frame coding with the interpolative concealment is much more robust to frame erasures and a typical improvement of 0.7 dB on spectral distortion can be obtained with 20% packet loss. Subjective listening tests indicate significant improvement as well.

Keywords: G.729, Concealment, VoIP, Intra-frame, Inter-frame, LSF, SVQ, frame erasures.

Résumé

Ce mémoire discute l'utilisation du codeur de la parole de l'ITU G.729 CS-ACELP sur l'Internet pour des applications de téléphonie. En particulier, le travail explore les problèmes de l'apparition et de la récupération des trames perdues. Puisque ce codec quantifie les paramètres LSF en s'intéressant particulièrement au débit binaire et à la complexité, le codage inter-trame des LSF adopté par le G.729 peut causer une propagation de l'erreur quand des effacements de trames se produisent. Nous étudions la performance du codage inter-trame des LSF dans le cas de pertes, et la comparons à celle d'une méthode de codage intra-trame, avec un masquage interpolatif au lieu de la méthode répétitive utilisée par le G.729. Nos résultats montrent qu'avec seulement 2,5% d'extra débit, le codage intra-trame avec le masquage interpolatif est beaucoup plus robuste, et une amélioration typique de 0,7 dB dans la distorsion spectrale peut être obtenue avec un taux de perte de paquet de 20%. Des tests subjectifs informels d'écoute indiquent aussi l'amélioration significativc.

Mots-clés : G.729, masquage, VoIP, Intra-trame, Inter-trame, LSF, SVQ, effacement de trame.

Sommaire

Liste des Figures

Liste des Tableaux

Abréviations

Le Tableau suivant résume les abréviations les plus utilisées dans ce mémoire.

Abréviation	**Signification**
ADM	Adaptive Delta Modulation
ADPCM	Adaptive Differential Pulse Code Modulation.
AR	Auto-Regressif.
ARMA	Auto-Regressif Moving Average
ARQ	Automatic Repeat reQuest
CELP	Code Excited Linear Prediction.
CS-ACELP	Conjugate Structutre Algebraic Code Excited Linear Prediction
DPCM	Differential Pulse Code Modulation.
EMBSD	Enhanced Modified Bark Ditorsion
FEC	Forward Error Correction.
IP	Internet Protocol.
ITU	International Telecommunacation Union
LP	Linear prediction.
LPC	Linear Predicion Coding.
LSP	Line Spectrum Pairs.
MOS	Mean Opinion Score.
PCM	Pulse Code Modulation.
PLC	Packet Loss Concealement
RTP	Real time Transfert Protocol
SD	Spectrtal Distortion.
SNR	Signal to Noise Ration
SQ	Scalar Quantization.
SVQ	Split Vector Quantization.
ULP	Uneven Level Protection.
VoIP	Voice cover IP network.
VQ	Vector Quantization.

Introduction

Ces dernières années, les groupes de recherches, et les simples utilisateurs, ont découvert l'intérêt considérable de la transmission interactive de la parole par Internet (VoIP[1]). Actuellement, la motivation principale de la Téléphonie par Internet, est le prix fixe et économique, comparé au tarif des services de la téléphonie traditionnels, qui est basé sur l'usage. Bien que ce prix ne peut pas rester réduit comme ça dans le futur, la transmission de la parole par Internet, reste très attirante, car elle peut être intégrée avec d'autres applications Internet pour fournir des services multimédias interactifs, qui sont impossibles (ou au moins très difficiles) à utiliser sur les réseaux de téléphonie traditionnel.

En outre, des codages et des décodages très complexes de la parole peuvent être menés avec un matériel qui peut être disponible chez tous les utilisateurs. Comme, par exemple, les deux codecs appelés "*frame-based*[2]" le G.729 [1], et le G.723.1 [2], qui sont très convenables pour la Téléphonie par Internet, car ils fournissent une qualité "*toll*[3]" de la parole avec des faibles débits binaires (bit-rate) (8 kbit/s et 5.3/6.3 kbit/s respectivement) comparés au PCM conventionnel (64 kbit/s). Donc les exigences sur la capacité du réseau pour une diffusion à grande échelle peuvent être réduites considérablement.

Cependant, les réseaux à commutation de paquets d'aujourd'hui, comme les réseaux IP, sont basés sur le principe dit "*best effort*", qui ne garantit pas le taux minimal de perte de paquets exigé par la *VoIP*, ni le délai minimal de transmission de ces derniers. Cela implique de diverses influences sur la qualité de la parole, par exemple, quand les routeurs ou les passerelles sont encombrés, des paquets de parole peuvent être abandonnés.

Dû à la nécessité du temps-réel pour la transmission interactive de la parole, généralement il est impossible que le récepteur demande la retransmission des paquets perdus. En plus, les paquets de parole qui n'arrivent pas avant leur temps du playout (le temps qu'ils doivent être écoutés) sont considérés perdus, et ne peuvent pas être joués quand ils seront reçus.

---

[1] Dans ce mémoire "*Voice over IP(Internet Protocol) network*" est abrégé VoIP.
[2] Pour la bonne compréhension du travail, nous avons gardé les mots anglais utilisés dans le domaine de la parole car ils sont plus significatifs.
[3] C'est-à-dire, aussi bon que la qualité fournie par la téléphonie traditionnelle

En outre, étant donné que le codage adopté par le G.729 est prédictif[1], 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.

Le but de notre travail est d'implémenter une méthode de masquage des trames effacées sur le standard de l'ITU (G.729), plus performante que celle adoptée par ce dernier, pour alléger le problème décrit dans le paragraphe précédent et améliorer les performances du codec.

Nous avons partagé notre travail en quatre chapitres :

Le **premier** chapitre comporte des généralités sur le modèle humain de production de la parole, le système auditif humain et les différentes méthodes de codage du signal parole.

Le **deuxième** chapitre est consacré à la transmission de la voix à travers les réseaux IP (*VoIP*), les problèmes de perte de paquets rencontrés et les les differentes méthodes connues de récupération (ou de masquage) de ces pertes.

Le **troisième** chapitre décrit le fonctionnement du standard de l'ITU, le codec G.729.

Le **quatrième** chapitre expose les simulations, les tests et les résultats obtenus en cours de notre implémentation d'une méthode interpolative pour le masquage des trames du signal parole effacées.

On **terminera** par une conclusion générale sur les méthodes que nous avons utilisé et les différentes perspectives.

[1] Quantification prédictive des paramètres LSF.

Chapitre 1

Codage de la parole

1.1 Introduction :

Ce chapitre se compose de deux parties, dans la première nous allons donner des généralités regroupant des notions fondamentales en traitement du signal de la parole, production, propriétés, et perception, nécessaires à la bonne compréhension de l'évolution des techniques de codage, qui seront exposées dans la deuxième partie.

1.2 Le signal vocal :

La parole peut être décrite comme le résultat de l'action volontaire et coordonnée d'un certain nombre de muscles. Cette action se déroule sous le contrôle du système nerveux central qui reçoit en permanence des informations par rétroaction auditive et par les sensations kinesthésiques.

1.2.1 Mécanisme de la phonation :

Les principaux organes composant l'appareil phonatoire sont : les poumons, la trachée-artère, le pharynx, les cavités buccales et nasales (Figure 1.1).

L'appareil respiratoire fournit l'énergie nécessaire à la production de sons, en poussant de l'air à travers la trachée-artère. Au sommet de celle-ci se trouve le ***larynx*** ou la pression de l'air est modulée avant d'être appliquée au conduit vocal. Le larynx est un ensemble de muscles et de cartilages mobiles qui entourent une cavité située à la partie supérieure de la trachée. Les ***cordes vocales*** sont en fait deux lèvres symétriques placées en travers du larynx. Ces lèvres peuvent fermer complètement le larynx et, en s'écartant progressivement, déterminent une ouverture triangulaire appelée ***glotte***. L'air y passe librement pendant la respiration et la voix chuchotée, ainsi que pendant la phonation des sons non voisés.

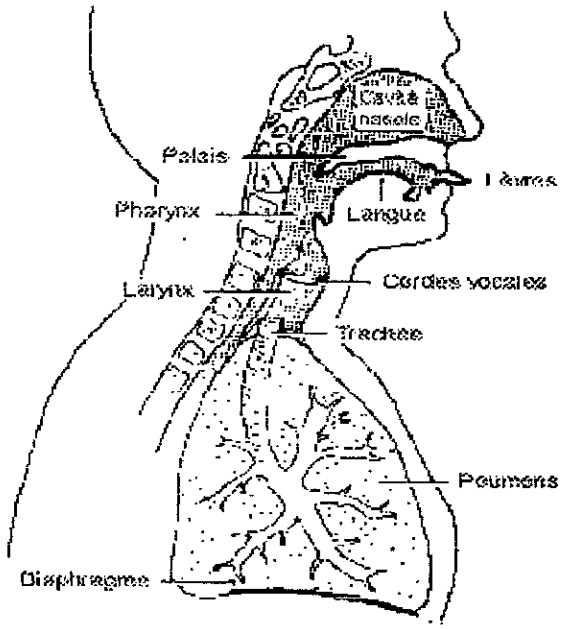

Figure 1.1 Appareil phonatoire

Les sons voisés résultent au contraire 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 les force à s'ouvrir, ce qui fait tomber la pression, et permet aux cordes vocales de se refermer; des impulsions périodiques de pression sont ainsi appliquées au conduit vocal, composé des cavités pharyngienne et buccale pour la plupart des sons. Lorsque la ***luette*** est en position basse, la cavité nasale vient s'y ajouter en dérivation. Notons pour terminer le rôle prépondérant de la langue dans le processus phonatoire. Sa hauteur détermine la hauteur du pharynx : plus la langue est basse, plus le pharynx est court. Elle détermine aussi le ***lieu d'articulation***, région de rétrécissement maximal du canal buccal, ainsi que l'aperture, écartement des organes au point d'articulation.

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

La fréquence du fondamentale peut varier [4] :

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

Un son voisé est un signal quasi périodique dont le spectre est tracé à la figure 1.2.On y observe les raies qui correspondent aux harmoniques du fondamentale F_0 (structure de *pitch*), l'enveloppe de ces raies présente des maximums appelés *formants* et qui correspondent aux fréquences propres F_i (i=1, 2, 3....) du conduit vocal (structure formantique).

Les trois premiers formants sont essentiels pour caractériser le spectre vocal, les formants d'ordre supérieur ont une influence plus limitée.

Un son non voisé ne présente pas de structure périodique, il peut être considère comme un bruit blanc filtré par la transmittance de la partie du conduit vocal située entre la constriction et les lèvres (Figure 1.3), son spectre ne présente donc pas de structure de pitch.

La classification qui vient d'être exposée est forcement un peut sommaire et surtout elle concerne la production normale de la parole. Ainsi une voyelle peut être chuchotée, c-à-d produite avec la glotte largement ouverte, dans ce cas le spectre du signal résulte de l'excitation du conduit vocal par une source aléatoire :c'est un spectre continu qui présente une structure formantique semblable à celle d'une voyelle voisée. Par contre, il ne possède pas de structure de pitch (raies dues aux harmoniques du fondamentale).

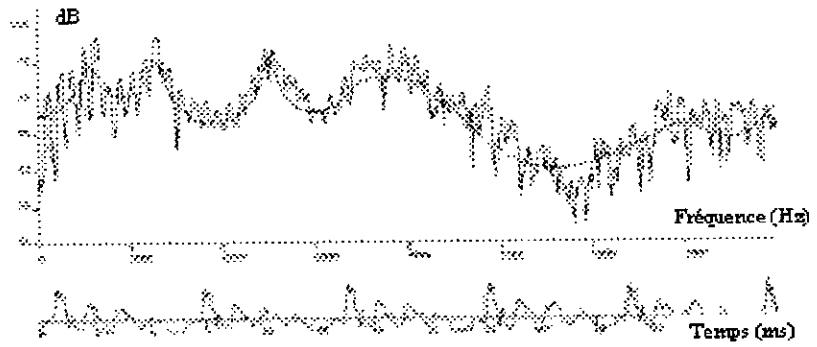

Figure 1.2 Un signal vocal voisé et son spectre

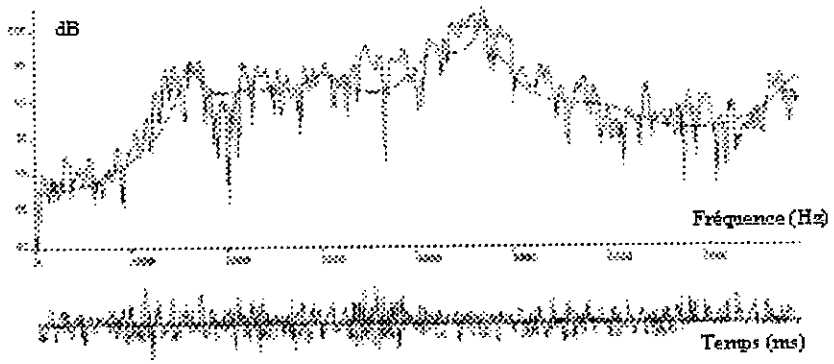

Figure 1.3 Un signal vocal non voisé et son spectre

Il reste très difficile de nos jours de dire comment l'information auditive est traitée par le cerveau. On a pu par contre étudier comment elle était finalement perçue, dans le cadre d'une science spécifique appelée ***psychoacoustique***. Sans vouloir entrer dans trop de détails sur la contribution majeure des psychoacousticiens dans l'étude de la parole, il est intéressant d'en connaître les résultats les plus marquants. Ainsi, l'oreille ne répond pas également à toutes les fréquences. Le seuil d'audition de l'oreille est non linéaire par rapport aux fréquences. L'oreille atteint sa sensibilité maximale entre 3 et 4 kHz.

1.2.2 Les redondances dans le signal de parole :

Le signal vocal est caractérisé par une très grande redondance, condition nécessaire pour résister aux perturbations du milieu ambiant. 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 sans nuire à son intelligibilité.

En pratique, cependant, il ne s'agit pas de redondance stricte, puisque nous avons négligé les nuances apportées par la prosodie et les caractéristiques propres à chaque locuteur. On parlera plutôt dans ce cas de variabilité. Le problème des techniques de reconnaissance de la parole sera précisément de retrouver le sens d'une phrase malgré l'extrême variabilité qui la caractérise.

Exemple : Soit trois types de codage binaires d'une source X contenant quatre messages $\{a, b, c, d\}$ (Tableau 1.1).

Tableau 1.1 Exemple de messages à coder

Mot	Probabilité	codage		
a	0.5	000000	00	0
b	0.25	010101	01	10
c	0.125	101010	10	110
d	0.125	111111	11	111

L'entropie de la source est:

$$H(X) = 0.5\log 2 + 0.25\log 4 + 0.125\log 8 + 0.125\log 8 = 1.75 bits.$$

Dans le deuxième type de codage (n=2) :

L'efficacité vaut :

$$\eta = \frac{H}{2} = 0.875$$

La redondance vaut :

$$\rho = 1 - \eta = 0.125$$

L'efficacité de 87.5% (ou la redondance de 12.5%) provient du fait qu'on avantage de la même manière des messages de fréquences différentes.

Le troisième codage, au contraire, affecte les mots les plus courts aux messages les plus fréquents.

On a: $n = 1 \times 0{,}5 + 2 \times 0{,}25 + 3 \times 0{,}125 + 3 \times 0{,}125 = 1{,}75$.

Donc : $\eta = 1$ et $\rho = 0$.

Le code est efficace à 100%. Sa redondance est nulle.

Dans la conversation courante, environ dix phonèmes sont prononcés chaque seconde; l'information moyenne est donc inférieure à 50 bits/s [5]. De l'autre coté pour garder une haute qualité de la parole avec une représentation numérique du signal parole, l'utilisation d'un système de conversion A/D réclame plus de 100000 bits par seconde. Il y a donc apparemment une redondance énorme dans le signal vocal. La suppression partielle des redondances permet une représentation plus efficace des données.

La compression des données peut se faire sans pertes d'information, ou avec pertes en exploitant dans ce cas la tolérance de l'organe récepteur (l'oreille). La compression du signal consistera à réduire les redondances du signal de la parole.

1.2.3 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. Celui-ci possède un ensemble de paramètres numériques, dont les plages de variation définissent l'ensemble des signaux couverts par le modèle.

Fant [6] a proposé en 1960 un modèle de production dont nous résumons ici la version numérique. Un signal voisé peut être modélisé par le passage d'un train d'impulsions *u(n)* à travers un filtre numérique récursif de type tout-pôles (AR[1]). On montre que cette modélisation reste valable dans le cas de sons non voisés, à condition que *u(n)* soit cette fois

[1] Abréviation de "*Auto Régressif*"

un bruit blanc. Le modèle final est illustré à la figure 1.4. Il est souvent appelé modèle auto régressif, parce qu'il correspond dans le domaine temporel à une régression linéaire de la forme :

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

(où *u(n)* est le signal d'excitation), ce qui exprime que chaque échantillon est obtenu en ajoutant un terme d'excitation à une prédiction obtenue par combinaison linéaire de p échantillons précédents.

Les coefficients du filtre sont d'ailleurs appelés coefficients de prédiction et le modèle AR est souvent appelé modèle de prédiction linéaire.

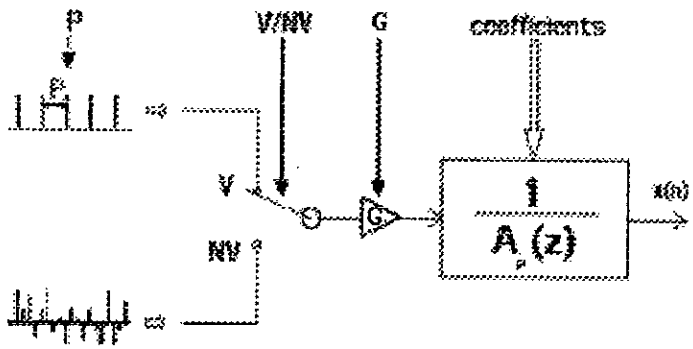

Figure 1.4 Modèle de production de la parole [3].

Les paramètres du modèle AR sont : la période du train d'impulsions (sons voisés uniquement), la décision Voisé / Non Voisé (V/NV), le gain σ, et les coefficients du filtre *1/A(z)*, appelé filtre de synthèse.

Les relations d'équivalence entre le modèle physique et le modèle mathématique sont :

Conduit vocal	⟺	*H(z)* (Le filtre LPC)
Le flux d'air	⟺	*U(n)* (L'excitation)
Vibration des cordes vocales	⟺	*V* (Voisé)
Périodes de vibration des cordes vocales	⟺	*T*
Fricatives et Plosives	⟺	*NV* (Non voisé)
Volume d'air	⟺	*G* (Gain)

Le problème de l'estimation d'un modèle AR, souvent appelée analyse *LPC*[1] revient à déterminer les coefficients d'un filtre tout pôles dont on connaît le signal de sortie, mais pas l'entrée. Il est par conséquent nécessaire d'adopter un critère, afin de faire un choix parmi l'infinité de solutions possibles. Le critère classiquement utilisé est celui de la minimisation de l'énergie de l'erreur de prédiction.

1.3 La prédiction linéaire :

La prédiction linéaire est l'une des méthodes les plus puissantes dans l'analyse du signal de la 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 la production de la parole.

Le principe fondamental de la prédiction linéaire est qu'un échantillon du signal *S*(n) peut être modéliser comme la sortie d'un système Auto Régressif à Moyenne Ajustée (ARMA) avec une entrée u(n) [3]et[5] :

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + G\sum_{l=0}^{q} b_l u(n-l), \qquad b_0 = 1, \tag{1.2}$$

Où $\{a_k\}$, (b_i), et le gain G sont les paramètres du système. L'équation (1.2) 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 :

$$H(z) = \frac{B(z)}{A(z)} = \frac{G[1+\sum_{l=1}^{q} b_l z^{-l}]}{1-\sum_{k=1}^{p} a_k z^{-k}} \tag{1.3}$$

Les racines du dénominateur et numérateur sont, respectivement, les pôles et les zéros du système ou modèle pôle-zéro H(z).

[1] Dans ce mémoire "*Linear Prediction Coding*" est abrégé LPC

Si $a_k = 0$ pour $1 \le k \le p$, $H(z)$ devient un modèle tout zéro ou modèle à moyenne ajustée (MA). Si $\{b_i = 0\}$ pour $1 \le i \le q$, $H(z)$ devient un modèle tout pôle ou modèle Auto Régressive (AR) :

$$H(z) = \frac{1}{A(z)} \tag{1.4}$$

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, ce modèle est préféré pour l'analyse par prédiction linéaire de la parole.

Ainsi, le signal prédit est égal à :

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) \tag{1.5}$$

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

$$e(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \tag{1.6}$$

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 due aux sons non voisés.

Quand la prédiction linéaire est basée sur les échantillons de parole passes $s(n)$, celle-ci, est dite « Prédiction Linéaire Adaptative Progressive (*Forward*) » et dans ce cas les coefficients de prédiction doivent être transmis au décodeur. Si la prédiction linéaire est basée sur les échantillons de parole reconstruits antérieurs $\tilde{S}(n)$, celle-ci, est dite « Prédiction Linéaire Adaptative Régressive (*Backward*) ». Pour avoir les coefficients du filtre court-terme $\{a_i\}$ du processus AR, la méthode classique des moindres carrés peut être utilisé. La variance ou l'énergie, du signal erreur $e(n)$ est minimisée sur une trame de parole. Deux grandes approches sont utilisées pour l'analyse *LPC* court-terme : la méthode d'*autocorrélation* et la méthode de *covariance*.

1.3.1 Méthode d'Autocorrélation :

La méthode d'Autocorrélation garantit la stabilité du filtre LP.

Les suppositions de cette méthode sont les suivants :

Le signal est défini pour toutes les valeurs du temps ; il est identiquement nul en dehors d'une séquence de N échantillons, où N est un entier; ceci équivaut à multiplier le signal de parole par une fenêtre de longueur finie correspondant à N échantillons.

$$\begin{cases} S_f(n) = W(n) \cdot S(n) & \text{pour } 0 \le n \le N-1 \\ S_f = 0 & \text{ailleurs} \end{cases} \tag{1.7}$$

La fonction de pondération la plus courante est la fenêtre de Hamming :

$$\begin{cases} W(n) = 0.54 - 0.46 \cos \dfrac{2\pi n}{N-1} & \text{pour } 0 \le n \le N-1 \\ W(n) = 0 & \text{ailleurs} \end{cases} \tag{1.8}$$

Chaque échantillon peut être prédit approximativement à partir de p échantillons précédents. Ceci est valable pour toutes les valeurs du temps : $-\infty < n < +\infty$.

L'erreur quadratique totale entre le signal fenêtré et le modèle (signal prédit) est minimisée sur l'ensemble des échantillons.

Après la multiplication du segment de parole par la fenêtre d'analyse, les coefficients d'autocorrélations du segment fenêtré sont calculés. La fonction d'autocorrélation du signal fenêtré $S_f(n)$ est :

$$R(i) = \sum_{n=i}^{N-1} s_f(n) s_f(n-i) \qquad 1 < i < p \tag{1.9}$$

La fonction d'autocorrélation est une fonction paire : $R(i) = R(-i)$

Pour trouver les coefficients du filtre LPC, l'énergie du résiduel de prédiction sur l'intervalle fini $0 \le n \le N-1$ doit être minimisée :

$$E = \sum_{n=-\infty}^{\infty} e^2(n) = \sum_{n=-\infty}^{\infty} \left(s_f(n) - \sum_{k=1}^{p} a_k s_f(n-k)\right)^2 \tag{1.10}$$

En annulant les dérivations partielles par rapport aux coefficients du filtre :

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

On obtient p équation linéaire avec "p" coefficient inconnus a_k :

$$\sum_{k=1}^{p} a_k \sum_{n=-\infty}^{\infty} s_f(n-i)s_f(n-k) = \sum_{n=-\infty}^{\infty} s_f(n-i)s_f(n) \,..\, 1 \le i \le p \tag{1.11}$$

Alors, les équations linéaires peuvent être écrites sous la forme :

$$\sum_{k=1}^{p} R(|i-k|)a_k = R(i) \qquad 1 \le i \le p \tag{1.12}$$

Sous la forme matricielle, l'ensemble des équations linéaires est représenté par $R \cdot a = v$ qui peut être réécrit comme la formule (1.13)

$$\begin{bmatrix} R(0) & R(1)\cdots\cdots R(p-1) \\ R(1) & R(2).....R(p-2) \\ R(2) & R(0)\cdots\cdots \\ . & ..... \\ . & ..... \\ R(p-1) & R(p-2)\cdot\cdot R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ . \\ . \\ . \\ a_p \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ . \\ . \\ . \\ R(p) \end{bmatrix} \tag{1.13}$$

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

1.3.2 Méthode de covariance :

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

$$E = \sum_{n=-\infty}^{\infty} e_f^{\,2}(n) = \sum_{n=-\infty}^{\infty} e^2(n)w^2(n) \tag{1.14}$$

En annulant les dérivations partielles par rapport aux coefficients du filtre $\delta E / \delta a_k = 0$

Pour $1 \le i \le p$, on a "p" équations linéaires.

$$\sum_{k=1}^{p} \Phi(i,k) = \Phi(i,0) \qquad 1 \le i \le p \tag{1.15}$$

Où la fonction de covariance $\Phi(i,k)$ est définie par l'équation (1.16)

$$\Phi(i,k) = \sum_{n=-\infty}^{\infty} w^2(n)s(n-1)s(n-k) \tag{1.16}$$

Sous la forme matricielle, les p équations deviennent $\Phi_a = \Psi$

$$\begin{bmatrix} \Phi(1,1) & \Phi(1,2)....\Phi(1,p) \\ \Phi(2,1).\Phi(2,2)....\Phi(2,p) \\ . \quad ..... \\ . \quad ..... \\ \Phi(p,1)\Phi(p,2)....\Phi(p,p) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ . \\ . \\ a_p \end{bmatrix} = \begin{bmatrix} \Psi(1) \\ \Psi(2) \\ . \\ . \\ \Psi(p) \end{bmatrix} \tag{1.17}$$

Où : $\Psi(i) = \Phi(i,0)$ pour $1 \le i \le p$.

La matrice Φ n'est pas une matrice Toeplitz, elle est symétrique et définie positive. La matrice de covariance peut être décomposée en matrices triangulaires supérieures et inférieures :

$$\Phi = LU \tag{1.18}$$

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

$$\Phi = CC^T \tag{1.19}$$

Ou C= L et C^T= U. le vecteur $\vec{a}$ est trouvé en résolvant d'abord l'équation (1.20)

$$Ly = \Psi \tag{1.20}$$

Puis :

$$Ua = y \tag{1.21}$$

1.3.3 Considération pratique :

Pour mener à bien une analyse LPC, il faut pouvoir choisir :

- La fréquence d'échantillonnage f_e.
- La méthode d'analyse et l'algorithme correspondant.
- L'ordre p de l'analyse LPC.
- Le nombre d'échantillons par tranche N et le décalage entre tranches successives L.

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

L'ordre de prédiction P, est choisi de façon à ce qu'il permette de bien représenter toute séquence de signal de parole.

Il a été montré que pour donner une représentation satisfaisante des pôles de la fonction de transfert du conduit vocal, la durée de mémorisation du prédicteur linéaire doit être le double du temps mis par l'onde de parole pour se propager de la glotte jusqu'aux lèvres.

Lorsque la fréquence d'échantillonnage est f_e(exprimée en échantillon/sec), la période de 1ms correspond à f_e/1000 échantillons. A la fréquence d'échantillonnage de 8 kHz, la valeur correspondante de P doit être au moins égale à 8. Elle trouve d'ailleurs une justification expérimentale dans le fait que l'énergie de l'erreur de prédiction diminue rapidement lorsqu'on augmente p à partir de 1, pour tendre vers une asymptote autour de ces valeurs : il devient inutile d'encore augmenter l'ordre, puisqu'on ne prédit rien de plus.

La durée des trames d'analyse et leur décalage sont souvent fixés à 30 et 10 ms respectivement. Ces valeurs ont été choisies empiriquement; elles sont liées au caractère quasi-stationnaire du signal parole.

Enfin, pour compenser les effets de bord, on multiplie en général préalablement chaque tranche d'analyse par une fenêtre de pondération $w(n)$, la plus souvent utilisée est celle de Hamming (1.22).

$$\begin{cases} W(n) - 0.54 - 0.46\cos\dfrac{2\pi n}{N-1} & \text{pour } 0 \le n \le N-1 \\ W(n) = 0 & \text{ailleurs} \end{cases} \qquad (1.22)$$

Où N est la longueur de la fenêtre

1.3.4 Représentation des paramètres de prédiction :

Les coefficients de prédiction linéaire ne sont pas toujours codés directement, mais sont transformés en un ensemble de paramètres qui ont des propriétés désirables. Plusieurs représentations des coefficients ont été proposées. Les plus populaires actuellement sont les paires de raies spectrales (LSP). D'autres représentations incluent le coefficient de réflexion, *log area ratio* (LAR), les coefficients cepstraux, la réponse impulsionnelle du filtre LP,....etc.

Les paramètres LSF[1] se prêtent mieux à la quantification que les autres représentations du filtre LPC à cause des propriétés suivantes:

- Les LSF ont de bonnes propriétés statistiques, et la stabilité du filtre de synthèse est assurée par la préservation de la propriété d'ordonnancement. En plus cette propriété permet la détection des erreurs de transmissions des LSF sans introduire de redondance.
- Il y a une relation évidente entre les LSF et le spectre du filtre LPC. Une concentration des LSF dans une certaine bande de fréquences correspond approximativement à une résonance dans cette bande.
- Les LSF entre deux fenêtres d'analyse adjacentes sont fortement corrélés.
- Un changement d'une LSF cause seulement un changement dans la forme du filtre de l'analyse dans une petite gamme de fréquence autour de cette LSF.

Dans l'annexe B, l'algorithme de conversion des paramètres de prédiction en LSP est détaillé

1.4 Codage de la parole :

Le but de l'opération du codage est de réduire le taux d'informations à envoyer à chaque seconde tout en gardant une qualité satisfaisante du signal reconstruit.

La première opération du codage est l'échantillonnage du signal analogique à une certaine fréquence d'échantillonnage et une certaine précision, cette précision étant caractérisée par le nombre de bits utilisés pour coder l'amplitude de chaque échantillon. Il est clair que le choix de la fréquence et du nombre de bits utilisés répond à un compromis débit/qualité du signal

[1] Abréviation de « Line spectrum frequencies », c'est une autre représentation des LSP ; LSF = acos(LSP), et c'est pour ça que, dans ce mémoire, on les considère comme une même représentation des coefficients LP.

codé. Plus grande est la qualité souhaitée, plus important est le débit obtenu après échantillonnage.

1.4.1 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 la plus voisine valeur tirée d'un ensemble 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.

La quantification peut être scalaire ou vectorielle selon que la variable x est à une ou plusieurs dimensions.

1.4.1.1 Quantification Scalaire :

La valeur quantifiée est représentée par l'une des valeurs discrètes fixes dites niveaux de quantification (Figure 1.5) Dans la quantification scalaire uniforme, ces niveaux sont régulièrement espacés. Dans la quantification logarithmique, l'espacement est uniforme dans le domaine logarithmique.

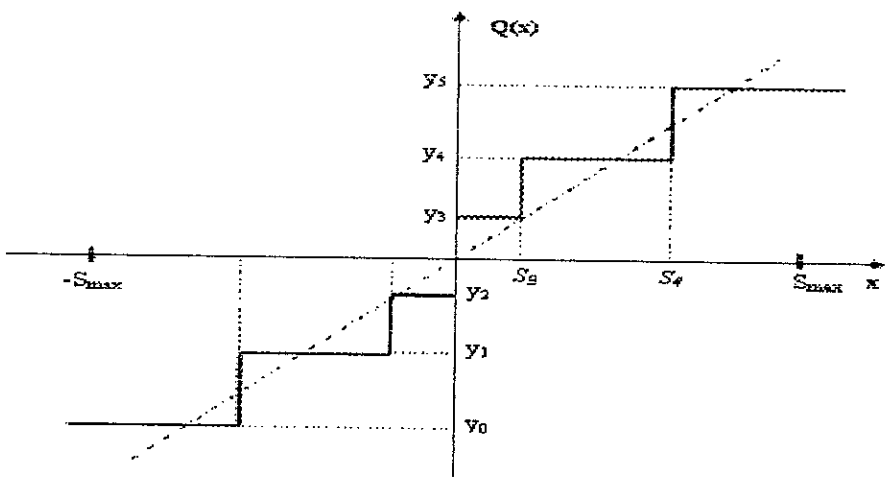

Figure 1.5 Caractéristiques typiques d'un quantificateur scalaire (N=6)

1.4.1.2 Quantification vectorielle :

Dans le cas ou la grandeur à quantifier est composée de plusieurs variables, on parlera de quantification vectorielle (QV) [5].

La collection des représentations possibles d'un vecteur est dite dictionnaire. On utilise en générale plus d'un dictionnaire pour représenter le vecteur. Plusieurs procédures ont été proposées pour créer, organiser et tester les dictionnaires.

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

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

Il n'y a rien de mystérieux à considérer des espaces de grandes dimensions, il suffit de savoir que tout s'organise autour des coordonnées des vecteurs et qu'il n'y a pas lieu de s'imposer une représentation mentale géométrique. A titre d'illustration, nous précisons qu'un vecteur de l'espace R^m est simplement une matrice colonne constituée de k nombres réel $x_i : x=(x_1,x_2,...,x_m)^T$, et que par exemple, une sphère entièrement caractérisée par son centre $u=(u_1,u_2,...,u_m)^T$, et son rayon ρ est constitué de points dont les coordonnées satisfont la relation (1.23)

$$\sum_{m=1}^{m}(x_i-u_i)^2=\rho^2 \qquad (1.23)$$

Nous appellerons distance entre x et $y_i = Q(x)$, généralement notée par $d(x, y)$ le degré de distorsion dû à l'approximation du vecteur d'entré x par le vecteur « arrondi » y_i. Une quantification vectorielle est alors complètement définie par le dictionnaire y et la distance d. en général, la fonction "d" nécessaire à la définition d'une distance entre deux éléments x et y, $d(x, y)$ est défini par l'application :

$$R^k \xrightarrow{d} D \qquad D=\{y_i \in R^k / i=1,2,...,k\}$$

Elle doit avoir les propriétés suivantes :

- $d(x,y) \geq 0$
- $d(x,y) = 0$ si x = y
- $d(x,y) = d(y,x)$ (Symétrie)
- $d(x,z) \leq d(x,y) + d(y,z)$ (inégalité triangulaire).

Dans le cas de la parole, la distance doit avoir deux propriétés supplémentaires :

- $d(x,y)$ doit avoir une interprétation physique
- $d(x,y)$ doit être simple et calculer.

La mesure de la distorsion doit avoir une certaine signification dans le domaine spectral selon les propriétés spectrale de la parole. Les différences, entre l'enveloppe spectrale du signal original et l'enveloppe spectrale du signal codé, qui peuvent conduire à des sons phonétiquement différents sont les suivantes :

- Les formants de l'enveloppe spectrale du signal original et ceux de l'enveloppe spectrale du signal codé se produisent à des fréquences différentes.
- Les bandes de ses formants différent significativement.

Exemples de distances : soit deux vecteurs $x = (\alpha_1, \alpha_2, ..., \alpha_n)$ et $y = (\alpha'_1, \alpha'_2, ..., \alpha'_n)$

$$d(x,y) = \sum_{i=1}^{n} |\alpha_i - \alpha'_i| \qquad \text{Distance de Minkow sky} \qquad (1.24.a)$$

$$d(x,y) = \left[\sum_{i=1}^{n} |\alpha_i - \alpha'_i|^2 \right]^{1/2} \qquad \text{Distance Euclidienne} \qquad (1.24.b)$$

$$d(x,y) = Max|\alpha_i - \alpha'_i| \qquad \text{Distance de Chebychev} \qquad (1.24.c)$$

D'autre mesures de distorsion spectrale peuvent être utilisées selon le contexte telles que: la mesure de la distorsion spectrale logarithmique, la mesure d'ITAKURA SAITO, la mesure cepstral, ...etc.

Par exemple, la distorsion d'ITAKURA-SAITO équation (1.25) 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.

$$d_{IS} = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[e^{V(\omega)} - V(\omega) - 1\right]d\omega \tag{1.25}$$

Avec:

$$V(\omega) = \log(S(\omega) - \log(\widetilde{S}(\omega))) \tag{1.26}$$

$$S(\omega) = \frac{G}{\left|A(e^{i\omega})\right|^2} \qquad \text{G: facteur gain du filtre LP} \tag{1.27}$$

En supposant que la grandeur d'entrée est un vecteur aléatoire distribué selon une loi $p(x)$, les performances du quantificateur peuvent être mesuré par la distorsion moyenne D_Q introduite, c'est à dire par l'espérance mathématique de la distance d :

$$D_Q = E\left[d(x,Q(x))\right] = \int d(x,Q(x).p(x).dx \tag{1.28}$$

Dans la pratique, la distribution des points d'entrée étant généralement inconnue, on approximera D_Q par une distorsion moyenne calculée sur un large nombre d'échantillons $\{x_1, x_2, ..., x_N\}$ de vecteurs d'entrée. L'ergodicité et la stationnarité nous permettent d'écrire :

$$D_Q \cong \frac{1}{N}\sum_{j=1}^{N} d(x_j, Q(x_j)) \tag{1.29}$$

La distance introduit implicitement une partition de l'ensemble des vecteurs d'entrée en k classes $\{S^i, i = 0,1,.....,k-1\}$, la classe S^i étant l'ensemble des vecteurs associés à y_i par le quantificateur (1.30).

$$S^i = Q^{-1}(y_i) = \{x; Q(x) = y_i\} \tag{1.30}$$

Nous appellerons centroïde de la classe S^i le vecteur c^i tel que sa distance moyenne à tout les éléments de la classe soit minimale (en géométrie euclidienne, le centroïde est le centre de gravité) :

$$E\left[d(x,c^i); x \in S^i\right] = \text{Inf}\left\{E\left[d(x,x^i); x \in S^i\right]\right\} \tag{1.31}$$

Étant donné une distance et une taille de dictionnaire, on cherche un quantificateur optimal qui minimise la distorsion moyenne ou qui se rapproche de l'optimalité.

1.4.1.3 Conditions d'optimalité :

Pour une distribution statistique donné de la source et un débit fixé:

Le quantificateur globalement optimal est celui qui minimise la distorsion moyenne.

Un quantificateur localement optimal a un dictionnaire qui peut être légèrement perturbé sans que la distorsion moyenne augmente.

Il n'existe pas de méthode qui décrit la façon de concevoir un dictionnaire globalement optimal pour les quantificateurs vectoriels. Seul des propriétés suffisantes sont connues permettent de construire des dictionnaires localement optimaux.

Un quantificateur se décompose en deux applications : un codeur et un décodeur (Figure1.6). Le quantificateur (localement) optimal est alors celui réunissant les points suivants [5] :

- Un codage optimal (pour un dictionnaire fixé) ; celui-ci respecte « la règle du plus proche voisin » que nous allons décrire ;
- Le décodage optimal (pour une partition S^i donnée), le vecteur représentant y^i doit minimiser la distorsion associée au voronoï S^i, y^i est donc le centroïde de cette cellule : $y^i = cent(S^i)$

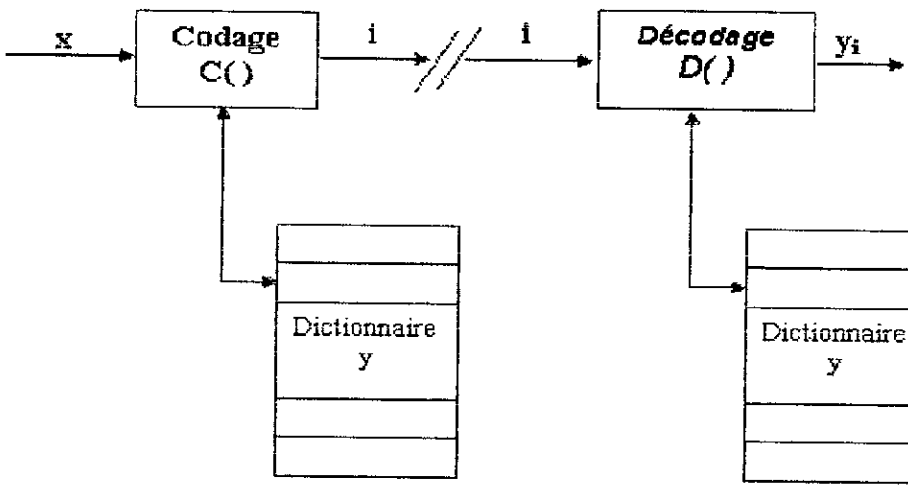

Figure 1.6 Schéma général d'un quantificateur vectoriel

Condition du plus proche voisin :

Etant donné un décodeur et son ensemble fini de mots codes de sortie C, les classes de partition S^i de l'encodeur sont optimales si, elles satisfassent :

$$S^i \subset \left\{x \middle| d(x, y_i) \leq d(x, y_j); \forall j\right\} \tag{1.32}$$

Les régions de partition sont définies par les mots codes $\{y_i\}$ dans C :

$$Q(x) = y_i \text{ seulement si } d(x, y_i) \leq d(x, y_j) \ \forall j \tag{1.33}$$

Condition du centroïde :

Etant donné une partition d'encodeur $P = \left\{S^i \middle| i = 1,.....N\right\}$, les mots codés optimaux y_i dans C sont les centroïdes dans chaque partition S^i :

$$\begin{aligned} &y_i - Cent(S^i) \\ &y_i = \min E(d(x, y) \middle| x \in S^i) \end{aligned} \tag{1.34}$$

1.4.1.4 Construction de quantificateurs statistiques :

Supposons que nous disposons d'une certaine distance d. Construire un quantificateur revient donc à établir une stratégie de choix du dictionnaire associé. Cette stratégie est intimement liée à la nature de la distribution des vecteurs à quantifier.

Dans le cas ou les poins d'entrée sont distribués d'une façon non uniforme, on adoptera une approche statistique visant à tirer partie de cette non-uniformité, Le dictionnaire sera construit par apprentissage : à partir d'une large base de vecteurs d'entrée où sera sélectionné un nombre réduit de points susceptible d'en refléter les propriétés statistiques.

En revanche, si la distribution des vecteurs d'entrée est plutôt uniforme, on aura intérêt à conférer à l'espace de représentation une structure mathématique forte, indépendamment de la réalité des donnée à traiter. Cette approche algébrique utilise généralement les propriétés des réseaux réguliers de points.

1.4.2 Classification des codeurs :

Le classement des codeurs de parole peut se faire selon différentes approches : le débit binaire, le type de codage ...etc.

1.4.2.1 Codeurs par formes d'ondes :

Dans cette catégorie, on distingue les codeurs temporels et fréquentiels. Ces derniers n'utilise aucune connaissance a priori sur la façon dans le signal est généré. Le codeur temporel fait correspondre à l'amplitude du signal analogique une suite d'éléments discrets.

Le signal reconstruit est sans doute le plus proche du signal original. Ces codeurs sont conçus pour être indépendant du signal à coder. Le débit de codage est généralement élevé. En utilisant les propriétés de corrélation du signal, il est possible de diminuer ce débit jusqu'à une certaine limite. En dessous de 16kbits/s la qualité se dégrade et la réduction de débit en bande étroite (2.4 – 4.8kbits/s) est peu envisagée.

L'algorithme de codage le plus simple est le codage appelé PCM[1] correspondant à la norme G.721, il est utilisé pour coder la voix dans le réseau téléphonique. La bande passante d'une paire torsadée étant d'environ 3,5 kHz, la fréquence d'échantillonnage a donc été fixée à 8 kHz afin de respecter le théorème de Nyquist. La quantification est faite avec une échelle logarithmique sur 8 bits, ce qui est équivalent à une quantification linéaire sur 13 bits.

Le codage PCM est à la base d'une famille de codages différentiels largement utilisés. On distingue :

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

1.4.2.2 Les vocodeurs

Utilisent une méthode dite par analyse et synthèse, où on essai d'extraire, du signal parole, un ensemble de paramètres liés à un modèle simplifié. Ces paramètres sont l'enveloppe du spectre court terme et les informations sur le signal d'excitation. On suppose donc qu'on a des connaissances a priori sur le mécanisme de production de la parole. Ces derniers sont sensibles au bruit de transmission et la qualité de la parole est limitée. Le débit de transmission est généralement faible (exemple : codeur LPC-10 à 2,4 kbits/s).

[1] Abréviation du mot anglais « *Pulse Coded Modulation* »

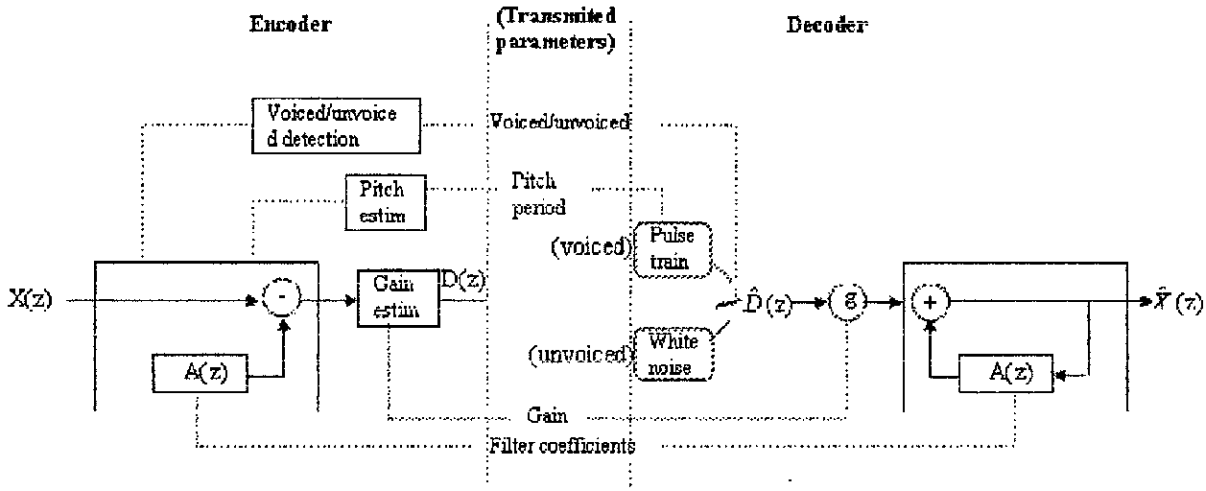

Figure 1.7 Principe du codage LPC

1.4.2.3 Codeurs hybrides:

Ces codeurs font intervenir les techniques d'analyse par synthèse et les techniques de codage par forme d'ondes. Au prix d'une complexité parfois élevée, ils permettent d'obtenir une bonne qualité de signal avec des débits intermédiaires.

1.4.3 Le codage CELP :

Le codage *CELP*[1] est une extension du codage *LPC*. Il comporte toujours deux phases, correspondant aux fonctions d'excitation et de transfert. L'identification de la fonction de transfert est identique à celle faite avec *LPC*. Par contre, la fonction d'excitation n'est pas seulement un bruit blanc ou un sinusoïde, mais une combinaison linéaire de fonctions stochastiques (c'est à dire de bruit) et périodiques. L'identification de ces fonctions est très coûteuse en temps *CPU* (et d'ailleurs les codeurs *CELP* sont en général implémentés avec l'aide des cartes spécifiques de traitement de signal (*DSP*)), mais la qualité obtenu est bien meilleure qu'avec le codeur LPC.

[1] Abréviation de « Code excited linear prediction »

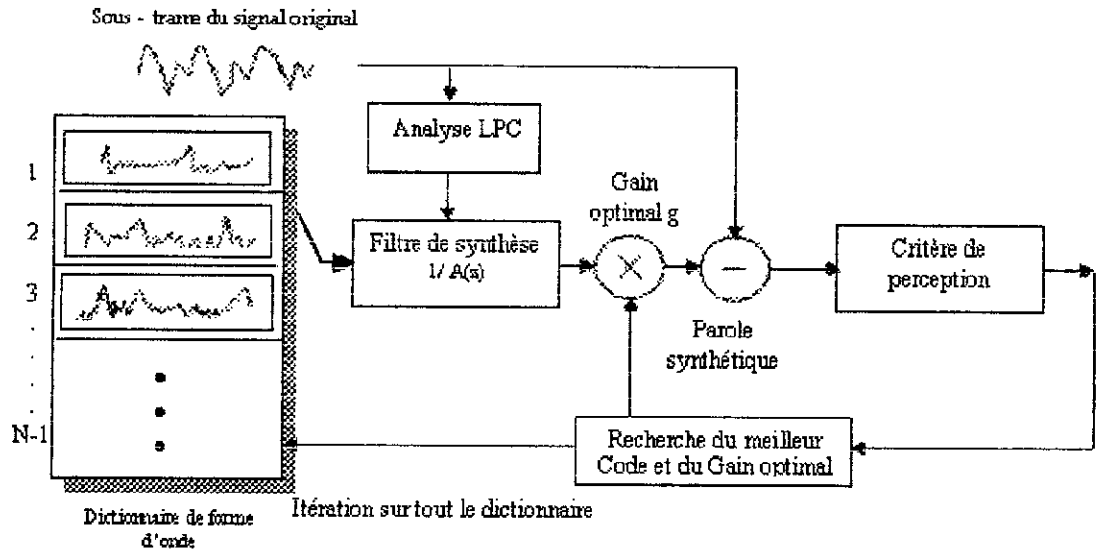

Figure 1.8 Le codage CELP

Dans chaque trame, une analyse spectrale par prédiction linéaire détermine le filtre de synthèse $1/A(z)$. On découpe chaque trame en sous-trames plus courtes (durée typique 5 ms) sur lesquelles on effectue une quantification vectorielle du signal par une technique d'analyse par synthèse. On compare à l'aide d'un critère dit « perceptuel » de type moindres carrés pondérés, le signal de parole original avec tous les signaux synthétiques possibles obtenus après quantification vectorielle. Ces signaux synthétiques sont générés en filtrant par le filtre de synthèse, un signal d'excitation choisi dans un dictionnaire de séquences d'excitation (on ajoute parfois la sortie de plusieurs dictionnaires) et en ajustant le signal résultant par le gain optimal. Le codeur transmet le ou les index des segments qui minimisent le critère ainsi que le ou les gains associés, les paramètres spectraux et le pitch fractionnaire. Le critère perceptuel prend en compte la propriété de masquage du bruit de quantification par les formants en pondérant plus fortement l'erreur de quantification dans les zones de faible amplitude du spectre et plus faiblement dans les zones de formants. Cette pondération s'effectue en filtrant le signal d'erreur par un filtre de type $A(z)/A(z/\gamma)$ où γ est compris entre 0 et 1 (typiquement γ = 0.85). Les dictionnaires utilisés sont appelés stochastiques ou adaptatifs selon qu'ils contiennent des séquences fixes de bruit ou bien les séquences d'excitation de trames précédentes. Le dictionnaire adaptatif permet de prendre en compte la redondance introduite par la quasi-périodicité des sons voisés.

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

Pour les sons voisés, le signal synthétique présente des harmoniques de F_0 jusqu'à $f_e/2$ même si le signal original n'a plus d'harmoniques au-delà d'une fréquence f_{max}.

1.5 Critères de performance dans le codage de la parole :

Le problème essentiel dans la compression du signal est de minimiser le débit binaire dans la représentation numérique toute en maintenant des niveaux adéquats de qualité du signal, de complexité d'implantation et de retard de communication [5].

1.5.1 Qualité du signal :

La qualité du signal perçu est souvent évaluée sur une échelle de 5 point qui est connue comme étant l'échelle MOS (Mean Opinion Score) dans les tests de la qualité de la parole : une moyenne à travers un grand nombre d'entrée parole, locuteurs d'écoute évaluant la qualité du signal. Les cinq points de la qualité sont associés à un ensemble d'adjectifs de description : mauvais, médiocre, inacceptable, bon, excellent. On attribue ainsi un seul niveau à chaque signal parole à évaluer durant la procédure d'évaluation subjective.

1.5.2 Débit binaire :

On mesure le début binaire d'une représentation digitale en bits par échantillon, ou bit 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.

1.5.3 Complexité :

La complexité d'un algorithme de codage est l'effort de calcul exigé pour implanter les processus de l'encodage et du décodage dans le hardware, mesuré en terme de la capacité arithmétique (évalué en *MIPS*) et l'espace mémoire utilisé. D'autres mesures de complexité peuvent être signalées telles que la taille physique du codeur ou du décodeur, le prix et la consommation de puissance (en Watt ou en mW) ce dernier étant un important critère dans un système portable.

1.5.4 Retard de communication :

La complexité dans un algorithme de codage est souvent accompagnée d'une augmentation de la durée de traitement dans le codeur et le décodeur. Bien que l'évolution des capacités des processeurs de traitement du signal, soit un facteur en faveur d'utilisation d'algorithme plus sophistiqué, le besoin de limiter le retard de communication ne doit pas être d'une importance moindre.

Le retard de codage est défini comme étant le temps écoulé entre l'instant où l'échantillon du signal de parole arrive à l'entrée du codeur et l'instant où le même échantillon apparaît à la sortie du décodeur, moins tout retard introduit par les autres équipements de communication, c'est-à-dire comme si le codeur et le décodeur sont directement connectés. Cette définition fait que le retard de codage ne dépend que de l'algorithme de codage. Pour les codeurs *CELP*, le retard de codage peut être grossièrement déterminé en fonction de la taille de la trame du signal de parole.

Le retard de codage consiste en trois catégories :

- Retard algorithmique de bufferisation ;
- Retard de traitement ;
- Retard de transmission binaire.

En pratique, on peut réduire le retard de traitement en utilisant des processeurs plus rapides.

1.6 Mesure de la Qualité :

- Pour mesurer la qualité du signal, il existe deux types de mesure, la mesure objective et la mesure subjective. Les mesures objectives de la qualité de la parole sont purement des mesures mathématiques évaluées en utilisant des distances euclidiennes, les mesures subjectives de qualité évaluent la qualité de codage par des tests d'écoute.

La mesure objective de la qualité la plus couramment utilisée, pour les codeurs qui essaient de préserver la forme du signal, reste le rapport signal à bruit (*SNR*).

Si : S est le signal de parole original, et $\overline{S}$ est le signal de parole synthétisé.

Alors le signal d'erreur est donné par (1.35)

$$e(n) = S(n) - \overline{S}(n) \tag{1.35}$$

Pour un signale de N échantillons, on définit l'énergie du signale

$$E_S = \sum_{n=0}^{N-1} S^2(n) \tag{1.36}$$

Et l'énergie de l'erreur :

$$E_e = \sum_{n=0}^{N-1} e^2(n) \tag{1.37}$$

Le *SNR* est alors donné par :

$$SNR = 10 \log\left(\frac{E_S}{E_e}\right) \text{en } dB \tag{1.38}$$

Le signal du parole est par nature non constant .Certains segments du signal peuvent avoir une énergie plus ou moins grande. En supposant que l'énergie de l'erreur soit à peu près constante, le *SNR* pourra être très important comme il peut être très faible. Alors, on utilise plutôt le *SNR* segmental, le signal est découpé en M segments de 15 à 30 ms puis on calcule une moyenne des *SNR*.

$$SNR_{seg} = \frac{1}{M} \sum_{i=10}^{M} 10 \log\left(\frac{\sum_{n=0}^{N-1} S^2(n)}{\sum_{n=0}^{N-1} e^2(n)}\right) \tag{1.39}$$

Les essais d'écoute sont nécessaires, car le récepteur humain représente le dernier bloc d'un système de codage de la parole. De plus, le *SNR* n'est pas nécessairement corrélé avec la qualité d'écoute. La méthode la plus utilisée dans les mesures subjectives est celle dite « Mean Opinion Score » (*MOS*), où des auditeur évaluent un codeur sur une échelle absolue allant de 1 à 5 (Tableau 1.2).

Tableau 1.2 Qualité du signal vocal avec la mesure MOS

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

1.7 Conclusion :

Le codage consiste à réduire le volume d'informations à transmettre en gardant une qualité acceptable de la parole. La connaissance de la façon de la production de la parole chez l'être humain permettent de pouvoir utiliser les propriétés de ce signal pour la réduction du débit de l'information.

Ainsi, la prédiction linéaire essaye d'exploiter la redondance dans le signal et d'extraire des coefficients (paramètres LPC) qui caractérisent le comportement du signal. La simplicité de concept, la résolution linéaire dans la prédiction linéaire, et ses performances dans le codage de la parole, sont sans doute celles qui la rendent la méthode la plus communément admise et la plus largement utilisée dans le codage du signal de parole.

Chapitre 2

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

2.1 Introduction :

Avec l'augmentation continue de la vitesse des microprocesseurs et le développement des techniques de traitement du signal, il est devenu réaliste de faire transmettre la voix, au même titre que des données informatiques, sur le réseau Internet. Hormis l'intérêt technologique, la téléphonie sur le réseau semble avoir un intérêt économique évident en autorisant des communications vocales à des tarifs pour le moment imbattables.

Or toutes les organisations -entreprises, administrations, associations- qui utilisent le téléphone sont à l'affût de sources d'économies. Elles sont donc naturellement intéressées par toutes les innovations dans ce secteur, d'autant que la mise en concurrence en matière de télécommunications devient la règle. Cependant, la téléphonie sur Internet est encore loin de satisfaire aux exigences de qualité de service attendues pour ce type de service, même si de fortes améliorations sont prévisibles.

2.2 La voix sur les réseaux IP :

La voix sur les réseaux IP, ou *VoIP*, est le transfert des conversations vocale sous forme de données sur un réseau IP. Contrairement aux réseaux traditionnels à commutation de circuit (RTCP) ; dans les appels VoIP, la connexion téléphonique est à commutation de paquets.

Avec un appel VoIP, l'établissement de l'appel doit être simulé c'est-à-dire la tonalité, les signaux de sonneries et les signaux d'occupation. En plus l'appel lui-même (c'est-à-dire la conversation) a besoin d'être converti de son format analogique à un format numérique, découpé en paquets, et envoyé à travers le réseau, reassemblé de nouveau, et reconverti du format numérique au format analogique. Les Codecs à chaque point font la conversion de l'analogique au numérique et vice vers ça.

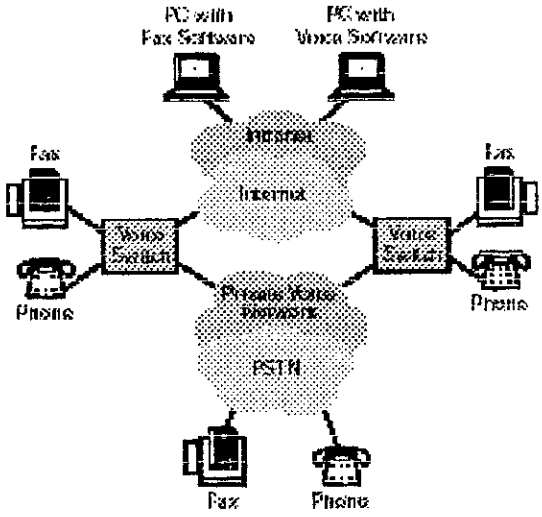

Figure 2.1 Infrastructure du VoIP [7]

2.3 Codecs employés en VoIP :

Un codec (acronyme de « codeur/décodeur ») est le hardware ou software qui échantillonne les sons analogiques, et les convertis aux formats numériques avec un débit binaire prédéterminé. Les codecs font aussi de la compression afin d'économiser la bande passante. Il existe sur le marché des douzaines de codecs, chacun avec ses propres caractéristiques.

Les codecs ont des noms, attribués par l' ITU, et qui décrivent les fonctions qu'ils accomplissent. Par exemple, les codecs nommé **G.711u** et **G.711a** font la conversion de l'analogique au numérique et du numérique à l'analogique en assurant, relativement, une bonne qualité. Comme avec la plupart des appareils numériques, plus on requiert de qualité et plus il y a de bits, donc ces deux codecs utilisent plus de ressource (bande passante) que des codecs dont la vitesse est inférieure.

Les codecs à vitesse inférieure, tel que **G.726**, **G.729**, et ceux dans la famille **G.723.1**, consomment moins de bande passante. Cependant, leur qualité sonore est beaucoup plus affaiblie par rapport aux codecs à grande vitesse, ceci est dû à la nature de la compression qu'ils effectuent, c'est-à-dire : la compression avec perte -compression qui perd quelques bits des données originales. Moins de bits sont envoyés, donc le récepteur fait de son mieux pour se rapprocher du signal vocal original émis, cependant cette recréation atteint très vite ses limites de fidélité.

Le tableau ci-dessous (Tableau 2.1) décrit les principaux codecs utilisés en VoIP. avec leurs techniques de codage, les débits avec lesquels les codecs génèrent leurs sorties, le degré de complexité, le retard introduit, et enfin la qualité résultante de chaque type de codage.

Tableau 2.1 Les principaux codecs en VoIP

Standards	Méthode	Débit (kbits/s)	Retard (ms)	Complexité (MIPS)	Qualité (MOS)
G.711	LOG PCM	64	0.125	0.01	4.1
G.726	ADPCM	32	0.125	2	3.85
G.728	LD-CELP	16	0.625	30	3.61
G.729	CS-ACELP	8	15	20	3.92
G.729A	CS-ACELP	8	15	10.5	3.7
G.723.1	ACELP	5.3	37.5	16	-
	MP-MLQ	6.3	37.5	14.6	-
IS-54	VSELP	7.95	20	14	3.54
GSM-FR	RPE-LTP	13	20	6	3.5
GSM-EFR	ACELP	12.12	20	-	-
GSM-HR	VSELP	5.10	20	-	-

2.4 Masquage des Paquets perdus :

La transmission de la voix sur Internet (réseau IP) se fait par des paquets. Au récepteur, certains paquets peuvent manquer, dû aux délais, à l'encombrement ou aux erreurs de transfert. Cette perte de paquets dégrade la qualité de la voix reçue dans un système de transmission IP. Etant donné que la transmission de la voix est effectuée en temps réel, le récepteur ne peut pas requérir à la retransmission des paquets perdus à cause des délais de transferts trop importants. Des algorithmes de masquage des pertes (Packet Loss Concealement) sont utilisés alors au niveau de l'émetteur ou du récepteur afin de combler la perte des paquets [8].

La PLC est une option supplémentaire disponible chez quelques codecs, les techniques PLC réduisent ou masquent les effets de perte de données.

Dans cette section nous exposons les différentes techniques utilisées pour récupérer les paquets perdus. Ces technique peuvent être divisées en deux classes : L'une est basée sur

l'émetteur (sender-based), l'autre est basée sur le récepteur (receiver-based). Comme indiqué sur la figure 2.2.

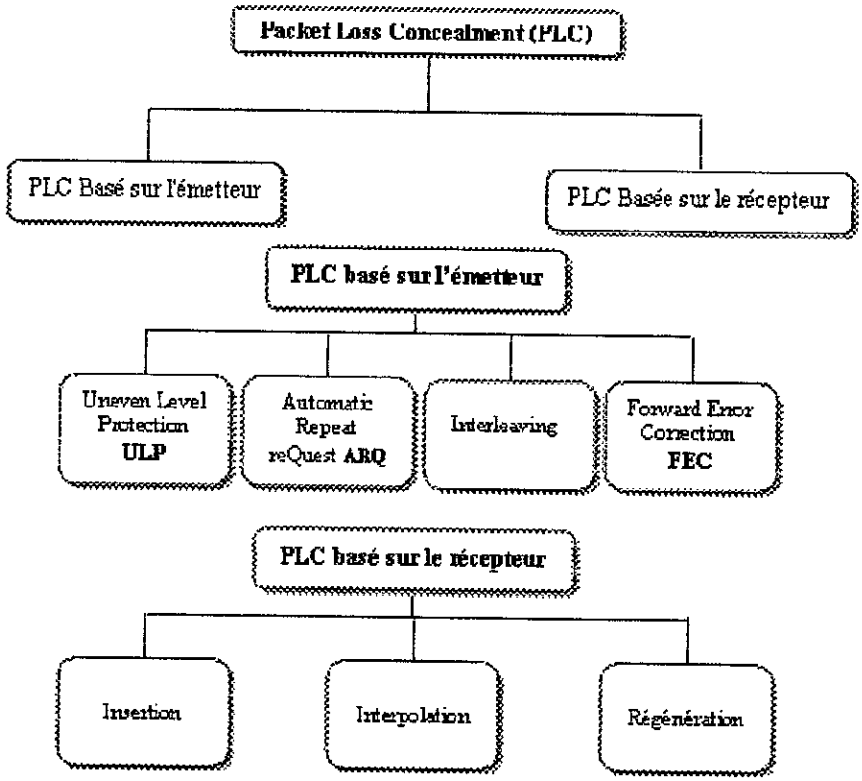

Figure 2.2 Les techniques de masquage des pertes [10]

2.4.1 Masquage basé sur l'émetteur :

Nous résumons ici, quelques techniques de masquage qui exigent la participation de l'émetteur afin de masquer les trames perdues.

2.4.1.1 Forward Error Correction (*FEC*) :

Plusieurs techniques *FEC* ont été développées pour masquer les pertes de données pendant la transmission. Ces techniques consistent à l'addition de données de redondance au flux binaire transmis à partir desquelles le contenu des paquets perdus peut être récupéré. Il y a deux genres d'informations de redondances 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 la connaissance de la donnée à transmettre.

Les données redondantes sont provenues des données originales en utilisant l'opération Ou exclusif (*XOR*) : un paquet de parité est généré pour les k paquets de donnés originaux.

La *FEC* transmet k paquets de données originaux (D), et h paquets supplémentaire de parité redondants (P). La figure 2.3 présente un exemple pour k=3 et h=2. Le codeur *FEC* produit deux paquets redondants (P_1, P_2) des trois paquets de donnée. Si un paquet de donnée (exp. D_3) et un paquet de parité (exp. P_1) sont mal reçus, le récepteur peut récupérer le paquet perdu (D_3) en utilisant les bonnes paquets reçus, D_1, D_2, et P_2.

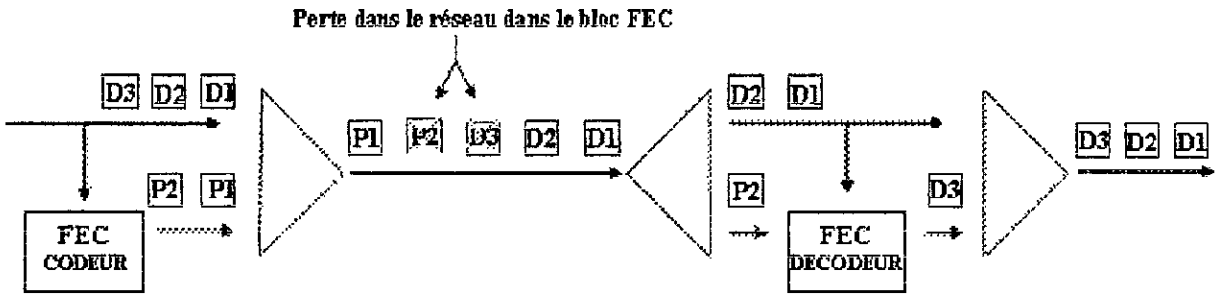

Figure 2.3 Exemple du *FEC* [11]

La *FEC* reste efficace pour un faible rapport de $\frac{h}{k}$. Pour le décodeur *FEC*, les pertes de paquets consécutifs peuvent être corrigées pour une grande valeur de k, si k augmente, le délai de la reconstruction au niveau du récepteur augmente aussi. On peut citer plusieurs avantages (👍) et inconvénient (👎) du *FEC*.

👍) L'opération de la *FEC* ne dépend pas du contenu des données originales, et la réparation est le remplacement exact du paquet perdu.

👍) Le paquet original de la donnée peut être utilisé par des récepteurs qui ne sont pas compatible avec la *FEC*, puisque les données redondantes sont envoyées habituellement comme un flux séparé.

👎) Le codeur *FEC* exige un délai supplémentaire et une bande passante additionnelle pour un codage et décodage efficaces.

2.4.1.2 Interleaving :

Interleaving est une technique utile pour réduire les effets de pertes [11]. Si la dimension de la trame de donnée est plus petite que la dimension des paquets transmis, alors quelques trames peuvent être combinées dans un seul paquet. Cependant afin de réduire les effets de

pertes des paquets, les trames originales de la donnée ne sont pas combinées dans le même ordre séquentiel tel qu'ils sont produites par le codeur, mais ils sont entrelacés par l'émetteur comme il est illustré sur la figure 2.4.

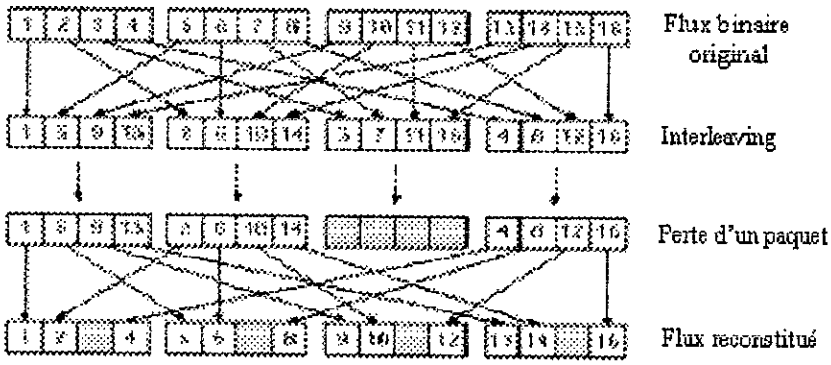

Figure 2.4 Exemple d'*Interleaving* [11]

Au récepteur, les trames de données sont rassemblées à leur ordre original. Comme peut être vu sur la figure 2.4, l'effet d'une perte de paquet est réparti sur des petits intervalles correspondant aux trames de données distribuées au lieu d'être adjacentes. La réduction de l'effet des pertes est dûe aux raisons suivantes:

- Les petits intervalles vides résultants correspondent typiquement aux intervalles de la parole qui sont considérablement plus courts qu'une longueur de phonème. Par conséquent, les humains peuvent interpoler mentalement les intervalles vides, et l'intelligibilité de la parole ne sera pas diminuée. C'est contrairement à la situation du no-interleaving où un simple paquet perdu peut avoir comme conséquence un phonème complet perdu, ce qui diminue l'intelligibilité de la parole.
- Si le récepteur emploie une certaine technique de masquage des erreurs (par exemple les lacunes dues à la perte de paquet sont remplies en utilisant l'interpolation des trames de données adjacentes reçues), alors un rendement plus élevé est obtenu, si l'interpolation est effectuée pour des petits intervalles au lieu de longs intervalles.

Au-dessous, nous récapitulons les avantages (👍) et les inconvénients (👎) de cette technique :

👍) La réduction de l'effet de perte est due à la distribution de l'éventuelle perte sur des petits intervalles disjoints.

- Aucune redondance de données.
- Un retard supplémentaire est introduit, ce qui pourrait être inacceptable pour quelques applications.

2.4.1.3 Automatic Repeat reQuest (*ARQ*) :

ARQ (*Automatic Repeat Request*) est une technique de retrtansmission , dont les stratégies de base consistent en trois parties :

- La détection du paquet perdu se fait par le récepteur ou par l'émetteur.
- La stratégie de Reconnaissance : Le récepteur envoie des informations qui indiquent quelles données sont reçues ou quelles données manquent.
- La stratégie de Rediffusion : Elle détermine quelles données seront retransmises par l'émetteur.

Bien que sa robustesse contre les brusques pertes, elle ne peut pas être utilisé dans les applications en temps réel, tel que *VoIP*, à cause du délai considérable et à la large bande passante nécessaire.

2.4.1.4 Uneven Level Protection (*ULP*) :

Lorsque les subdivisions constituant la donnée, ne sont pas en même niveau d'importance (Les données de la parole, en particulier), une technique dite *ULP* (*Uneven Level Protection*) peut être appliquée, cette technique attribut aux données les plus importantes plus de protection. Elle est très souvent employée avec la technique *FEC* (Pg.2.4.1.1), Les unités de données sont arrangées dans un paquet de type *RTP*[1] (*Real-time Transfert Protocol*) par ordre d'importance descendante, et plus de protection est appliquée aux débuts des données, c.-à-d. aux données les plus importantes.

2.4.2 Masquage basé sur le récepteur :

Dans cette section, nous résumons plusieurs techniques de masquage de pertes qui peuvent être effectuées au niveau du récepteur et qui n'exigent pas la contribution de l'émetteur [11]. Ces techniques sont en générale moins efficaces que les techniques précédentes (Basées sur l'émetteur).

[1] Protocole d'Internet, fournit un moyen uniforme de transmettre sur IP des données soumises à des contraintes de temps réel (audio, vidéo,...).

Elles consistent à produire des remplacements semblables aux paquets perdus originaux. Ainsi, ces techniques fonctionnent bien pour des taux de pertes relativement faibles (< 15%) et pour de petites lacunes (< 40ms). Quand la longueur des pertes consécutives approche de la longueur d'un phonème, ces techniques ne fonctionnent pas correctement. Il existe trois catégories de méthodes de dissimulation : Insertion, Interpolation, et Régénération.

2.4.2.1 Insertion :

Cette technique de réparation génére un remplacement pour un paquet perdu en insérant une simple donnée d'appoint. Il est à mentionner que cette technique ne prend plus en compte les caractéristiques du signal, ce qui la rend simple à implémenter, la donnée remplaçante peut être de natures différentes, à savoir : un silence, un bruit, ou bien une version répétée de la dernière bonne trame reçue, telles techniques sont faciles à implémenter, mais à l'exception de la technique répétitive, ont de pauvres performances.

Substitution par un silence : La substitution consiste à combler la lacune avec un silence afin de maintenir la succession temporelle des paquets. Elle est seulement efficace avec des longueurs courtes de paquet (< 4ms) et de bas taux de perte (< 2 %). Sa performance se dégrade rapidement quand les tailles de paquets augmentent (la qualité est mauvaise pour une taille de paquets de 40 ms). Elle est couramment utilisée dans les réseaux de communication audio. En dépit de ceci, l'utilisation de ce type de substitution est répandue, parce qu'il est simple à implémenter.

Substitution par un bruit : Puisque la substitution de pertes par un silence présente une mauvaise performance, une autre méthode a été introduite, et consiste à remplacer la trame perdue par un bruit de fond. En outre, une fois comparée au silence, l'utilisation du bruit blanc a donné une qualité subjective meilleure [11] et une intelligibilité améliorée.

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

2.4.2.2 Interpolation :

Une autre méthode très intéressante peut être appliqué dans ce domaine, 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 interpolatives par rapport aux celles d'insertion est qu'elles prennent en compte le changement des caractéristiques du signal.

2.4.2.3 Régénération :

Les techniques de régénération profite de la connaissance à priori de l'algorithme de compression des signaux audio pour récupérer les paramètres du codec, donc le signal audio dans un paquet perdu peut être synthétisé. Ces techniques sont nécessairement dépendantes du codec, mais ils sont plus performants, en raison de la grande quantité d'informations d'état utilisées dans la réparation.

2.5 Conclusion :

Au cours de ce chapitre, nous avons donné un petit aperçu sur la transmission de la voix via les réseaux *IP*, et on a décrit les principaux codecs les plus employés dans ce secteur, chaque codec a sa propre méthode de codage qui définit sa qualité et ses performances.

Nous avons abordé aussi un des problèmes affectant la qualité de service, c'est la perte de trames lors de la transmission, et nous avons discuté les différentes méthodes existant et qui essaient de masquer ces pertes, Ces techniques peuvent être appliquées au niveau de l'émetteur ou de récepteur, chaque technique présente une certaine complexité, et requière des exigences liées à la méthode de masquage.

Chapitre 3

Le codec de l'ITU « G.729 »

3.1 Introduction :

Ce chapitre décrit un des codecs les plus utilisés en codage de signaux vocaux dans les applications VoIP (Pg.2.3). Ce codeur est fondé sur le modèle de codage prédictif linéaire à excitation par séquences codées à structure algébrique conjuguée (CS-ACELP) (*conjugate-structure algebraic-code-excited linear-prediction*).

Il est conçu pour fonctionner avec un signal numérique que l'on obtient en effectuant d'abord un filtrage du signal analogique d'entrée dans la bande téléphonique (3400 Hz), puis en l'échantillonnant à 8000 Hz et en le convertissant en signal PCM linéaire à mots de 16 bits, qui est injecté dans le codeur. Et inversement, on reconvertira le signal de sortie du décodeur en signal analogique.

3.2 Description générale du codec G.729 :

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

La méthode utilisée, pour déterminer les coefficients du filtre et l'excitation, est appelée « analyse par synthèse » : Le codeur cherche les paramètre de codage, en appelant la procédure de décodage dans chaque boucle de la recherche, et en comparant le signal décodé (le signal synthétisé) avec le signal original, les paramètres les plus proches de l'original sont choisis, codés, et puis transmis aux récepteurs, selon l'allocation des bits représentée sur le tableau 3.1. Au niveau du récepteur, ces paramètres sont utilisés pour reconstruire le signal de parole original.

3.2.1 Le Codeur :

Le schéma bloc du codeur est représenté sur la figure 3.1. Pour chaque trame de 10 ms, le codeur exécute une analyse prédictive linéaire pour calculer les coefficients du filtre LP. Cette analyse est basée sur l'idée, qu'un échantillon de parole peut être approximé comme la combinaison linéaire des échantillons passés. En minimisant la somme des différences carrées (l'erreur quadratique) entre les échantillons de parole réels et ceux approximés, un ensemble de coefficients de filtre peut être trouvé. Le filtre linéaire composé par cet ensemble de coefficients est appelé « le filtre d'analyse », c'est à dire lorsqu'il est attaqué par le signal de parole, sa sortie est le vecteur d'excitation de ce signal. Le filtre de synthèse est obtenu en inversant le filtre d'analyse. Quand nous filtrons l'excitation par ce filtre, la sortie est une approximation du signal vocal original.

Pour des raisons de stabilité et efficacité, les coefficients du filtre de prédiction linéaire ne sont pas quantifiés directement, mais sont transformés en paires de raies spectrales (LSP), puis quantifié en utilisant une quantification vectorielle prédictive à deux étages. L'excitation du signal vocal, est calculée pour chaque sous-trame de 5 ms (ce qui correspond à 40 échantillons PCM), et elle a deux composants : contribution du codebook fixe et celle du codebook adaptatif. Premièrement, un délai tonale est estimé en boucle ouverte pour la trame de 10 ms. Cette estimation est basée sur l'autocorrélation du signal vocal pondéré, qui est obtenu par un filtrage du signal de parole avec un filtre de pondération perceptive, la contribution du codebook adaptatif modélise la corrélation à long terme des signaux vocaux, et elle est présentée par le délai tonal de la boucle fermée et un gain [1]. Le délai tonal de la boucle fermée est cherché autour du délai tonal de la boucle ouverte en minimisant l'erreur quadratique pondéré entre le signal vocale originale et le signal reconstitué. La différence entre l'excitation trouvée, filtrée par le filtre de synthèse, et le signal original, est utilisée pour trouver la contribution du codebook fixe. Le vecteur et le gain du codebook fixe sont onbtenus en minimisant l'erreur quadratique moyenne entre le signal d'entrée pondéré et le signal vocal reconstitué, en utilisant un train d'impulsions comme excitation. Le gain du codebook adaptatif et celui du codebook fixe sont conjointement[1] quantifiés avec une quantification vectorielle.

[1] Ce type de quantification conjointe représente le *CS* (Conjugate structure) dans le nom du codec

Tableau 3.1 Affectation des bits dans l'algorithme de codage CS-ACELP à 8kbit/s (Trame de 10ms)

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 dictionnaire de code adaptatif	P1, P2	8	5	13
Parité du délai tonal	P0	1	-	1
Index du dictionnaire fixe	C1, C2	13	13	26
Signe du dictionnaire fixe	S1, S2	4	4	8
Gains du dictionnaire (étage 1)	*GA1, GA2*	3	3	6
Gains du dictionnaire (étage 2)	*GB1, GB2*	4	4	8
Total				80

Le tableau 3.2 énumère les symboles utilisés dans les deux figures, 3.1, et 3.2

Tableau 3 2 Glossaire des symboles utilisés sur les figures du codec G.729

Désignations	Description
A(z)	Filtre d'analyse (coefficients LP).
P(z)	Pré-filtre du dictionnaire fixe.
c(n)	Contribution du dictionnaire fixe.
d(n)	Correlation entre le signal cible et la réponse Impultionnelledu filtre ded synthèse pondéré.
h(n)	Reponse impultionnelle du filtre de synthèse pondéré.
$S(n)$	Signal vocal pré-traité.
x(n)	Le signal cible.
v(n)	Contrribution du dictionnaire adaptatif.
g_p	Gain de la contribution du dictionnaire adaptative.
g_c	Gain de la contribution du dictionnaire fixe.
γ	Coefficient de ponderation.
L_0	Modee de predicteur a mpyenne ajustée.
L_1	Première étage du dictionnaire des LSP.
L_2	deuxième étage du dictionnaire des LSP(partie inferieure).
L_3	deuxième étage du dictionnaire des LSP(partie superieure).
P_0	Bit dee parité de code du délai tonal.
P_1, P_2	Mot de code du délai tonal de la 1ère et la 2ème sous-trame respectivement.
S_1, S_2	Signes des impultions du dictionnaire fixe.
C_1, C_2	Mots de codes du dictionnaire fixe de la 1ère et la 2ème sous-trame resp.

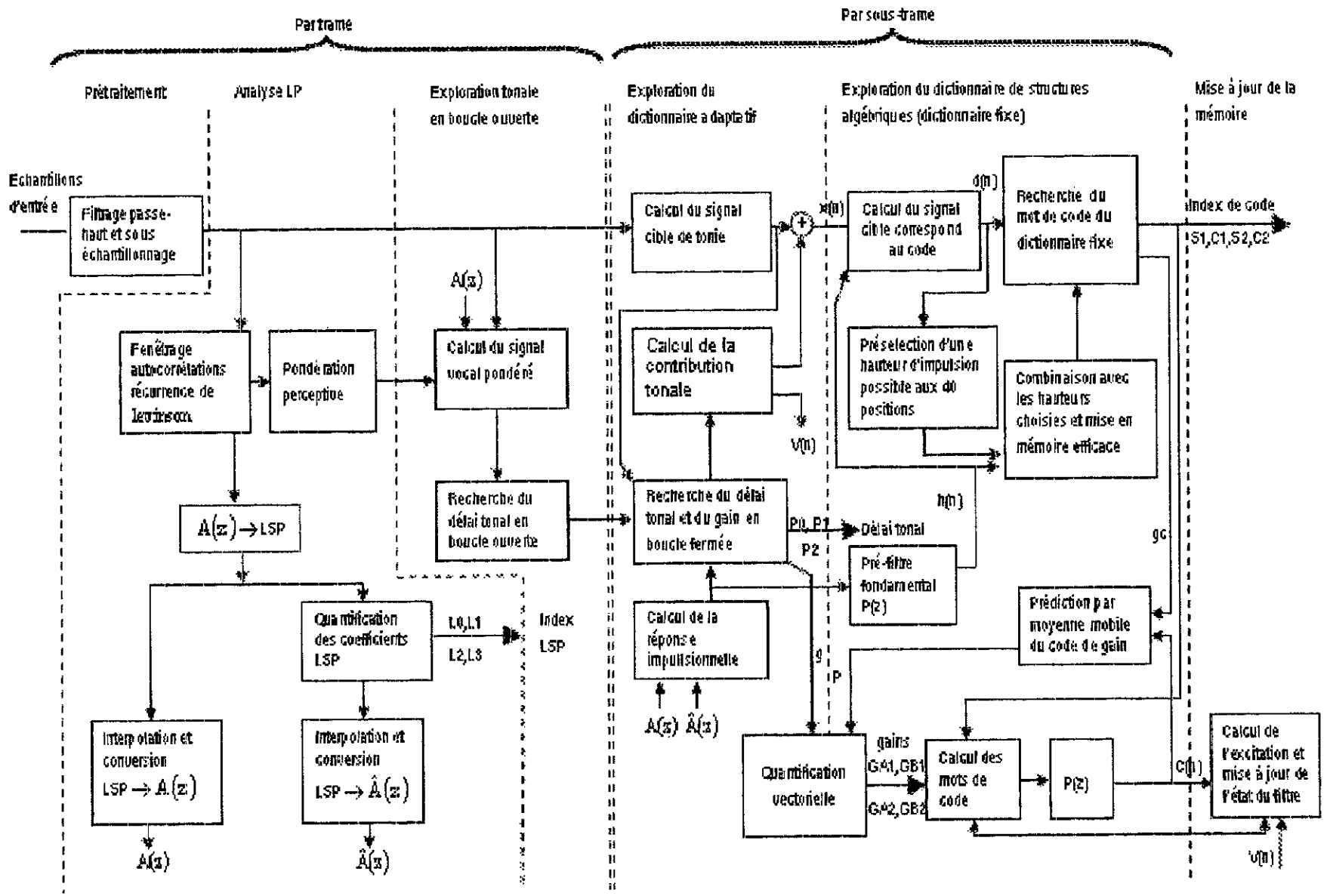

Figure 3.1 Schéma bloc du codeur du G.729

3.2.2 Le Décodeur :

Le principe de fonctionnement du décodeur est représenté sur la figure 3.2. A l'arriver du signal codé au récepteur, le décodeur du G.729, extrait les paramètres suivants :

- Les coefficients de paires de raies spectrales (LSP) ;
- Les deux délais tonals ;
- Les deux mots de code représentant le vecteur du codebook fixe ;
- Les gains des codebooks fixe et adaptatif.

Pour chaque sous-trame, les LSP sont transformé en coefficients LP du filtre de prédiction linéaire, et puis la reconstitution du signal de parole suit les étapes suivantes :

- L'excitation est la somme des deux vecteurs, du codebook fixe et celui du codebook adaptatif, multipliés par leurs gains respectifs.
- Le signal vocal est obtenu en attaquant le filtre de synthèse par l'excitation.
- Le signal de parole reconstitué est filtré par un filtre de post-traitement qui comprend un post-filtre adaptatif basé sur les filtres de synthèse à long terme et à court terme, suivi d'un filtre passe-haut. Cette opération de filtrage réduit la distorsion perceptuelle et accroissent la qualité de la parole synthétisée [1].

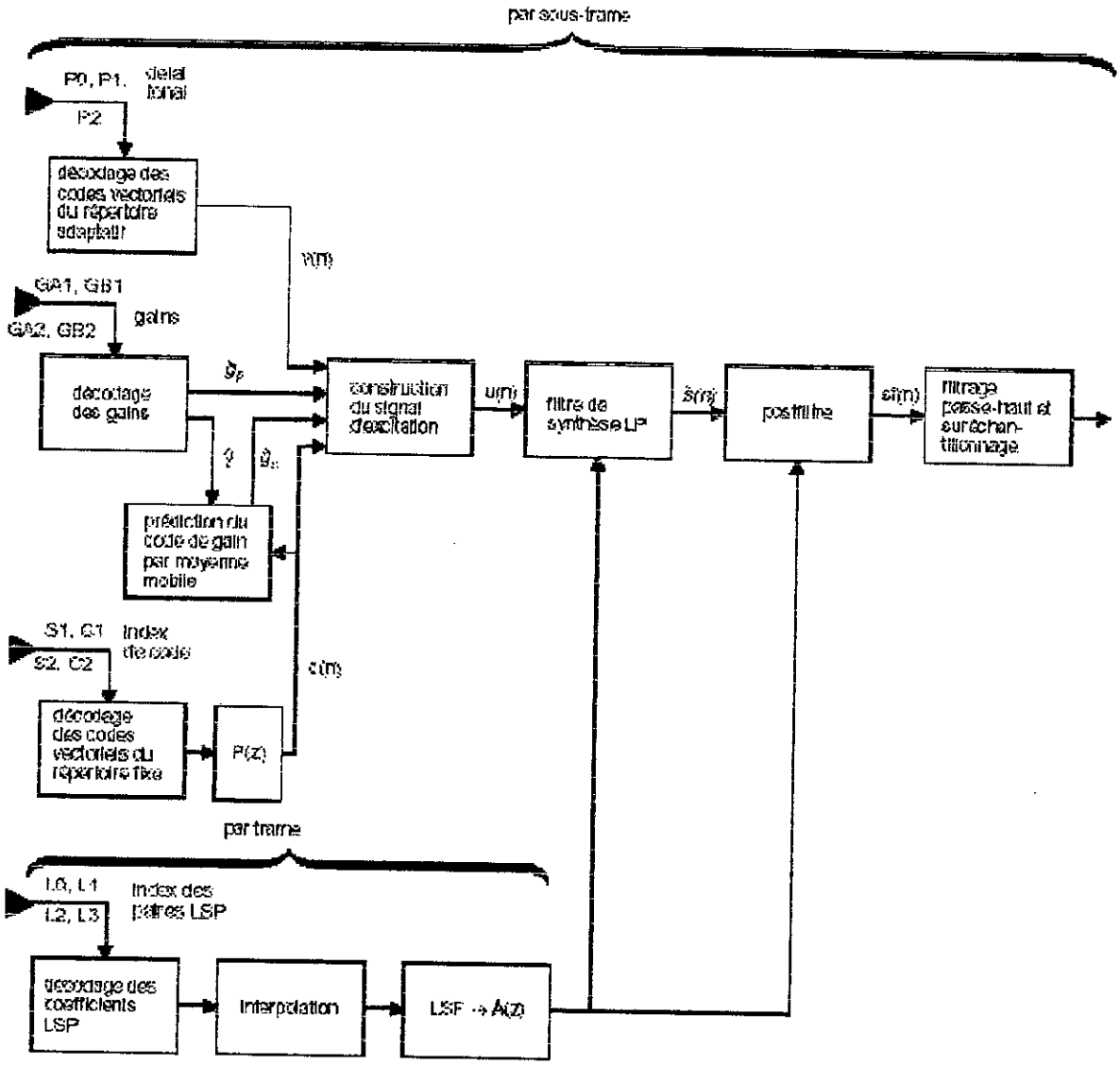

Figure 3.2 Schéma bloc du décodeur du G.729

3.3 Quantification des coefficients LP :

Les coefficients des paires de raies spectrales, q_i, sont quantifiés par application de la représentation des fréquences LSF, ωi, dans le domaine fréquentiel normalisé [0, π]; c'est-à-dire: $\omega_i = \arccos(q_i)$ $i = 1,...,10$ (18), et afin de réduire la bande passante, le codec G.729 utilise une prédiction par moyenne mobile périodique du 4ème ordre pour prédire les coefficients LSF de la trame courante. La différence entre les coefficients calculés et les coefficients prédits est quantifiée au moyen d'un quantificateur vectoriel à deux étages. Le premier étage est un quantificateur vectoriel à 10 dimensions qui utilise le dictionnaire L1

avec 128 niveaux (7 bits). Ce vecteur à 10 dimensions est alors soustrait du vecteur LSF original. Le vecteur résultant est devisé en deux vecteurs à cinq dimensions qui seront quantifiés séparément, avec deux dictionnaires à 5 dimensions, L2 et L3 contenant 32 entrées (5 bits) chacun.

3.4 Procédure de masquage des trames effacées du G.729 :

Une procédure de masquage des erreurs a été incorporée dans le décodeur afin de réduire la dégradation dans le signal vocal reconstitué en raison d'effacements de trame dans le flux binaire. Ce processus de masquage des erreurs est fonctionnel lorsque la trame des paramètres du codeur (correspondant à une trame de 10 ms) a été identifiée comme étant effacée.

La stratégie de masquage consiste à reconstruire la trame actuelle sur la base de l'information déjà reçue. Cette méthode remplace le signal d'excitation manquant par un signal de caractéristiques similaires, tout en diminuant progressivement son énergie. Pour cela, on utilise un classificateur d'éléments voisés utilisant le gain de prédiction à long terme, qui est calculé dans le cadre de l'analyse par post-filtre à long terme. Celui-ci trouve le prédicteur à long terme pour lequel le gain de prédiction est supérieur à 3 dB [1]. Pour cela, on fixe un seuil de 0,5 pour le carré de la corrélation normalisée. Pour le processus de masquage d'erreur, une trame de 10 ms est déclarée « périodique » si au moins une sous-trame de 5 ms possède un gain de prédiction à long terme supérieur à 3 dB, et dans ce cas seul le dictionnaire de code adaptatif est utilisé et la contribution du dictionnaire de code fixe est mise à zéro, Le délai tonal est fondé sur la partie entière du délai tonal contenu dans la trame précédente. Ce délai est répété pour chaque trame successive, Sinon, la trame actuelle est considérée également comme « apériodique » et la contribution du dictionnaire de code adaptatif est mise à zéro, la contribution du dictionnaire de code fixe est construite par sélection aléatoire d'un index de dictionnaire et d'un index de signe.

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

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

2) affaiblissement des gains du dictionnaire adaptatif et celui du dictionnaire fixe.

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

4) production de l'excitation de remplacement.

Répétition de paramètres du filtre de synthèse :

Le filtre de synthèse pour une trame effacée utilise les paramètres de prédiction linéaire de la dernière bonne trame.

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

Affaiblissement des gains du dictionnaire adaptatif et celui du dictionnaire fixe :

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

$$g_c^{(m)} = 0.98 g_c^{(m-1)} \tag{3.1}$$

où m est l'index de sous-trame.

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

$$g_p^{(m)} = 0.9 g_p^{(m-1)} \text{ avec la limite } g_p^{(m)} < 0.9 \tag{3.2}$$

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

Le prédicteur de gain utilise l'énergie des vecteurs de code du dictionnaire fixe qui ont été précédemment sélectionnés, c(n). Afin d'éviter des effets transitoires dans le décodeur, la mémoire du prédicteur de gain est rafraîchie dès que des trames normales sont reçues, au moyen d'une version affaiblie de l'énergie de la contribution du dictinnaire[1].

Production de l'excitation de remplacement :

L'excitation utilisée dépend de la classification de périodicité. Si la dernière trame reconstituée a été classifiée comme étant périodique, la trame actuelle est également considérée comme périodique. Dans ce cas, seul le dictionnaire adaptatif est utilisé et la contribution du dictionnaire fixe est mise à zéro. Le délai tonal est fondé sur la partie entière du délai tonal contenu dans la trame précédente. Ce délai est répété pour chaque trame successive. Afin d'éviter une périodicité excessive, le délai est augmenté de 1 à chaque sous-trame successive mais jusqu'à une limite de 143[1]. Le gain de la contibution adaptative est fondé sur une valeur affaiblie selon l'équation (3.2).

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

3.5 Conclusion

Dans ce chapitre, nous avons décrit un des codecs les plus populaire en transmission de la voix sur les réseaux IP.

Nous avons essayé de résumer son principe de fonctionnement, et surtout les partie qui nous interesse « la quantification des LSF, et la methode de madquage des pertes ».

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

Chapitre 4

Simulation, résultats et interprétation

4.1 Introduction :

Quand des paquets de parole sont envoyés en temps réel à travers des réseaux IP, il n'y a aucune garantie de les recevoir dans une manière appropriée, ce qui est dû à la nature « *best effort* » des réseaux IP. Quand un ou plusieurs paquets sont perdus, et aucun effort n'est fait pour les récupérer, la qualité perceptuelle de la parole reçue peut se détériorer considérablement.

Plusieurs méthodes peuvent être proposés pour alléger cet effet, et elles sont souvent classées en deux catégories : masquage basé sur le codeur "encoder-based" ou sur le décodeur "decoder-based".

La technique *FEC* (Pg.2.4.1.1) [11] est la plus connue, où des trames de parole de redondance sont enchaînés, avec un retard, avec les paquets sélectionnés. Si une trame est perdue, la version redondante retardée de cette trame peut être reçue correctement pour la décoder.

Les méthodes *FEC* sont efficaces, si la perte dans le réseau est prévisible, et si une bande passante supplémentaire est disponible. Pour les applications à bande passante limitée, les méthodes de masquage par le décodeur (*decoder-based*) deviennent importantes. Ces dernières sont convenables pour les trames de parole codées par le codage CELP puisque plusieurs paramètres de ce type de codage présentent une bonne corrélation *inter-trame.*

Le tableau 4.1 présente l'autocorrélation normalisée des paramètres LSF pour des trames de parole de 10ms codées par le G.729. On voit bien que les corrélations sont bonnes ce qui fait que les LSF possèdent de bonnes propriétés pour la quantification et l'interpolation.

Tableau 4.1 Autocorrélations des paramèters LSF

τ	φ(τ)
0	1.0000
1	0.8027
2	0.7021
3	0.6048
4	0.5133
5	0.4299

Le codeur de l'ITU G.729 [1] possède une procédure de traitement des trames effacées basée sur une méthode de masquage prédictive (Pg. 3.4). Ce type de méthodes n'introduit aucun délai supplémentaire, car les paramètres des trames perdues seront récupérés des bonnes trames précédentes. Cependant, ce codeur quantifie les paramètres LSF par une méthode prédictive, donc la procédure de masquage utilisée peut causer une propagation de l'erreur aux autres trames comme illustré par la figure 4.1, où les distorsions spectrales d'une séquence de parole codée avec et sans effacement de trame sont représentées. Sur le graphe, on peut voire très bien la propagation de l'erreur de la distorsion après chaque perte (voire l'indicateur des trames effacées et la divergence des deux graphes).

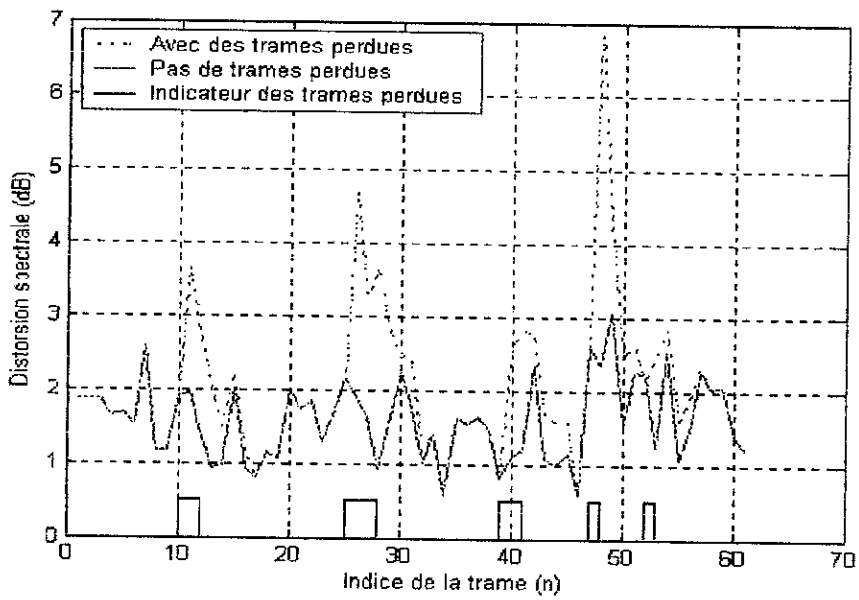

Figure 4.1 Propagation de l'erreur de la distorsion spectrale dans le G.729

Dans notre travail, nous avons, en premier temps, donné une description de l'approche interpolative, après, et pour l'implémentation de la stratégie de masquage par interpolation au G.729, nous avons quantifié les paramètres LSF avec une SVQ intra-trame à la place de la quantification prédictive utilisée par le standard, puis nous avons fait une étude comparative des distorsions spectrales causées par le masquage prédictif et interpolatif des trames effacées.

4.2 Masquage par Interpolation :

Si les futures données de la parole sont disponibles, ou peuvent être générées, alors une approche interpolative pour masquer les trames effacées devient possible. Cela devrait intuitivement produire un meilleur rendement que l'approche répétitive simple, en payant un délai supplémentaire.

L'approche interpolative, pour les codeurs CELP, a été à peine exploitée. La raison pour telle négligence relative est probablement due au délai supplémentaire imposé par cette technique, ce qui n'est pas acceptable dans quelques applications, comme le cas de l'émission sans fil où le délai est fortement contrôlé.

L'apparition d'une nouvelle, et importante application, la Voix sur des réseaux IP (*VoIP*), a rendu la méthode interpolative très attirante. Dans les systèmes VoIP, en fait, un ou plusieurs futurs trames sont, au moins la plupart du temps, disponible au décodeur, chargées dans un tampon appelé le "tampon du playout".

Un tel tampon, est introduit pour minimiser les effets d'instabilité du délai, et c'est un composant essentiel pour tous les récepteurs VoIP, donc on peut exploiter le délai introduit par ce tampon pour masquer les trames effacées avec l'approche interpolative, et donc améliorer les performances du codec sans aucun coût supplémentaire en terme de délai.

La figure 4.2 illustre l'application du masquage par interpolation dans un récepteur VoIP typique.

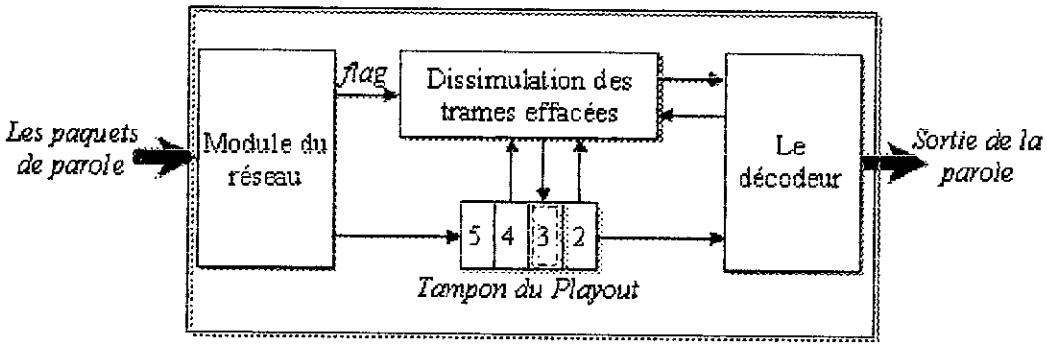

Figure 4.2 Récepteur VoIP typique : Masquage par Interpolation

Les paquets arrivant du réseau sont traités d'abord par le module du réseau. Les statistiques sont collectées, les paquets sont rangés et transféré au tampon du playout. Si, près du temps du playback, le paquet n'a pas arrivé, il est déclaré perdu et le module du masquage des trames effacées le reconstruit en utilisant les bonnes trames futures et précédentes. Sur la figure, il nous manque le paquet 3, alors on le reconstruit en interpolant le précédent (2) et le suivant (4).

4.3 Application du masquage par Interpolation au G.729 :

Les paramètres LSF sont bien connus par leur propriété d'être ordonné d'une façon que pour chaque trame, ils sont strictement en ordre ascendant avec leurs index. Ils sont connus aussi par leurs *inter-trame* et *intra-trame* corrélations. A cet effet, et pour appliquer le masquage des trames effacées par interpolation au G729, nous allons chercher une quantification *intra-trame* qui donne des performances égales ou meilleurs que la quantification prédictive (*inter-trame*), utilisée par le G729.

4.3.1 Quantification *Intra-trame* des LSF :

Nous avons choisit SVQ (*Split Vector Quantization*) comme technique de quantification des vecteurs LSF, son idée de base est de diviser un vecteur de grande dimension en des vecteurs de dimensions inférieures qui seront quantifiés séparément. Typiquement un seul quantificateur doit être conçu et utiliser pour chaque sub-vecteur. Un avantage de la SVQ est la possibilité de modifier facilement les allocations des bits entre les différents sub-quantificateurs. Cette propriété est très utile si quelque partie des vecteurs d'entrée exige une quantification plus exacte que les autres. Les questions principales pour concevoir une SVQ

concernent : ❶ La division (*Splitting*) -de combien de parties devrait le vecteur être divisé et combien de composants devrait chaque partie contenir- et ❷. L'allocation des bits.

Dans la SVQ, les vecteurs LSF sont divisés habituellement en deux ou trois parties qui seront quantifiées séparément. La Construction des dictionnaires pour la SVQ est simple car les sub-quantificateurs peuvent être traités comme étant des quantificateurs vectoriels conventionnels séparés, par exemple, comme nous l'avons fait dans notre travail, en utilisant l'algorithme de Lloyd généralisé « *GLA* » (Annexe *C*). En revanche, l'allocation des bits est la tâche la plus compliquée dans cette technique. Mis à part les longueurs des différents sub-vecteurs, l'allocation des bits est compliquée par le fait que les LSFs en fréquences intermédiaires varient plus que les LSFs en hautes et basses fréquences comme montre la figure 4.3.

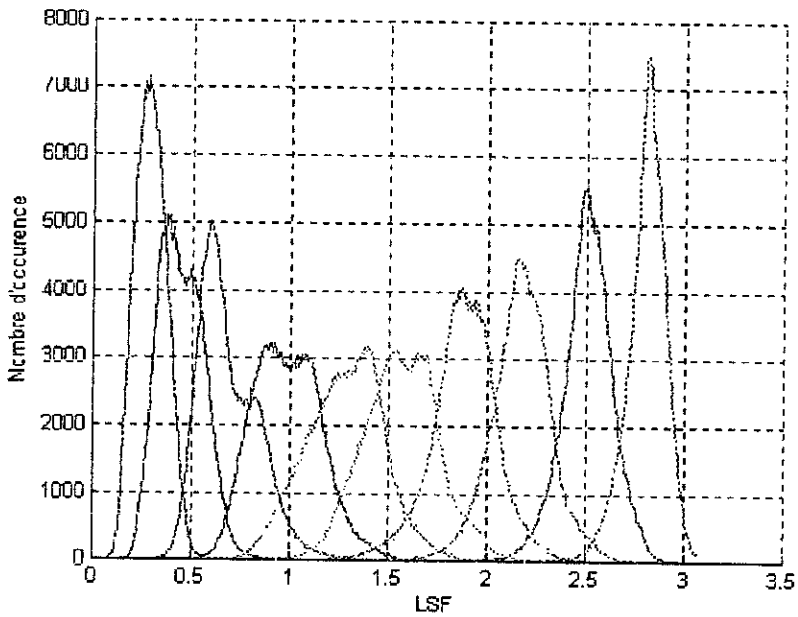

Figure 4.3 Distribution des paramètres LSF

4.3.1.1 Bases de données utilisée et Mesure des distorsions :

La matière de la parole utilisée dans les expériences consiste de deux bases de données séparées qui incluent 229829 vecteurs LSF pour l'entraînement (*Training*) et 72839 vecteurs ont été réservés pour l'évaluation ou les tests.

Les vecteurs LSF ont été produits de la base de données TIMIT [23] qui contient un total de 6300 phrases, 10 phrases parlées par chacun des 630 orateurs des 8 régions du dialecte majeures des États-Unis. La fréquence d'échantillonnage des fichiers de parole est de 8 kHz. Les phrases sont parlées par les deux sexes male et femelle.

Une analyse LPC de 10ème ordre basée sur la méthode d'autocorrélation a été appliquée pour chaque trame de 10ms en utilisant une fenêtre asymétrique de 30ms composée d'une demi-fenêtre de Hamming et un quart de période d'une fonction cosinus. Les coefficients du polynôme A(z) résultants ont été converti en paramètres LSF.

Des mesures de qualité objectives sont menées, en essayant d'estimer la qualité subjective aussi précisément que possible en modelant le système auditif humain.

Dans notre évaluation nous avons employé deux mesures objectives de qualité: la distorsion spectrale (SD) et la EMBSD (*Enhanced Modified Bark Spectral Distortion*) [17].

Pour évaluer la performance des quantificateurs, une mesure de la distorsion spectrale (SD) est effectuée. Cette dernière est l'une des mesures les plus fréquemment utilisées pour l'évaluation des performances des quantificateurs LSF. Elle est définie en dB comme :

$$SD^2 = \frac{1}{f_u - f_l} \int_{f_l}^{f_u} \left(20 \log_{10} \frac{\left| H\left(e^{j2\pi f / f_s}\right) \right|}{\left| \hat{H}\left(e^{j2\pi f / f_s}\right) \right|} \right)^2 df \qquad (4.1)$$

Où $H(z)$ et $\hat{H}(z)$ présentent respectivement, le filtre de synthèse original et le filtre de synthèse quantifié, donné par H(z) = 1/A(z), f_l et f_u définissent la fréquence limite inférieure et la fréquence limite supérieure d'intégration, et f_s est la fréquence d'échantillonnage.

La mesure objective "EMBSD" est rapportée d'avoir une corrélation très élevée avec les essais subjectifs (Tableau 4.2) et conviennent à l'évaluation de la parole dégradée par des erreurs de transmission dans des environnements réels de réseau [17], tels que des erreurs de bit et des effacements des trames.

Tableau 4.2 Table provisoire de conversion des valeurs du MOS au EMBSD

Category	*Speech quality*	*EMBSD Perceptual Distortion*
1	Unsatisfactory	8
2	Poor	6
3	Fair	4
4	Good	2
5	excellent	0

4.3.1.2 Résultats et interprétations de la quantification des LSF :

Pour trouver la partition optimale des vecteurs LSF, la corrélation *intra-trame* a été calculée pour les 229829 vecteurs, c'est-à-dire la corrélation entre LSFi et LSFj de la même trame, i, j = 1,2,...;10. Les coefficients de corrélation *intra-trame* sont présentés sur le tableau 4.3. Ces résultats montrent que la corrélation entre les LSFs consécutif est considérable. La méthode de division la plus commune est (3,3,4) dans laquelle la première tranche contient les trois premiers composants du vecteur LSF, LSF1-LSF3, la deuxième consiste en LSF4-LSF6, et LSF7-LSF10 constituent la troisième. Cependant, le quatrième LSF se corrèle plus fortement avec LSF3 qu'avec LSF5 et la corrélation entre LSF4 et LSF1 est approximativement la même que celle entre LSF4 et LSF6, alors il a été supposé que le quatrième LSF devrait être déplacé de la deuxième tranche à la première. De plus, à cause de la corrélation faible entre LSF8 et LSF9, (4,4,2) est théoriquement une bonne division.

Tableau 4.3 La corrélation entre LSF_i et LSF_j de la même trame

j \ i	1	2	3	4	5	6	7	8	9	10
1	1.000	0.721	0.427	0.472	0.069	0.015	0.094	0.104	0.095	-0.009
2	0.721	1.000	0.772	0.576	0.323	0.274	0.325	0.364	0.276	0.195
3	0.427	0.772	1.000	0.745	0.480	0.491	0.450	0.509	0.411	0.300
4	0.472	0.576	0.745	1.000	0.728	0.512	0.490	0.432	0.441	0.259
5	0.069	0.323	0.480	0.728	1.000	0.775	0.586	0.491	0.335	0.279
6	0.015	0.274	0.491	0.512	0.775	1.000	0.757	0.629	0.456	0.301
7	0.094	0.325	0.450	0.490	0.586	0.757	1.000	0.740	0.525	0.399
8	0.104	0.364	0.509	0.432	0.491	0.629	0.740	1.000	0.606	0.398
9	0.095	0.276	0.411	0.441	0.335	0.456	0.525	0.606	1.000	0.533
10	-0.009	0.195	0.300	0.259	0.279	0.301	0.399	0.398	0.533	1.000

Nous avons entraîné des quantificateurs en utilisant l'algorithme GLA, pour différents débits binaires, et comme le G.729 utilise 18 bits/trame pour la quantification des LSF, nous avons démarré avec 18 bits/trame en essayant de trouver une distorsion spectrale moyenne égale, ou meilleur que celle du standard.

La distorsion spectrale moyenne pour le G.729, calculée pour les vecteurs de la base de donnée de test est :

$$SD_{G.729} = 1.543dB \tag{4.2}$$

Pour 18 bits/trame et 19 bits/trame nous avons trouvé des distorsions spectrales >1.8dB, les résultas qui ont des *SD* proches de $SD_{G.729}$ sont mentionnés dans le tableau 4.4.

Tableau 4.4 Distorsion spectrale pour les différentes divisions et les différentes allocations de bits.

bits	Division	Allocation des bits	SD (dB)
20	3-3-4	7-6-7	1.679
20	4-4-2	9-9-2	1.645
20	4-4-2	9-8-3	1.556
20	4-6	10-10	1.503

Alors d'après ces résultats la meilleure division est celle qui nous a donné $SD = 1.503dB < SD_{G.729}$ c'est-à-dire (4,6) avec une allocation de bits 10-10, ici on doit préciser qu'on a ajouté 2 bits/trame pour réaliser la quantification *intra-trame*, ce qui va apporter 0.2 kbits/sec comme extra débit pour le G.729, mais nous allons montrer que cela peut être une solution attirante avec l'application de l'approche interpolative pour masquer les trames effacées.

4.3.2 Interpolation des paramètres LSF :

4.3.2.1 Espérance de l'erreur quadratique du masquage interpolatif :

Premièrement, on calcule l'erreur quadratique pour une récupération interpolative des LSF à partir des LSF codés par une quantification *intra-trame*. Commençant à l'instant n + 1, soit L trames consécutives sont perdues. La méthode d'interpolation récupère les vecteurs LSF perdus par interpolation linéaire entre les bonnes trames "antérieures" et "suivantes." Soit le

vecteur de dimension P, $F_n = (f_1, f_2, ..., f_P)$ le vecteur LSF de la nième trame et $\hat{F}_n$ le vecteur LSF quantifié ou interpolé correspondant; alors le vecteur LSF perdu interpolé peut être écrit :

$$\hat{F}_{n-x} = \frac{L+x-1}{L+1}\hat{F}_n + \frac{x}{L+1}\hat{F}_{n-L+1} \tag{4.3}$$

Les paramètres LSF peuvent être considérés comme stationnaire en sens large. Alors on peut approximer les vecteurs LSF quantifier par leur version non quantifiée, et prendre l'espérance de la distorsion quadratique moyenne :

$$D_L = \frac{1}{L}\sum_{x=1}^{L}\sum_{p=1}^{P}(f_{n+x,p} - \hat{f}_{n+x,p})^2 \tag{4.4}$$

On peut écrire l'espérance de la distorsion de ces L trames :

$$\begin{aligned} ED_{\text{int}} = \frac{\Phi(0)}{L}\sum_{x=1}^{L}[1 + \frac{(L+1-x)^2 + x^2}{(L+1)^2} - \frac{2(L+1-x)}{L+1}\phi(x) \\ - \frac{2x}{L+1}\phi(L+1-x) + \frac{2x(L+1-x)}{(L+1)^2}\phi(L+1)] \end{aligned} \tag{4.5}$$

Où $\Phi(.)$ et $\phi(.)$ sont, respectivement, la somme des autocorrélations, et la somme normalisée des autocorrélations des vecteurs LSF. Et sont définies comme :

$$\begin{aligned} \Phi(\tau) &= \sum_{p=1}^{P} E[f_{n,p} f_{n+\tau,p}] \\ \phi(\tau) &= \frac{\sum_{p=1}^{P} E[f_{n,p} f_{n+\tau,p}]}{\sum_{p=1}^{P} E[f_{n,p}^2]} \end{aligned} \tag{4.6}$$

4.3.2.2 Espérance de l'erreur quadratique du masquage prédictif :

Pour le masquage prédictif, les LSF perdus sont récupérés par un estimateur scalaire fixe, à partir des vecteurs, codés par une quantification *inter-trame* prédictive, reçus des "bonnes" trames antérieures, comme :

$$\hat{F}_{n+x} = B^x \hat{F}_n \tag{4.7}$$

Noter que l'erreur de masquage peut propager aux autres trames. Cette propagation peut être oubliée après plusieurs "bonnes" trames. Pour simplifier le calcule, on suppose que la propagation n'affecte qu'une seule trame. Soit e_n le vecteur résiduel reçu, le vecteur LSF résultant peut être écrit :

$$\hat{F}_{n+L+1} = \beta^{L+1}\hat{F}_n + e_{n+L+1} \tag{4.8}$$

L'erreur quadratique totale de ces $L+1$ trames sera la somme de la partie prédit et de celle propagée.

Donc l'espérance de la distorsion de la partie prédit est :

$$L \times ED_{L,pred} - \Phi(0)\sum_{x=1}^{L}[1+\beta^{2x} - 2\beta^{x}\phi(x)] \tag{4.9}$$

Pour la partie propagée :

$$D_{prop} = \sum_{p=1}^{P}(f_{n+L+1,p} - \beta^{L+1}f_{n,p} - e_{n+L+1,p})^2 \tag{4.10}$$

On prend l'espérance sur les deux côtés. Tout les termes avec e_n+L+1 égal à zéro puisque e_n+L+1 est indépendant de f_n et l'espérance de e_n+L+1 égale à zéro. En négligeant le petit terme, on obtient :

$$ED_{prop} - \Phi(0)[1+\beta^{2(L+1)} - 2\beta^{L+1}\phi(L+1)] \tag{4.11}$$

Donc l'espérance de la distorsion moyenne des $L+1$ trames est :

$$\begin{aligned} ED_{Pred} &= \frac{1}{L+1}(L \times ED_{L,pred} + ED_{prop}) \\ &\quad - \frac{\Phi(0)}{L+1}\sum_{x=1}^{L-1}[1+\beta^{2x} - 2\beta^{x}\phi(x)] \end{aligned} \tag{4.12}$$

La méthode répétitive utilisée par le G729 est un cas spécial de la méthode prédictive où l'estimateur $\beta = 1$ donc l'espérance de la distorsion moyenne devient :

$$ED_{rep} = \frac{\Phi(0)}{L+1}\sum_{x=1}^{L+1}[2 - 2\phi(x)] \tag{4.13}$$

4.3.2.3 Comparaison des deux méthodes :

Nous avons calculé la somme des autocorrélations des 229829 vecteurs LSF (voire Tableau 4.5)

Tableau 4.5 Sommes des autocorrélations normalisées des paramètres LSF

τ	φ(τ)
0	1.0000
1	0.8027
2	0.7021
3	0.6048
4	0.5133
5	0.4299

Le rapport $\frac{ED_{int}}{ED_{rep}}$ pour L = 1, 2, 3 est calculé et représenté dans le tableau 4.6.

Tableau 4.6 Rapport des distorsions moyennes de la récupération prédictive et interpolative des LSF

L	ED_{rep}	ED_{int}	ED_{rep}/ED_{int}
1	0.4952	0.2457	2.0155
2	0.5936	0.2860	2.0755
3	0.6886	0.3248	2.1201

On remarque que l'espérance de la distorsion moyenne de la méthode répétitive, est plus grande que celle de la méthode interpolative d'un facteur de 2. On peut approximer la relation entre la distorsion spectrale (*DS*), et l'erreur quadratique *SE* par une fonction linéaire (dans

l'échelle log-log, c'est une relation statistique calculée pour un très grand nombre de vecteurs LSF [9]) c'est-à-dire : $\log(SD) = r\log(sqe) + b$ où r est positive.

Alors :

$$\frac{DS_{rep}}{DS_{int}} - (\frac{ED_{rep}}{ED_{int}})^{r} > 1 \qquad (4.14)$$

D'après ces résultats qui ont l'air théorique plus qu'expérimental, on peut voire réellement l'efficacité de la méthode interpolative appliquée au G.729.

4.3.3 Simulation et Résultats :

Nous avons trouvé qu'avec une quantification SVQ *intra-trame* 20 bits/trame, on réalise une distorsion minimale pour le codeur G.729, cette quantification qui est 2 bits/trame plus que le codage prédictive original du G.729, mais qui présente des propriétés adéquates à l'application de la récupération interpolative des LSF. Nous allons simuler la voix en temps-réel sur des paquets en réseaux où chaque paquet contient une trame.

4.3.3.1 Modèle de réseau :

Nous avons employé un modèle simple de réseau, dit modèle de Markov, pour abandonner des paquets de voix, qui est acceptable pour modeler le processus point-à-point de perte des paquets sur Internet ([13], [14]). Ce modèle a deux états refléter, si le paquet précédent est reçu (état 0) ou perdu (état 1). Soit p la probabilité pour que le modèle de réseau abandonne un paquet sachant que le paquet précédent est livré, c.-à-d. la probabilité pour que le modèle aille de l'état *0* à l'état *1*. Soit q dénote la probabilité pour que le modèle de réseau abandonne un paquet sachant que le paquet précédent est abandonné, c.-à-d. la probabilité pour que le modèle reste dans l'état *1*. Cette probabilité est également connue comme *la probabilité conditionnelle de perte (CLP)*. Soit P_0 et P_1 dénotent la probabilité pour être dans l'état *0* et l'état *1* respectivement nous avons :

$$\begin{aligned} P_1 &= P_0 \cdot p + P_1 \cdot q \\ P_1 &+ P_0 = 1 \end{aligned} \qquad (4.15)$$

$$\Rightarrow P_0 = \frac{1-q}{p+1-q} \qquad P_1 = \frac{p}{p+1-q} \qquad (4.16)$$

La probabilité pour qu'un paquet soit abandonné sans connaître si le paquet précédent est livré ou abandonné, c.-à-d. *La probabilité de perte sans conditions* (*ULP*) est exactement la probabilité pour que le modèle de réseau soit dans l'état 1 (P1). La figure 4.4 présente le modèle de Markov avec ses probabilités de transition, et le tableau 4.7 cite les taux de perte utilisés dans notre simulation.

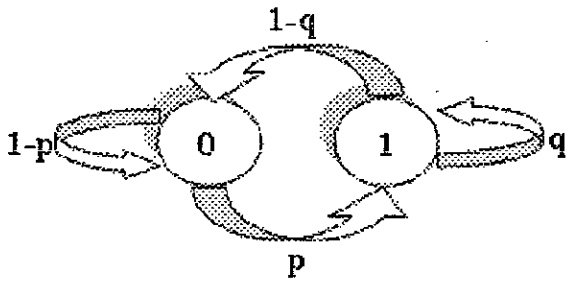

Figure 4.4 Perte des paquets modélisée par un processus aléatoire de Markov

Tableau 4.7 Les taux de Pertes simulés

Taux(%)	p	q
00	0.0	0.00
10	0.1	0.15
20	0.2	0.30
30	0.3	0.35
40	0.3	0.50

4.3.3.2 Procédure de masquage implémentée :

Le processus complet de masquage peut être résumé ici.

Si une trame est déclarée perdue :

1. Interpolation linéaire des paramètres LSF de la bonne trame "précédente" et la bonne trame "suivante";
2. Interpolation du délai tonale ;
3. En se basant sur la bonne trame précédente, Prendre une décision sur le type de la trame (voisée ou non voisée V/UV);
4. Si la trame précédente est voisée : -Mettre la contribution du dictionnaire fixe à zéro;

5. Si la trame antérieur est non voisée: -Mettre l'information du dictionnaire adaptatif à zéro, -Utiliser l'information précédente du gain, -Remplacer les signaux d'excitation par une séquence de nombres aléatoires normalisée par le gain atténué.

4.3.3.3 Résultats :

La figure 4.5 montre les performances de la méthode interpolative appliquée aux paramètres LSF quantifiés avec SVQ *intra-trame*, comparées à la méthode prédictive adoptée par le G.729.

Les pourcentages des distorsions spectrales *SD* comprises entre 2-4dB et celle >4dB (*Outliers*) sont des paramètres importants qui affectent la qualité perceptuelle de la parole décodée et par conséquent sont présentés dans le tableau 4.8.

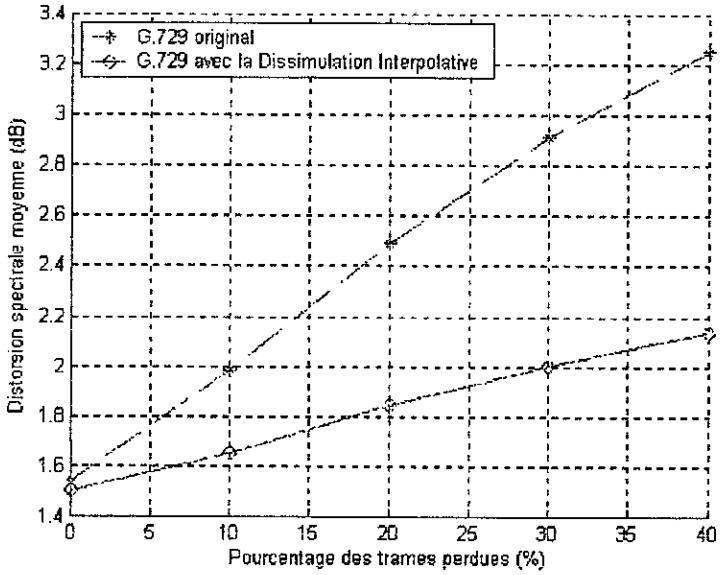

Figure 4.5 Distorsion spectrale moyenne avec des trames effacées

On voit bien que, avec un débit de 0.2 kbps de plus, c'est-à-dire 2,5% du débit total, notre méthode de quantification et de récupération des LSF ainsi appliquée, réalise 0.3 à 1.1 dB de distorsion spectrale de moins, comparée à la méthode adoptée par le standard G.729. Les pourcentages des *Outliers* sont aussi beaucoup plus petits, ce qui permet d'avoir une qualité perceptuelle considérable quand des trames effacées se produit.

Tableau 4.8 Distorsion spectrale moyenne et les *Outliers* avec des trames effacées

Trames perdues (%)	G729			SVQ *Intra-trame*		
	SD_{moy} (dB)	*Outliers* (%)		SD_{moy} (dB)	*Outliers* (%)	
		2-4 dB	> 4dB		2-4 dB	> 4dB
0	1.543	19.60	0.62	1.503	13.89	0.02
10	1.989	32.00	5.46	1.655	18.88	1.73
20	2.490	40.82	12.69	1.845	24.04	4.26
30	2.913	46.04	19.58	2.003	28.37	6.29
40	3.249	56.15	23.72	2.141	32.55	8.02

La distribution, des distorsions spectrales, représenté sur la figure 4.6, montre que la plupart des trames perdues sont interpolées avec des petites distorsions, avec la quantification *intra-trame* SVQ.

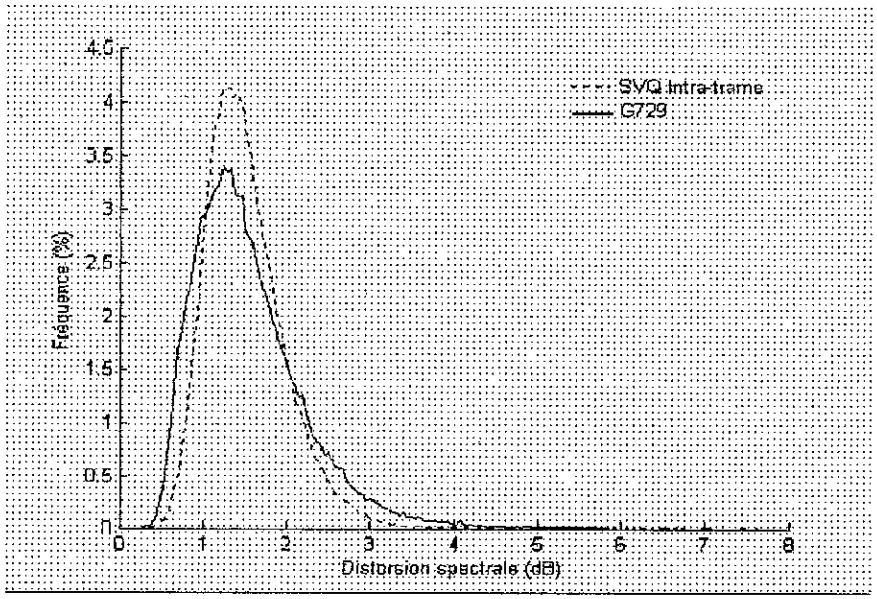

Figure 4.6 Histogrammes des Distorsions Spectrales (SD)

Nous avons effectué aussi des mesures comparatives avec la distorsion spectrale modifiée « EMBSD » (qui peut nous donner une idée plus précise sur la qualité de la parole), la figure 4.7 montre, pour une nouvelle fois, les performances du masquage par interpolation appliqué aux paramètres LSF quantifiés avec une SVQ *intra-trame*, comparées à la méthode prédictive adoptée par le G.729.

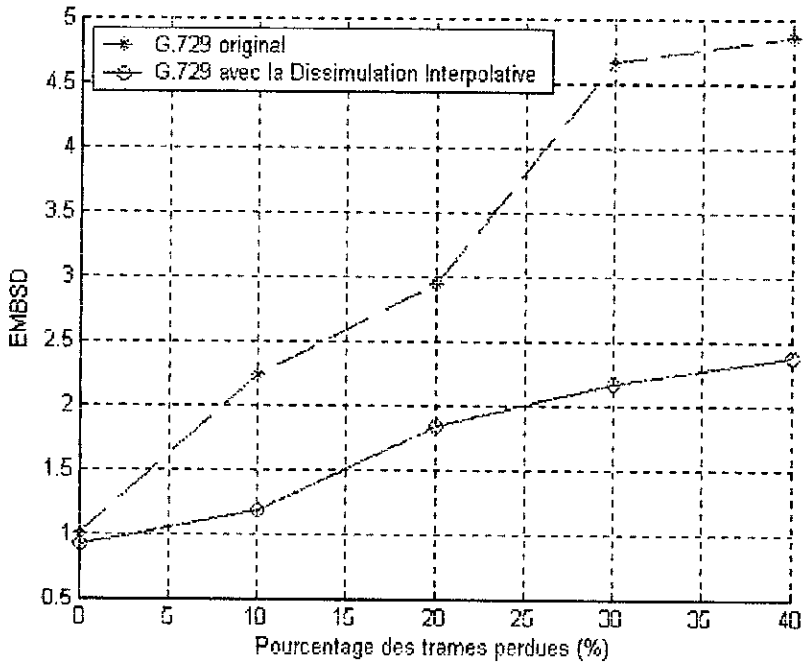

Figure 4.7 EMBSD avec des trames effacées

On voit bien, pour une deuxième fois, que avec un débit de 0.2 kbps de plus, notre méthode de quantification et de récupération des LSF ainsi appliquée, réalise jusqu'à 2.5 de moins de la distorsion perceptuelle EMBSD, comparée à la méthode adoptée par le standard G.729. Et si on compare les résultats de la figure 4.7 avec ceux du tableau 4.2 on peut conclure que notre méthode donne une qualité « *good* » jusqu'au 30% de taux de perte ce qui correspond à une qualité « *Fair* » pour le G.729 original.

Des testes d'écoute informels (simples) montrent que l'application de la quantification *intra-trame*, et de la récupération des LSF par interpolation (*Interpolative Concealment*), améliore considérablement la qualité des trames de parole effacées.

Le délai total d'interpolation est la multiplication des délais des trames effacées. Si, par exemple, on a trois trames effacées, alors le délai sera 3 * 10ms+5ms + RTT/2 où RTT (*Round Trip Time*) est le temps moyen de l'aller-retour des paquets sur le réseau, généralement compris entre 10 et 700 ms pour un réseau typique. Le retard maximal acceptable pour les applications VoIP (*Voice over IP networks*) est moins de 800ms. Par conséquent, le délai causé par l'interpolation peut être insignifiant comparé à l'amélioration apportée à la qualité de la parole.

4.4 Conclusion :

En ajoutant la quantification SVQ intra-trame des LSF, et en permettant aux trames effacées d'être interpolées des bonnes trames « précédente » et « suivante », nous avons présenté une méthode efficace pour récupérer les trames de parole codées par le codage CELP.

Des tests d'écoute informels (tests subjectifs simples) montrent que l'application de la quantification *intra-trame*, et de la récupération des LSF par interpolation, améliore considérablement la qualité des trames de parole effacées.

L'inconvénient de cette méthode peut être le délais supplémentaire requis pour l'interpolation, mais comme nous l'avons vu, nous pouvons exploiter le délai introduit par le « *tampon du playout* » (qui est un composant essentiel dans les applications VoIP), pour implémenter cette méthode.

L'extra débit (2,5% du débit total) de la quantification SVQ intra-trame des paramètres LSF peut être insignifiants, comparé aux performances obtenues, et à la qualité du signal vocal reconstitué après le masquage des trames effacées.

Conclusion Générale

Dans ce travail, nous avons abordé le problème de perte des trames lors de la transmission de la parole sur les réseaux *IP*, et nous avons mis l'accent sur l'implémentation d'une méthode interpolative afin de masquer ces pertes, cette méthode consiste à interpoler à partir des bonnes trames antérieures et futures les paramètres *LSF*, qui représentent les coefficients du filtre de prédiction linéaire.

Nous avons appliqué cette méthode d'interpolation au standard de l'ITU, le **G.729**, un des codecs les plus utilisé dans ce domaine.

Ce codec quantifie les paramètres LSF avec une quantification prédictive (quantification inter-trame), nous avons montré que cette quantification ne convient pas à l'application de la méthode de masquage proposée, ce qui nous a amené à changer complètement la méthode de quantification adoptée par ce standard.

Des tests d'écoute informels (tests subjectifs simples) montrent que l'application de la quantification *intra-trame*, et de la récupération des LSF par interpolation, améliore considérablement la qualité des trames de parole effacées.

L'inconvénient de cette approche interpolative peut être le délais supplémentaire requis pour l'interpolation, mais comme nous l'avons vu, on peut exploiter le délai introduit par le « *tampon du playout* » (qui est un composant essentiel dans toutes les applications VoIP), pour implémenter cette méthode.

L'extra débit (2,5% du débit total) de la quantification SVQ intra-trame des paramètres LSF peut être insignifiants, comparé aux performances obtenues, et à la qualité du signal vocal reconstitué après le masquage des trames effacées.

Etant limité par le temps consacré au PFE, nous n'avons pas pu implémenter qu'une seule méthode de quantification des paramètres LSF et une seule méthode de masquage, et les comparées aux celles du standard G.729. Le travail reste ouvert à d'autres techniques surtout de la quantification intra-trame des LSF à fin de réduire l'extra débit et la complexité.

Annexe A

Algorithme de Levinson-Durbin

Les coefficients d'autocorrélation $R(k)$, $k = 0,1,\ldots\ldots\ldots, p$ sont utilisées pour obtenir les coefficients du filtre LP après résolution du système linéaire (1.13)

Il s'agit donc d'inverser une matrice d'ordre "p". Les méthodes algébriques classiques exigent pour cela un nombre d'opérations (multiplication+ addition) de l'ordre de p^3, ce que l'on note $O(p^3)$

L'algorithme qui va être décrit profite de la structure particulière (Toeplitz symétrique) de la matrice d'autocorrélation pour résoudre (1.13) par une récursion sur l'ordre de prédiction: autrement dit, ils fournissent toutes les solutions d'ordre

M=1,2,...p, le nombre d'opérations est seulement $O(p^2)$.

La variance de l'erreur de prédiction αp sera obtenue également par une récurrence sur l'ordre m.

Rappelons que la fonction d'autocorrélation est supposée connue et que pour un signal stationnaire, on a :

$$R(i,j) = R(|i-j|) = R(k) \tag{A.1}$$

Initialisation:

$$a_m(0) = 1, \qquad (m = 1,2,\ldots\ldots, p) \qquad E_0 = R(0) = \sigma_x^{\,2} \tag{A.2}$$

Récursion :

Pour : $m = 1,2,\ldots p.$

$$k_m = -\frac{1}{E_{m-1}}\left[R(m) - \sum_{k=1}^{m-1} \alpha_{m-1}(k) R(m-k)\right] \tag{A.3}$$

Pour $k = 1,2,...m-1$.

$$\alpha_k(m) = \alpha_k(m-1) - k_m \alpha_{m-k}(m-1) \tag{A.4}$$

$$E_m = E_{m-1}(1 - k_m{}^2) \tag{A.5}$$

Les coefficients $a_k(m)$ résultant, quand m = p représentent les coefficients de prédiction d'un prediteur linéaire d'ordre p :

La valeur de k_m joint à la propriété : $-1 \le k_m \le 1$

Cette relation est une condition nécessaire et suffisante pour que le filtre soit stable.

La méthode d'autocorrélation garantit la stabilité du filtre, de plus le calcul de $R(i)$ nécessite un fenêtrage de $S(n)$ par un la fenêtre de Hamming.

Annexe B

Conversion des coefficients de prédiction linéaire (LP) en coefficients de paires de raies spectrales (LSP)

Les coefficients du filtre LP, a_i, $i = 0,.....,10$, sont convertis en coefficients de paires de raies spectrales (LSP) aux fins de la quantification et de l'interpolation. Pour un filtre LP du 10e ordre, les coefficients LSP sont définis comme étant les racines des polynômes sommateurs et différentiateurs suivants:

$$F'_1(z) = A(z) + z^{-11}A(z^{-1}) \tag{B.1}$$

$$F_2'(z) = A(z) - z^{-11}A(z^{-1}) \tag{B.2}$$

Respectivement. Le polynôme $F'_1(z)$ est symétrique, tandis que le polynôme $F'_2(z)$ est antisymétrique (conjugué de $F'_1(z)$). On peut démontrer que toutes les racines de ces polynômes se trouvent sur le cercle unité et qu'elles alternent l'une avec l'autre.

Le polynôme $F'_1(z)$ possède une racine $z = -1$ ($\omega = \pi$) et le polynôme $F'_2(z)$ possède une racine $z = +1$ ($\omega = 0$). On élimine ces deux racines en définissant les deux nouveaux polynômes suivants:

$$F_1(z) = F'_1(z)/(1 + z^{-1}) \tag{B.3}$$

Et

$$F_2(z) = F'_2(z)/(1 + z^{-1}) \tag{B.4}$$

Chacun de ces polynômes possède 5 racines conjuguées sur le cercle unité ($e^{\pm j\omega_i}$) et on peut les écrire comme suit:

$$F_1(z) = \prod_{i=1,3,.....,9} (1 - 2q_i z^{-1} + z^{-2}) \tag{B.5}$$

Et

$$F_2(z) = \prod_{i=2,4,\ldots,10} (1 - 2q_i z^{-1} + z^{-2}) \tag{B.6}$$

Où $q_i = \cos(\omega i)$. Les coefficients ωi sont les fréquences des raies spectrales (LSF) et satisfont à la relation d'ordre $0 < \omega i < \omega 2 < \ldots < \omega 10 < \pi$. Les coefficients q_i sont désignés comme étant les coefficients LSP dans le domaine cosinusoïdal. Etant donné que les deux polynômes $F_1(z)$ et $F_2(z)$ sont symétriques, il suffit de calculer les 5 premiers coefficients de chaque polynôme, au moyen des relations de récurrence suivantes:

$$\begin{aligned} f_1(i+1) &= a_{i+1} + a_{10-i} - f_1(i) \qquad \text{i} = 0,\ldots,4 \\ f_2(i+1) &= a_{i+1} + a_{10-i} - f_2(i) \qquad \text{i} = 0,\ldots,4 \end{aligned} \tag{B.7}$$

Où $f_1(0) = f_2(0) = 1{,}0$. Les coefficients LSP sont trouvés par évaluation des polynômes $F_1(z)$ et $F_2(z)$ à 60 points équidistants entre 0 et π puis en vérifiant les changements de signe. Chaque changement de signe correspond à l'existence d'une racine et l'intervalle du changement de signe est alors divisé par 4 afin d'affiner la recherche de la racine. On se sert des polynômes de Tchebycheff pour évaluer les polynômes $F_1(z)$ et $F_2(z)$. Dans cette méthode, on trouve directement les racines dans le domaine cosinusoïdal. Le polynôme $F_1(z)$ ou $F_2(z)$, évalué à la valeur $z = \exp(j\omega)$, peut s'écrire comme suit:

$$F(w) = 2e^{-j5w} C(x) \tag{B.8}$$

Avec

$$C(x) = T_5(x) + f(1)T_4(x) + f(2)T_3(x) + f(3)T_2(x) + f(4)T(x) + f(5)/2 \tag{B.9}$$

Où le terme $T_m(x) = \cos(m\omega)$ est le polynôme de Tchebycheff du nième ordre et où les $f(i)$, $i = 1,\ldots,5$, sont les coefficients du polynôme $F_1(z)$ ou $F_2(z)$, calculés par l'équation (B.7) de récurrence suivante:

Pour k = 4 à 1

$$b_k = 2xb_{k+1} - b_{k+2} + f(5-k)$$

Fin

$$C(x) = xb_1 - b_2 + f(5)/2$$

Avec les valeurs initiales : $b_5 = 1$ et $b_6 = 0$

Annexe C

Algorithme de Lloyd Généralisé (GLA)

Les conditions d'optimalité citées précédemment conduisent à la conception d'un algorithme qui réalise, à partir d'une séquence d'apprentissage représentative de la statistique de la source à coder, la construction d'un dictionnaire (localement) optimal. Cet algorithme de classification, appelé aussi algorithme des K-moyens (K-means) est l'extension au cas vectoriel de l'algorithme de Lloyd-Max (cas scalaire).

Il s'agit d'un algorithme d'optimisation itérative opérant à partir d'un dictionnaire initial. A chaque itération (dite « itération de Lloyd »), deux opérations distinctes sont appliquées :

- Une classification suivant la règle du plus proche voisin ;
- Une optimisation suivant la condition du centroïde.

Chaque itération de Lloyd, en modifiant localement le dictionnaire, réduit ou laisse inchangée la distorsion moyenne. L'algorithme converge en un nombre fini d'itérations vers un minimum local. Ce minimum local varie en fonction du choix du dictionnaire initial. Le choix de ce dernier est donc capital.

étape 1 : on commence par un dictionnaire initial C_i, mettre m=1,

étape 2a : ayant un dictionnaire $C_m=\{y_i\}$, partitionner la séquence d'entraînement en un ensemble de classe (cluster) S^i en utilisant la condition du plus proche voisin, où

$$S^i = \left\{ x \in T \middle| d(x, y_i) \le d(x, y_j);\ \text{pour tout } j \neq I) \right\}$$

étape 2b :en utilisant la condition de centroïde, calculer les centroïdes pour l'ensemble des classes trouvés en étape 1 pour obtenir le niveau dictionnaire :

$$C_{m+1} = \left\{ Cent(S_i) \middle| i = 1,...N \right\}$$

étape 3 : calculer la distorsion moyenne pour C_{m+1}, si elle a change d'une petite quantité par rapport à l'itération précédente stop, sinon, mettre m= m+1 et répéter étape 2 et 3.

La figure C.1 illustre la procédure de fonctionnement de l'algorithme de Lloyd.

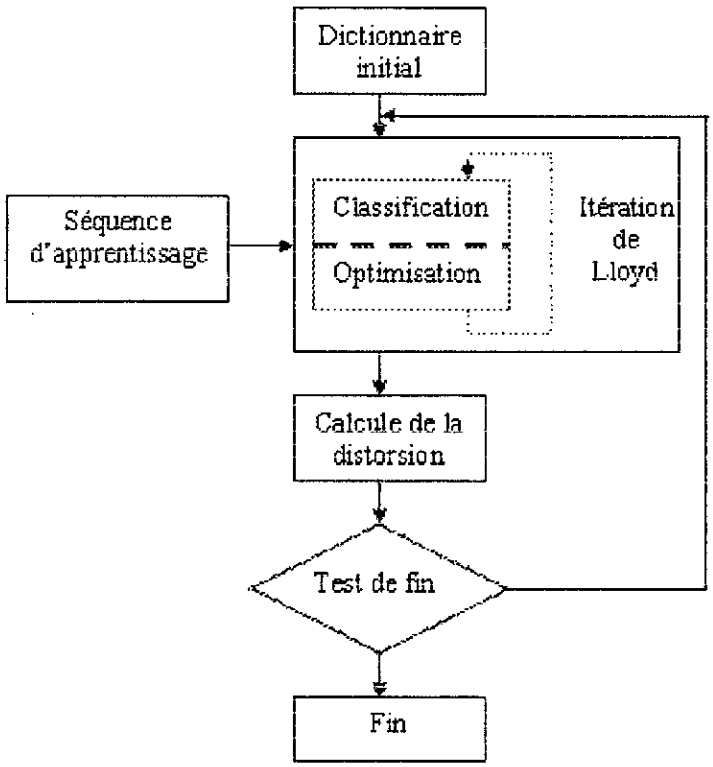

Figure C.1 Shéma de fonctionnement de l'algorithme de Lloyd

Bibliographie

[1] ITU," *ITU-T G.729: CS-ACELP Speech coding at 8 Kbits/s*", ITU 1998.

[2] ITU,*"ITU-T G.723.1: Dual rate speech coder for Multimedia Communication Transmitting at 5.3 and 6.3 Kbits/s"*, ITU 1996.

[3] T.Dutoit," *Introduction au Traitement Automatique de la Parole*", Faculté Polytechnique de Mons 1989.

[4] R.Boite et M. kunt, " *Traitement de la Parole*", presses polytechniques Romandes.1987.

[5] M.Djeddou," *Conception et réalisation d'un codeur / Décodeur de la parole à bande étroite (300-3400 Hz).à16 Kbits/s et àfaible retard (< 5 ms)*", Thèse magistère ENP.1997.

[6] G. Fant, *"Acoustic Theory of Speech Production,"* Mounton and Co., Gravenhage, The Netherlands, 1960.

[7] E. Mahfuz,*"Packet Loss Concealment for Voice Transmission over IP Networks"*, Thesis Master of Engineering. Department of Electrical & Computer Engineering McGill University. Montreal, Canada. September 2001.

[8] JUAN Carlos DE Marten, TAKAHIRO Unno and VISHU Viswanathan, " *Improved frame erasure concealment for CELP- based coders*, " DSPS R&Dm Texas Instrtument, DALASm TEXAS.

[9] J.Wang and J.D.Gibson, *"Parameter interpolation to Enhance the Frame Erasure Robustness of CELP Coders in Packet Networks"*, Department of Electrical Engineering Southern Methodist University 2001.

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

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

[12] J.Wang and J.D.Gibson, *"Performance Comparison of Intraframe and Interframe LSF Quantization in Packet Networks"*, Proc.2000 IEEE Workshop on Speech Coding, Delavan , WI, USA, septembre 2000.

[13] J. C. Bolot, S. Fosse-Parisis, and D. Towsley. Adaptive FEC-Based Error Control for Interactive Audio in the Internet. Proceedings IEEE Infocom 1999, New York, NY, March 1999.

[14] H.Sanneck[1], N.Tuong Long Le[2], M.Haardt[1], and W.Mohr [1] ,*"Selective Packet Prioritization forWireless Voice over IP"*,[1]Siemens AG, Information and Communication Mobile, Networks, D-81359 Munich, Germany.[2]Department of Computer Science, University of North Carolina, Chapel Hill, NC 27599- 175, USA.

[15] J.Rosenberg,*"G.729 Error Recovery for Internet Telephony"*, Project Report, Columbia University, Mai 1997.

[16] M.Oueld-Cheikh,*"Conception et Réalisation d'un Codeur/Décodeur de la Parole à Large Bande (50-7000 Hz), et à Faible Débit (13 Khits/s) "*, Thèse Magistère, ENP1999.

[17] W.Yang,*"Enhanced Modified Bark Spectral Distortion (EMBSD): An Objective Speech Quality Measurment Based on Audible Distortion and Cognition Model"*, Ph.D Dissertation, May 1999, Temple University, USA.

[18] Paxton J. Smith,*" Voice Conferencing over IP Networks", Thesis* Master of Engineering. Department of Electrical & Computer Engineering McGill University Montreal, Canada 2002.

[19] A.Kaddai, *" Utilisation du treillis pour le codage de la parole"*, Thèse PFE, ENP1999

[20] A.Vega Garcia,*"Mécanismes de contrôle pour la transmission de l'audio sur l'Internet"*, Thèse doctorat, L'universitéé de Nice-Sophia Antipolis, Ecole doctorale SPI 1996.

[21] H.Sanneck and N.T.L.Le, *"Speech Property-Based FEC for Internet Telephony Applications"*, SPIE/ACM SIGMM Multimedia computing and Networking conference 2000(MMCN 200), San Jose, CA, Janury 2000.

[22] J.pan and T.R.fischer,*" Vector Quantization of Speech Line Spectrum Pair Parameters and Reflection Coefficients"*, IEEE transaction on speech and audio processing, vol.6, No.2, P106, 1998.

[23] NIST,*"Timit Speech Corpus"*, NIST 1990.