N° d'ordre : 11/2005-M/EL

République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université des sciences et de la technologie Houari BOUMEDIENE

جامعة هواري بومدين
للعلوم والتكنولوجيا
U S T H B

Faculté d'Electronique et d'Informatique

MEMOIRE

Présenté pour l'obtention du diplôme de MAGISTAIRE

EN : ELECTRONIQUE

Spécialité : Traitement du Signal et des Images

Par : CHAOUCH Hocine

Sujet

Amélioration des performances du codec G.729

Soutenu le : 30/06/2005, devant le jury composé de :

Mr Youcef SMARA,	Professeur	USTHB	Président
Mr Daoud BERKANI,	Professeur	ENP	Directeur de thèse
Mme Fatima BOUMGHAR,	Professeur	USTHB	Examinateur
Mme Amina SERIR,	Maître de conférence	USTHB	Examinateur
Mlle Fatiha MERAZKA,	Docteur	USTHB	Examinateur

Remerciements

Je voudrais exprimer ma reconnaissance et mes remerciements les plus sincères à toutes les personnes qui ont eu l'amabilité de me prodiguer aide, critiques, suggestions et encouragements dans l'accomplissement du présent travail.

En particulier, je voudrais exprimer ma gratitude à :

- Monsieur le professeur D. BERKANI de l'ENP pour sa patience, ses encouragements et ses suggestions.
- Melle F. MERAZKA, Docteur à l'USTHB pour son aide, conseils et avoir mis à ma disposition toute la documentation nécessaire.
- Tous les enseignants qui ont accepté de faire partie de mon jury : Mme Fatima BOUMGHAR, Mme Amina SERIR et Mr Youcef SMARA.

ملخص

لوحظ في الرّامزة النّموذجية ج.729 (G.729) المقدمة عند اتحاد التّولي للموا صلات البعيدة (I.T.U) أنّ عند ضياع فلح من الكلام، الأخطاء المتولدة لا تتحصر في الفلح المفقودة فقط بل تنتشر إلى الفلح الموالية

و بما أنّ الأصل في هذا الانتشار هو التكميم بين الفلح "التكميم بالمتنبأ" (quantification Inter-trame) لمعاملات أزواج الخطوط الطيفية (LSP)، فبدلنا هذا التكميم بتكميم داخل الفلح (quantification Intra-trame).

نهايته، النتائج المحصل عليها تظهر، في حالة ضياع فلح من الكلام، تحسن واضح في فعالية الرامزة ج.729

مفاتيح الكلمات

ترميز الكلام، الكلام عبر شبكة انترنت، تكميم بين الفلح، تكميم داخل الفلح، إخفاء فقدان الفلح.

Résumé

Dans le codec standard de l'ITU, le G.729, nous avons observé que, lors de l'occurrence de pertes de trames, les erreurs engendrées ne se limitaient pas aux trames perdues mais se propageaient aux trames suivantes. Comme l'origine de cette propagation est la quantification Inter-trames (quantification prédictive) des coefficients de paires de raies spectrales LSP, nous l'avons remplacé par une quantification Intra-trame (Intra-DQ).

Les résultats obtenus montrent, dans des conditions de pertes de trames, une nette amélioration des performances du codec G.729.

Mots clès

Codage de la parole, Voix sur IP, Quantification Inter-trame, quantification Intra-trame, Masquage de perte.

Abstract

In the codec G.729, error propagation can be observed with the occurrence of loss frames. As a matter of fact, the errors propagate to the next frames following those which are lost and don't stop within the lost frames. Since that propagation originates from the Inter-frame quantization (predictive quantization) of Line Spectral Pairs (LSP), we have replaced this method by an Intra-frame quantization(Intra-DQ).

The results have shown a clear improvement of the performances of the G.729 in lossy conditions.

Key words

Speech coding, Voice Over Internet protocol, Inter-frame quantization, Intra-frame quantization, packet loss concealment.

Sommaire

Lexique

ADM	Adaptative Delta Modulation Modulation Delta adaptative
ADPCM	Adaptative Differential Pulse Code Modulation Modulation différentielle adaptative par impulsions codées
CELP	Code-Excited Linear-Prediction Prédiction linéaire avec excitation par code
CS-ACELP	Conjugate-Structure Algebraic Code-Excited Linear-Prediction Prédiction linéaire avec excitation par séquence codées à structure algébrique conjuguée
EMBSD	Enhanced Modified Bark Spectral Distortion Distortion spectrale modifiée
FEC	Forward Error Correction Correction d'erreurs en avance
Intra-DQ	Intraframe Differential Quantization Quantification différentielle intra-trame
IP	Internet Protocol Protocole d'Internet
ITU	International Telecommunication Union Union internationale de télécommunication
LAR	Log Area Ratio Ecart d'aire logarithmique
LP	Linear Prediction Prédiction linéaire
LSF	Line Spectral Frequency Fréquence de raie spectrale
LSP	Line Spectral Pair Paire de raie spectrale
MOS	Mean Opinion Score
PCM	Pulse Code Modulation Modulation par impulsions codées
PSTN	Public Switched Telephone Network Réseau téléphonique publique
PSVQ	Predictive Split Vector Quantizer Quantification vectorielle par Split prédictive
RTP	Real Time Protocol Protocole en temps réel
SD	Spectral Distortion Distorsion spectrale
SIP	Session Initiation Protocol Protocole d'initiation de session
SVQ	Split Vector Quantization Quantification vectorielle par Split
TCP/IP	Tranmission Control Protocol / Internet Protocol
UDP	User Datagramme Protocol
VQ	Vector Quantization Quantification vectorielle
VoIP	Voice over Internet Protocol Voix sur IP

Introduction

Ces dernières années, les chercheurs et les simples utilisateurs, ont découvert l'intérêt considérable de la transmission interactive de la parole par Internet (VoIP: Voice over Internet Protocol). 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 difficile) à utiliser sur les réseaux de téléphonie traditionnels.

En plus, des codages et des décodages très complexes de la parole peuvent être menés avec un matériel économique, et peuvent être disponibles chez tous les utilisateurs. Comme par exemple, les deux codecs appelés "frame-based" le G.723.1[30] et le G.729[18], qui sont très convenable pour la téléphonie par Internet, car ils fournissent une qualité téléphonique de la parole avec des faibles débits binaires (5.3 / 6.3 kBit/s et 8 kBit/s respectivement) comparés au PCM (Pulse Coded Modulation) 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 l'Internet, sont basés sur le principe dit "best effort", qui ne garantit pas le taux minimal de perte de paquets exigé par la Téléphonie par Internet, 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 outre, étant donné le codage prédictif adopté par le G.723.1 et le G.729, 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 re-synchronisé avec le codeur.

L'objectif de notre travail est d'implémenter une nouvelle technique de quantification, et une méthode de dissimulation des trames perdues dans le réseau, pour alléger le problème décrit dans le paragraphe précédent et améliorer les performances de ce codec.

Ce mémoire est constitué de quatre chapitres :

Le **premier** chapitre comporte des généralités sur le modèle de production de la parole humaine et les différentes techniques de codage de la parole.

Le **deuxième** chapitre est consacré à la transmission de la voix sur les réseaux IP(VoIP), les codecs utilisés et les différentes méthodes de recouvrement de pertes.

Le **troisième** chapitre décrit la norme du codec G.729, principes de codeur et de décodeur.

Le **quatrième** chapitre regroupe toutes les études, simulations et interprétations des résultats obtenus.

Chapitre 1

Codage de la parole

1.1- Le signal parole

Le signal de parole est généré par l'appareil phonatoire. C'est un organe d'une grande complexité mécanique. Il se compose de deux parties anatomiquement distinctes. Le poumon et le larynx, partie supérieure de la trachée artère, constituent l'essentiel du générateur sonore. Le larynx est un ensemble de muscles et de cartilages mobiles qui entourent une cavité située à la partie supérieure de la trachée.

1.2- Processus 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).

Le larynx a une fonction qui lui est propre: c'est la production des sons, ou "phonation". 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, déterminer une ouverture triangulaire appelée glotte. Pendant la respiration, l'air y passe librement et aussi pendant la phonation des sons sourds ou non voisés. Les sons voisés résultent au contraire d'une vibration périodique des cordes vocales; des impulsions périodiques de pression sont ainsi appliquées au conduit vocal qui s'étend du pharynx jusqu'aux lèvres.

La voix résulte du fonctionnement simultané des poumons, du larynx et de la cavité de la bouche et du nez qui modifie sa forme et ses dimensions suivant le son émis et qui, avec la poitrine, jouent le rôle de caisse de résonance. Toutes les voix se ressembleraient si la voix était seulement laryngée. Or, ce sont les modifications de forme et de dimensions que subissent la bouche, le pharynx pendant l'émission de la voix, qui donnent au contraire à celle-ci un timbre qui est particulier à chacun d'entre nous.

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

Un son voisé est un signal quasi périodique et un son non voisé peut être considéré comme un bruit blanc.

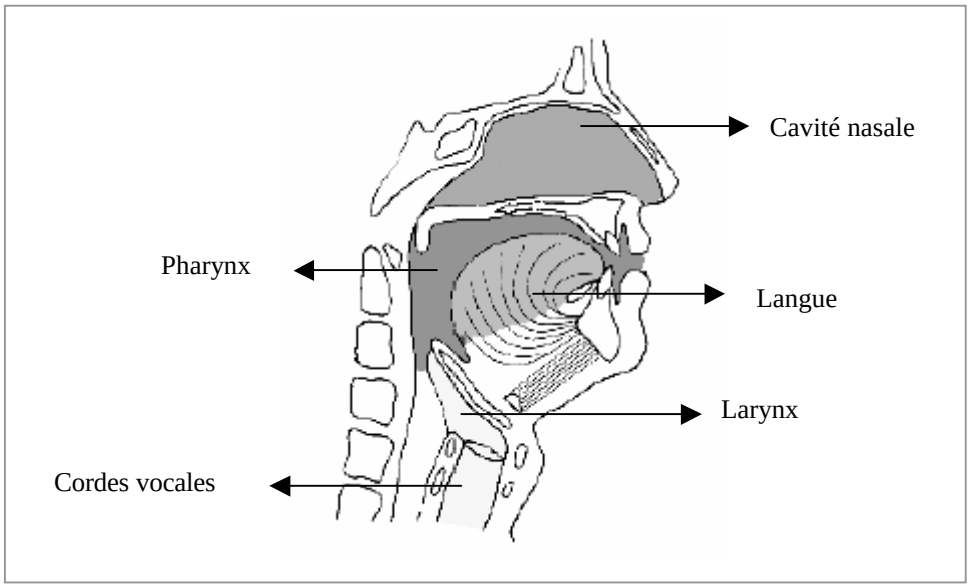


Figure 1.1 : Appareil phonatoire

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 [1]:

- De 80 à 200 *Hz* pour les hommes.
- De 150 à 450 *Hz* pour les femmes.
- De 200 à 600 *Hz* pour les enfants.

Un son voisé est un signal quasi périodique dont le spectre est illustré 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).

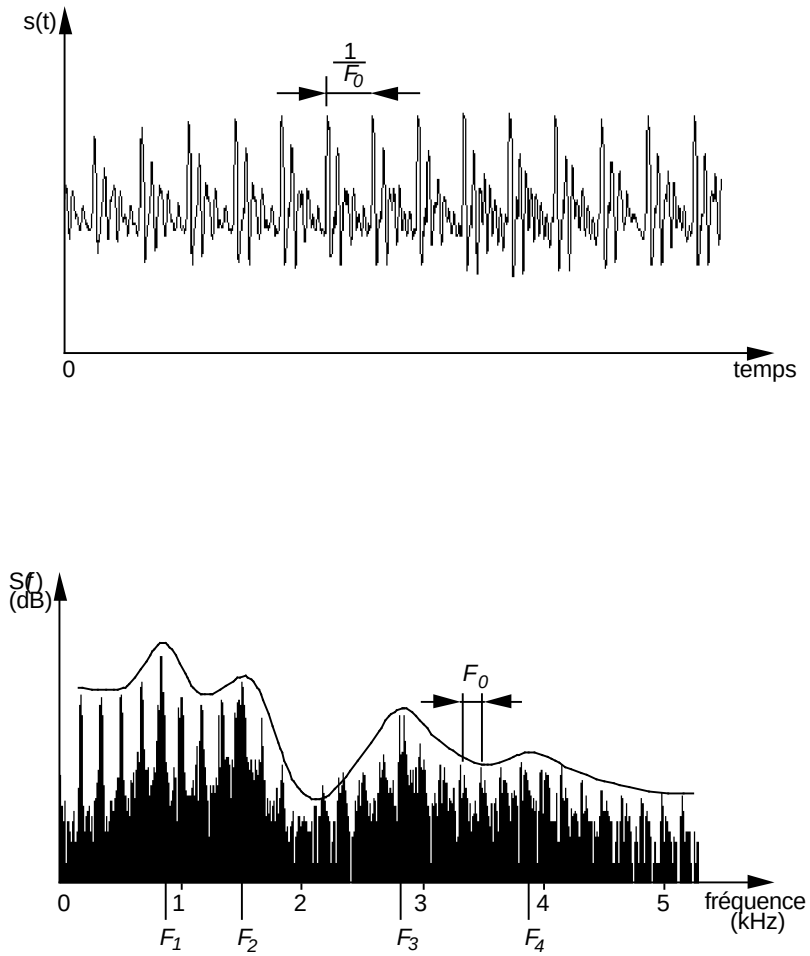


Figure 1.2: Forme d'onde et spectre du son voisé [1]

Un son non voisé ne présente pas de structure périodique, il peut être considéré comme un bruit blanc filtré par la transmittance de la partie du conduit vocal situé entre la constriction et les lèvres (figure1.3), son spectre ne présente donc pas de structure de pitch.

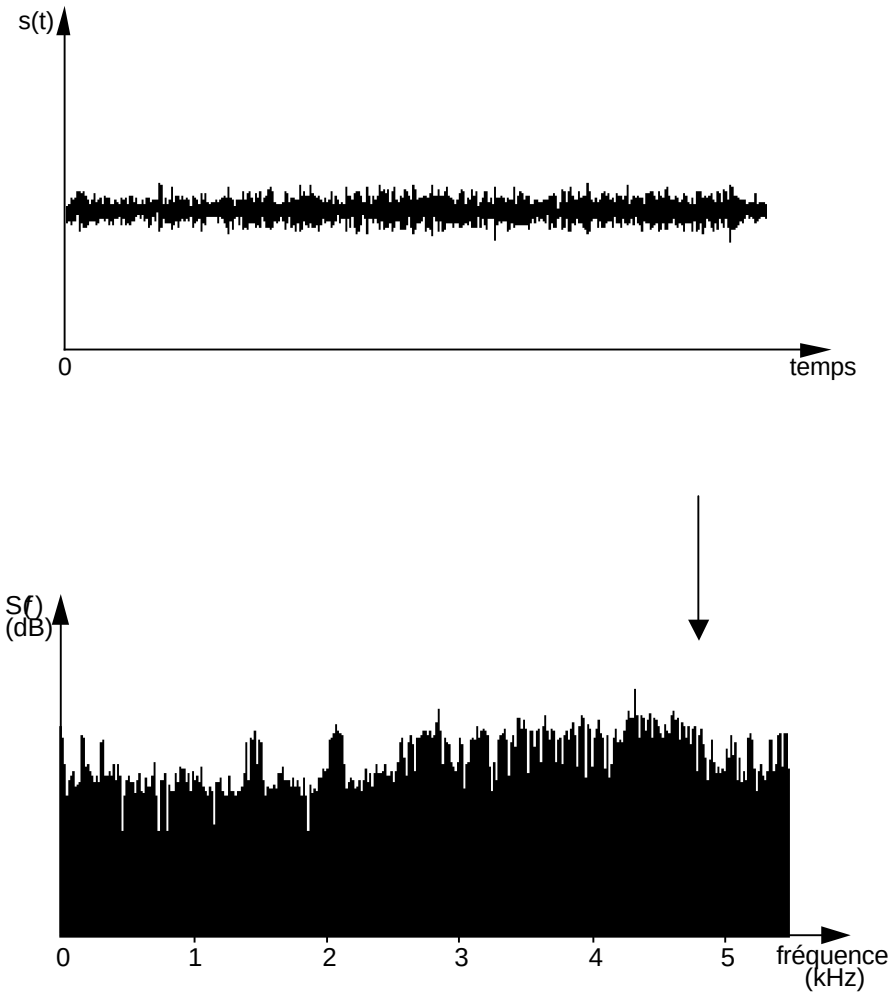


Figure 1.3 : Forme d'onde et spectre du son non voisé [1]

1.3- Modèle de production de la parole

Fant a proposé en 1960[2] 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*. On montre que cette modélisation reste valable dans le cas de sons non voisés, à condition que *u(n)* soit cette fois un bruit blanc. Le modèle final est illustré à la figure 1.4. Il est souvent appelé modèle auto régressif (AR), parce qu'il correspond dans le domaine temporel à une régression linéaire de la forme :

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

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.

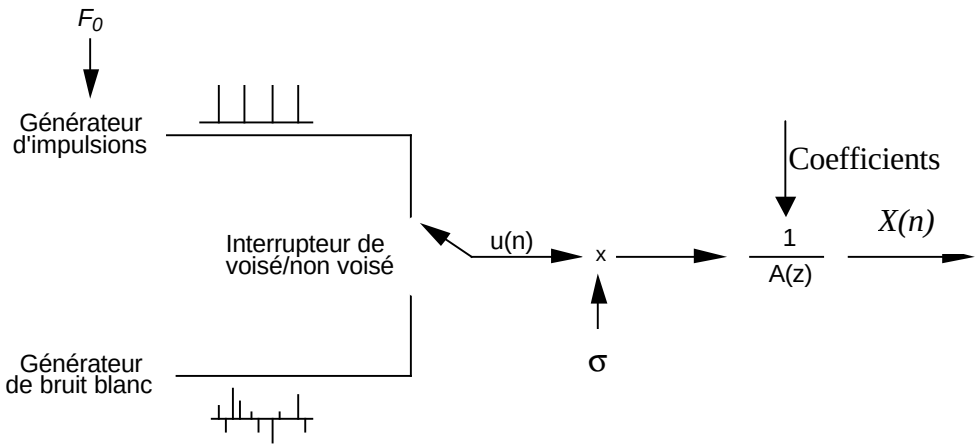


Figure 1.4: Modèle simplifié de production de la parole

Ce modèle comprend :

- un générateur périodique d'impulsions;
- un générateur de bruit blanc;
- un interrupteur servant à choisir les sons voisés ou non;
- un gain proportionnel à la valeur efficace du signal *s*[*n*];
- un filtre tous pôles $H(z) = 1/A(z)$.

Le problème de l'estimation d'un modèle AR, souvent appelée analyse LP (Linear Prediction) 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.

1.4- 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[3][4].

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) :

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

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 :

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

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

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 Régressif(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 passés 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 le codage par prédiction linéaire *LP* court-terme : la méthode d'auto corrélation et la méthode de covariance[3][4].

1.4.1- Méthode d'autocorrélation

La méthode d'autocorrélation garantit la stabilité du filtre LP. Les suppositions de cette méthode sont les suivantes :

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 fenêtre et le modèle (signal prédit) est minimisée sur l'ensemble des échantillons.

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

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

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

Pour trouver les coefficients du filtre LP, l'énergie du résiduel de prédiction sur l'intervalle fini 0• n• 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 :

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

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 $\boldsymbol{R}$. $\boldsymbol{a}$= $\boldsymbol{v}$ qui peut être ré-écrit en :

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

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

1.4.2- Méthode de covariance

La méthode 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 $dE/da_k = 0$

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:

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

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

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

où : • (i)= • (i.0) pour $1 \le i \le p$.

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

$$\bullet = LU \quad (1.18)$$

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

$$\bullet = CC^T \quad (1.19)$$

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

$$Ly = \bullet \quad (1.20)$$

puis :

$$Ua = y \quad (1.21)$$

1.4.3- Considération pratique

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

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

La fréquence d'échantillonnage est de 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 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 tranches d'analyse et leur décalage sont souvent fixées à 30 et 10 ms respectivement. Ces valeurs ont été choisies empiriquement; elles sont liées au caractère quasi-stationnaire du signal de parole.

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

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

Où N : est la longueur de la fenêtre

1.4.4- 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 fréquences spectrales *LSF (Line Spectral Frequency*). D'autres représentations incluent le coefficient de réflexion, les rapports d'aires logarithmiques LAR (Log Area Ratio), les coefficients spectraux, la réponse impultionnelle du filtre LP,….etc. Les paramètres *LSFs* se prêtent mieux à la quantification que les autres représentations du filtre LP[2].

- **Paires de raies spectrales(*LSP*)**

Connue aussi sous le nom de fréquence de raies spectrales (*LSF*), la représentation *LSP* (*Line Spectral Pair*) a été introduite par Itakura[2][19].

Soient les polynômes $P(z)$ et $Q(z)$ définis par:

$$\begin{cases} P(z) = A(z) + z^{-(p+1)} A(z^{-1}) \\ Q(z) = A(z) - z^{-(p+1)} A(z^{-1}) \end{cases} \quad (1.23)$$

Donc: $$A(z) = \frac{1}{2}[P(z) + Q(z)] \quad (1.24)$$

Les zéros des polynômes $P(z)$ et $Q(z)$ sont appelés les *LSP*. Ces polynômes ont les propriétés suivantes:

1.tous les zéros de $P(z)$ et $Q(z)$ se trouvent sur la cercle unité.

2.les zéros de $P(z)$ et $Q(z)$ sont entrelacés les uns aux autres, les *LSPs* sont dans un ordre croissant.

Il a été montré [3]que le filtre LP $A(z)$ est à phase minimum si et seulement si les *LSPs* satisfont les deux propriétés citées plus haut, donc la stabilité du filtre de synthèse est facilement vérifiable. De plus, les caractéristiques suivantes ont été relevées de [2][3][4]:

- Comme illustré à la figure 1.5, il y a une relation évidente entre les *LSPs* et le spectre du filtre LP. Une concentration des *LSPs* dans une certaine bande de fréquences correspond approximativement à une résonance dans cette bande.
- Sensibilité spectrale: Un changement d'une *LSP* cause seulement un changement dans la forme du filtre d'analyse dans une petite gamme de fréquence autour de cette *LSP*.

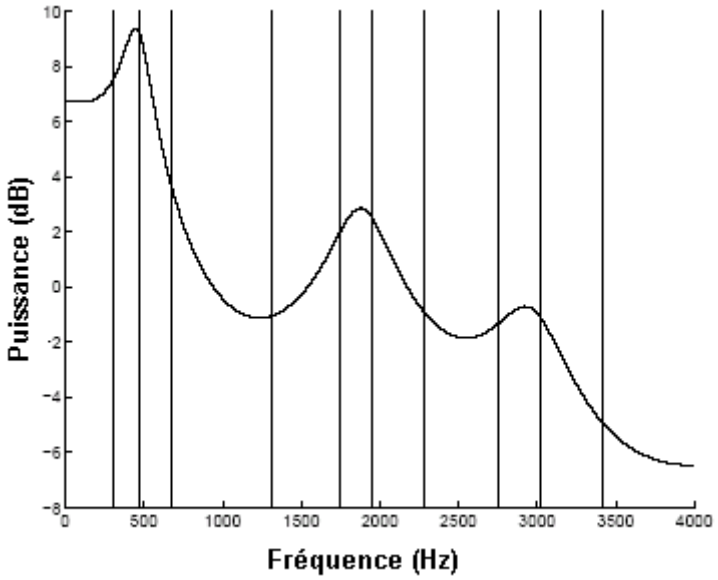


Figure 1.5 : Spectre *LP* avec *LSF* superposés[2]

1.5- Principe de la quantification

La quantification est l'opération de numérisation d'une ou plusieurs variables. Pour une seule dimension, le processus est bien connu sous le nom de quantificateur scalaire. Pour plusieurs dimensions, le quantificateur est dit vectoriel.

1.5.1 - Quantification vectorielle

Contrairement à la quantification scalaire, le quantificateur vectoriel (QV) s'applique sur des vecteurs. Il fait correspondre à tout vecteur d'entrée x de dimension k décrit comme suit :

x = (x (0), x (1), ..., x (k-1)) un vecteur de même dimension

yi = (yi(0), yi(1), ..., yi(k-1)) choisi parmi un ensemble fini **B** de N vecteurs de reproduction ou mots de code. Dans la littérature, on attribue à cet ensemble **B** l'appellation de dictionnaire.

Le débit binaire R d'un QV utilisant un dictionnaire B est défini par l'équation suivante:

$$R = \frac{1}{k}\log_2 N \qquad (1.25)$$

Le débit représente le nombre de bits par échantillon à coder. ($\log_2 N$) bits sont nécessaire pour représenter l'indice d'un vecteur de dimension k. Cette formulation du débit permet de faire des études comparatives de quantificateurs opérant sur des vecteurs de dimension différentes. La figure 1.6 illustre le principe de la quantification vectorielle[5].

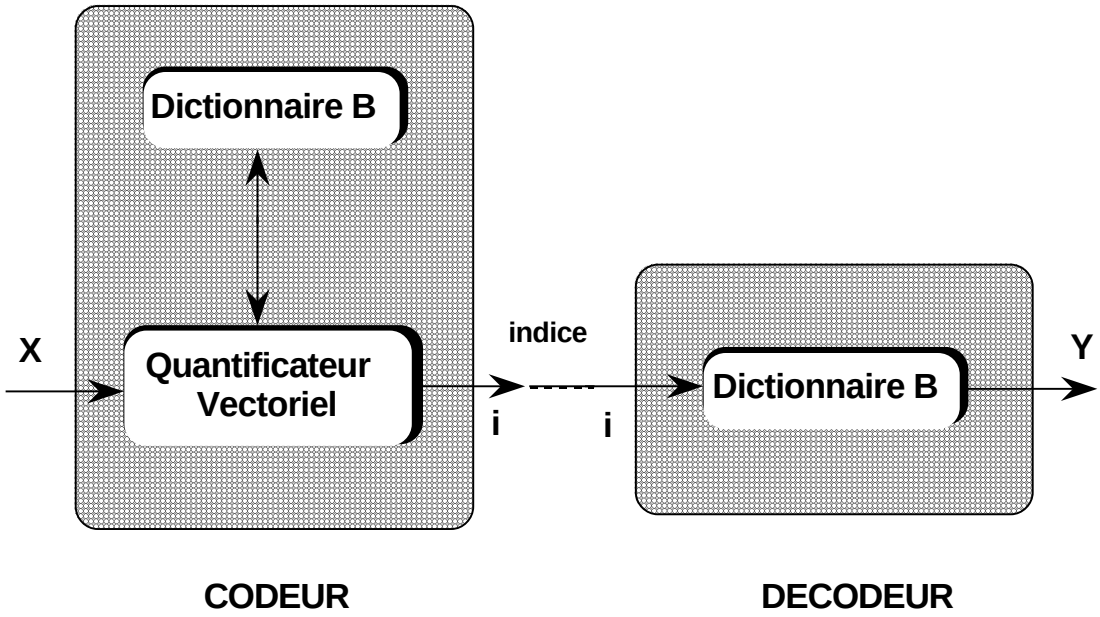


Figure 1.6 : Quantificateur vectoriel

Dans le codeur, on associe au vecteur d'entrée x un mot de code Y_i du dictionnaire selon le critère du plus proche voisin. Seul l'indice i est transmis au décodeur. On a N indices correspondant aux N mots de code du dictionnaire **B**. Le même dictionnaire est utilisé au décodage. Le mot de code y_i, représentant du vecteur x, est retrouvé à partir de l'indice i reçu.

1.5.2- Quantification vectorielle par Split

Dans la quantification par Split (SVQ), le vecteur des paramètres à quantifier x est divisé en plusieurs (k) blocs [x0 x1 ... xk-1] et chaque bloc est quantifié séparément par un quantificateur vectoriel à recherche exhaustive [3][5][6].

On sait que la SVQ réduit la complexité au prix d'une dégradation de performance. D'où il y a un compromis entre la complexité et les performances, qui détermine le nombre de blocs à prendre dans le split[31].

1.6- Techniques de codage de la parole

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

Les algorithmes de codage de la parole peuvent être divisés en trois catégories[7] :

- Codage de forme d'onde (waveform coding).
- Codage paramétrique (parametric coding).
- Codage hybride (hybrid coding).

La figure 1.7 montre la différence de qualité de parole qui existe entre les codecs.

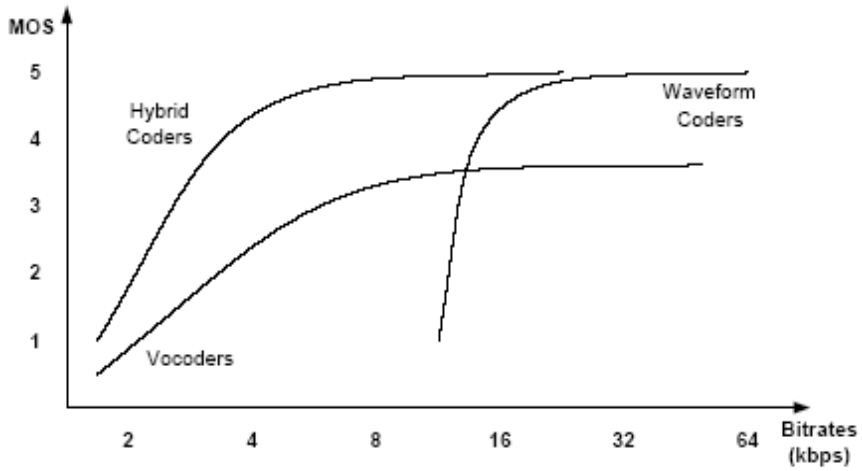


Figure 1.7 : Comparaison de la qualité de codage de parole [7]

1.6.1- Codage de forme d'onde

Les codeurs de ce type fonctionnent à des hauts débits (supérieurs à 16 kb/s).Ils essaient de reconstruire la forme du signal de manière aussi proche que le signal d'origine et basés sur des échantillons du signal d'origine. En théorie, cela signifie que ces codeurs sont indépendants du signal, et peuvent fonctionner avec des signaux non vocaux, de type modem ou fax par exemple. Ces codeurs sont relativement simple à mettre en œuvre, et produisent une qualité acceptable jusqu'à des débits de 16 Kb/s. En deçà, la qualité du signal reconstruit se dégrade rapidement.

La PCM (Pulse Coded Modulation) est un exemple de cette technique. Dans le cas d'une quantification linéaire, au moins 12 bits par échantillon sont nécessaires pour assurer une bonne qualité, ce qui conduit à un débit de 96 Kb/s (8000 échantillons de 12 bits). Cependant, la nature de la parole et de l'oreille humaine ne suit pas une échelle linéaire. La plupart des signaux vocaux sont de faible amplitude, et l'oreille humaine n'est pas sensible à l'amplitude absolue d'un son, mais au logarithme de l'amplitude. La représentation binaire (nombre de bits) est alors plus importante pour les signaux de faible amplitude que pour les signaux de forte amplitude. Cela conduit à un débit de 64 Kb/s.

D'autres techniques de ce type existent comme la DPCM (Differential PCM), ADPCM (Adaptative Differential Pulse Code Modulation) et ADM (Adaptive Delta Modulation).

1.6.2- Codage paramétrique

Les codeurs paramétriques sont destinés à fonctionner pour des débits de quelques dizaines de bits par seconde à 4Kb/s. La performance de ce type de codage; connu aussi sous le nom de codage de source ou vocodeurs, dépend aux modèles de production de la parole. Ces codeurs sont désignés pour des applications à bas débit et sont destinés à maintenir l'intelligibilité de la parole. La plupart de ces codeurs sont basés sur le codage linéaire prédictif LP.

Le codage LP consiste à synthétiser des échantillons à partir d'un modèle d'un système de production vocal et d'une excitation. Pour la voix humaine, le système de production vocal est l'ensemble poumons-cordes vocales -trachée -gorge -bouche -lèvres. En pratique, on modélise ce système par un ensemble de cylindres de diamètres différents, 10 dans le cas de LP-10, excités par un signal qui est soit une sinusoïde, soit un bruit blanc. Le choix de la fonction d'excitation (sinusoïde ou bruit blanc) dépend des caractéristiques, voisée ou non voisée, du signal.

1.6.3- Codage hybride

Les codeurs hybrides sont destinés à fonctionner pour des débits moyens de 4 kb/s à 16 kb/s, ils combinent les caractéristiques des deux techniques précédentes. Le principal représentant de cette classe est le codage Prédictif Linéaire à Excitation par séquences Codées CELP (Code Excited Linear Prediction).

Le codage CELP est le plus efficace et le plus utilisé actuellement. C'est est une extension du codage LP. L'idée de base est que la série des impulsions utilisées comme excitation du filtre prédictif est un ensemble discret et fini. On peut donc essayer d'opérer une quantification vectorielle de cet ensemble, de constituer un dictionnaire. Il suffit alors de transmettre le numéro dans le dictionnaire de la forme d'excitation la plus proche [8][9].

1.7- 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 du signal toute en maintenant des niveaux adéquats de qualité du signal, de complexité d'implantation et de retard communication.

1.7.1- 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.

1.7.2- Débit binaire

On mesure le débit 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. La fréquence d'échantillonnage doit être au moins deux fois plus grande que la largeur de bande du signal correspondant. Dans le cas de la téléphonie, pour une bande de 3.2 KHZ (200-3400 HZ), la fréquence d'échantillonnage de 8 KHZ est utilisée.

1.7.3- 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 les cartes de traitement du signal (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 de l'encodeur ou du décodeur ou codec, le prix et la consommation de puissance, ce dernier étant un important critère dans un système portable.

1.7.4- Retard de communication

La complexité dans un algorithme de codage est souvent accompagnée d'une augmentation de la durée de traitement dans l'encodeur et le décodeur. Bien que l'évolution des capacités des processeurs de traitement du signal est 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. Selon l'environnement de communication, le retard total permis à un sens peut être aussi bas qu'une milliseconde (comme en réseaux téléphoniques sans annulateur d'écho).

Le retard de codage à un seul sens est défini comme étant le temps écoulé entre l'instant où l'échantillon du signal de parole arrive à l'entrée de l'encodeur et l'instant où le même échantillon apparaît à la sortie du décodeur, moins tout retard introduit par les autres équipements de communication (comme les MODEM) entre la paire encodeur-décodeur et le retard de propagation du signal qui dépend de la distance. En d'autres termes, c'est comme si l'encodeur et le décodeur sont directement connectés par fils sans aucun équipement entre eux. Cette définition fait que le retard de codage dépend seulement de l'algorithme de codage.

Avec cette définition, le retard de codage des codeurs CELP peut être grossièrement déterminé en fonction de la taille de la trame du signal de parole utilisée.

1.8- 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 (*RSB*).

Si : S est le signal de parole original.

$\overline{S}$ est le signal de parole synthétisé.

Alors le signal d'erreur est donné par :

$$e(n) = S(n) - \overline{S}(n) \tag{1.26}$$

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

$$E_S = \sum_{n=0}^{N-1} S^2(n) \tag{1.27}$$

Et l'énergie de l'erreur :

$$E_e = \sum_{n=0}^{N-1} e^2(n) \tag{1.28}$$

Le *RSB* est alors donné par :

$$RSB = 10 \log\left(\frac{E_S}{E_e}\right) \text{en } dB \tag{1.29}$$

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 *RSB* pourra être très important comme très faible.

On utilise plutôt le *RSB* segmental. Le signal est découpé en M segments de 15 à 30 ms puis on calcule une moyenne des *RSB*.

$$RSB_{seg} = \frac{1}{M}\sum_{i=10}^{M} 10\log\left(\frac{\sum_{n=0}^{N-1} S^2(n)}{\sum_{n=0}^{N-1} e^2(n)}\right) \tag{1.30}$$

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 *RSB* n'est pas nécessairement corrélé avec la qualité d'écoute.

Les méthodes les plus utilisées sont les suivantes[7]:

- **D**iagnostic **R**hyme **T**est (DRT) qui mesure l'intelligibilité sur un grand nombre de mots.
- **D**iagnostic **A**cceptability **M**esure (DAM) qui mesure le naturel perçu de la parole.
- **M**ean **O**pinion **S**core (MOS) ou l'auditeur évalue un codeur sur une échelle absolue allant de 1 à 5 avec :

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

Table 1.1 : Qualité avec la mesure MOS

1.9- Conclusion

Le codage consiste à réduire le volume d'informations à transmettre en gardant une qualité de la parole acceptable. 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 LP) qui caractérisent le comportement du signal. La simplicité de concept et 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 parole.

Chapitre 2

Transmission de la voix sur les réseaux IP(VoIP)

2.1- Introduction

La téléphonie par Internet est une technologie récente et comporte certains nombres de problèmes de développement dû au fait que Internet n'était pas conçu pour une transmission à temps réel tel que la voix et la vidéo.

La voix sur IP utilise le Protocole Internet pour transmettre notre voix en paquets dans un réseau. Des protocoles de signalisations sont utilisés pour mettre en place ou arrêter les appels.

Les bénéfices majeurs de cette technologie sont:

- L'intégration de la voix et des données: en effet ce premier sera demandé par des programmes de multi-applications, l'évolution inévitable étant des serveurs web permettant de faire de l'interaction avec la voix, les données et les images.
- La simplification des infrastructures: une infrastructure intégrée supportant tout type de communication revient à avoir une meilleure standardisation et moins de gestion d'équipement.
- Un réseau plus efficace: l'intégration de la voix et des données remplira les chaînes de communication de données efficacement donnant donc une bande passante renforcée. En effet, aucune bande passante n'est donnée pendant les périodes silencieuses.
- Une réduction de prix: le coût des services PSTN (Public Switched Telephone Network, le réseau téléphonique publique) peut être contourné en utilisant Internet permettant ainsi des coûts d'appel à longue distance moins coûteux.

2.2- Le Protocole IP

Le protocole IP (Internet Protocol) permet aux paquets de se déplacer sur Internet, indépendamment les uns des autres, sans liaison dédiée. Chacun d'entre eux, envoyé sur le réseau, se voit attribuer une adresse IP. Cette dernière est un en-tête accolé à chaque paquet et contenant certaines informations, notamment, l'adresse source, l'adresse destinataire, son temps de vie, le type de service,... ,etc.

2.3- La Voix sur IP

La voix sur IP ou VoIP (Voice over Internet Protocol) est le transfert de conversations vocales sous forme de données sur un réseau IP. Contrairement aux PSTN (Public Switched Telephone Network); dans les appels VoIP, la connexion téléphonique est à commutation de paquets.

Avec un appel VoIP, la partie de l'établissement de l'appel doit être simulée: la tonalité, les signaux de sonneries et les signaux occupés. La partie audio de l'appel elle-même (la conversation) a besoin d'être converti de son format analogique à un format numérique, découpée en paquets, envoyée à travers le réseau dans un format paquet, ré-assemblée, et à présent, reconverti du format numérique au format analogique. Les codecs (codeur-décodeur) à chaque bout font la conversion de l'analogique au numérique et vice vers ça.

2.4- Les composants VoIP

Pour transférer de la voix sous forme de données sur le même réseau transportant les e-mails et les pages Web, un nouvel ensemble de composant est rajouté à ceux déjà existants. Parmi ces composants, on peut citer les suivants :

- Les codecs.
- Les TCP/IP(Transmission Control Protocol)/Internet Protocol).
- Les protocoles VoIP.
- Les serveurs de téléphonie IP et les PBXs(Private Branch eXchange).
- Les routeurs et les gateways VoIP.
- Les gatekeepers.
- Les téléphones IP et les softphones.

2.4.1- Les Codecs

Pour faire passer la voix dans un réseau IP, il faut tenir comptes de certains paramètres. On estime la bande passante de la voix à 4 kHz (400-3400 Hz). Ce qui donne après numérisation une bande passante de 8 kHz et après codage un débit de 64 kb/s. Ceci entraîne, que si on veut transmettre la voix sur un réseau IP sans mécanisme d'optimisation de la taille, il nous faudra une bande passante « continue » de 64 kb/s rien que pour la partie données à transmettre.

Toutefois, ceci est rarement le cas, car à l'aide de mécanisme de codage optimisé et grâce aux lois de compression de l'information, on arrive à réduire ce débit nécessaire de plus de 8 fois pour les meilleurs algorithmes de codages. L'abréviation de ces algorithmes généralement utilisée est codec (Codeur-Décodeur). Afin d'avoir la meilleure qualité possible de la voix et après avoir passé à travers un codec, on peut définir les exigences suivantes[10]:

- Robustesse contre les erreurs binaires: masquage d'erreur nécessaire.
- Robustesse contre les pertes de paquets: généralement, pas de temps pour demander la retransmission de paquets; dégradation progressive de la qualité vocale en cas de perte.

La table 2.1 rassemble quelques codecs et leurs performances.

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	-	-

Table 2.1 : Les codecs et leurs performances

2.4.2- Le Protocole de Contrôle de Transmission (TCP/IP)

Le protocole TCP établit un mécanisme d'acquittement et de re-émission de paquets manquants. Ainsi, lorsqu'un paquet se perd et ne parvient pas au destinataire, TCP permet de prévenir l'expéditeur et lui réclame de ré-acheminer les informations non parvenues. Il assure d'autre part un contrôle de flux en gérant une fenêtre de congestion qui module le débit d'émission des paquets. Il permet donc de garantir une certaine fiabilité des transmissions.

2.4.2.1- Le Protocole de Contrôle de Transmission (TCP)

Le rôle de ce protocole est d'éviter que des données soient perdues, dupliquées, ou désordonnées. Il est, aussi, connu comme étant un protocole en mode connecté parce qu'il a besoin d'une requête et d'un acquittement avant de commencer le transfert de données.

2.4.2.2- Le protocole datagramme utilisateur (UDP)

Le protocole UDP (User Datagramme Protocol), permet aux applications d'échanger des datagrammes. Ce protocole utilise la notion de port qui permet de distinguer les différentes applications qui s'exécutent sur une machine. En plus, du datagramme et de ses données, un message UDP contient, à la fois un numéro de port source et un numéro de port destination. Le protocole fournit un service en mode non connecté et sans reprise sur erreur. Il n'utilise aucun acquittement, ne reséquence pas les messages, et ne met en place aucun contrôle de flux. Les messages UDP peuvent être perdus, dupliqués, remis hors séquence ou arriver trop tôt pour être traités lors de leur réception.

2.4.3- Les protocoles VoIP.

Ils sont de deux sortes:

2.4.3.1- Les protocoles de signalisation

Les protocoles de signalisation d'appel utilisent les protocoles TCP et UDP pour encapsuler les phases d'établissement et de libération des ressources d'un appel. Par exemple, dans la téléphonie sur IP, ils prennent en charge différentes fonctions comme la correspondance entre les numéros téléphoniques et les adresses IP, génération de la tonalité et des signaux occupés. Le protocole H.323 est le protocole de signalisation d'appel le plus largement déployé. Le protocole MGCP (Media Gateway Control protocol) est moins flexible pour une utilisation avec un équipement traditionnel comme les téléphones de maison.

La famille des protocoles H.323 est une famille de protocoles robustes et flexibles ceci a été la conséquence de plusieurs années de raffinage. Mais le coût de cette robustesse est qu'il faut toute une série d'acquittements et de données échangées pour chaque fonction exécutée pendant une session d'appel.

Le SIP (Session Initiation Protocol) et MGCP sont des protocoles légers développés par l'IETF(Internet Engineering Task Force) dans les réseaux de données. Le SIP en particulier représente typiquement la logique des réseaux de données qui demande pourquoi utiliser de lourds protocoles (tel que H.323) quand un protocole léger (tel que le SIP) accomplira, la plupart du temps, le même travail.

2.4.3.2- Les protocoles temps réels (steaming protocols)

Il existe deux types de protocoles:

- ***Protocole de transport en temps réel(RTP: Real Time Transport Protocol)***

Le RTP est un protocole adapté aux applications présentant des propriétés temps réel. Il permet ainsi de :

- Reconstituer la base de temps des flux (horodatage des paquets: possibilité de re-synchronisation des flux par le récepteur).

- Détecter les pertes de paquets et en informer la source.

- Identifier le contenu des données pour leurs associer un transport sécurisé.

- ***Protocole de contrôle de transport en temps réel (RTCP: Real Time Transport Control Protocol)***

Ce protocole a pour but de transmettre périodiquement des paquets de contrôle à tous les participants d'une session. Pour expliquer les différents paquets de contrôle fournis par RTCP, il faut voir dans quel contexte RTP est utilisé. Par exemple, une application de visioconférence pourra l'utiliser pour transmettre les caractéristiques de chacun des participants.

Ce protocole défini quatre principaux paquets de contrôle :

SR (Sender Report) : ce rapport regroupe des statistiques concernant la transmission (pourcentage de perte, nombre cumulé de paquets perdus, variation de délai, …etc). Ces rapports sont issus de l'émetteurs actifs d'une session.

RR (Receiver Report) : ensemble de statistiques portant sur la communication entre les participants. Ces rapports sont issus des récepteurs d'une session.

SDES (Source Description) : carte de visite de la source (nom, e-mail, localisation).

BYE : message de fin de participation à une session.

2.5- Techniques de recouvrements de paquets

Vu l'impacte très néfaste qu'ont les pertes de paquets sur la qualité des transmissions des flux sonores, plusieurs techniques de recouvrement de paquets perdus ont été mises en œuvre. Ces techniques résumées se divisent en deux parties complémentaires[11][14][15]:

-Réparation basée au niveau de l'émetteur (Sender-Based repair).

-Réparation basée au niveau du récepteur (Reciever-Based repair).

2.5.1- Réparation basée au niveau de l'émetteur (Sender-Based repair)

La figure 2.1.rassemble les techniques de réparation basées au niveau de l'émetteur. Pour éviter les malentendus dans ce qui suit, nous avons distingué une trame de donnée d'un paquet de donnée. Une trame représente un intervalle du flux sonore. Un paquet peut contenir une ou plusieurs trames encapsulées afin d'être envoyées sur le réseau.

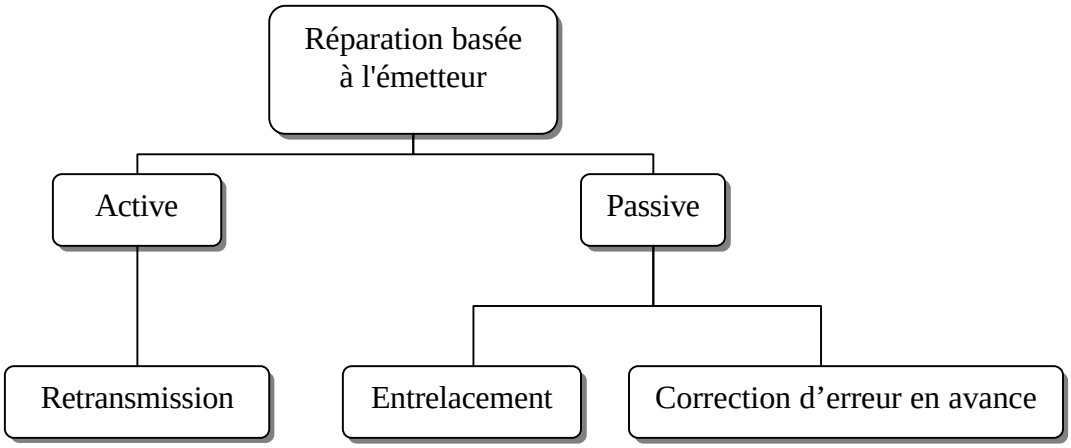


Figure 2.1 :. Classification des techniques de réparations basées à l'émetteur

2.5.1.1- Correction d'erreur en avance(FEC : Forward Error Correction)

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

Le principe est illustré dans la figure 2.2

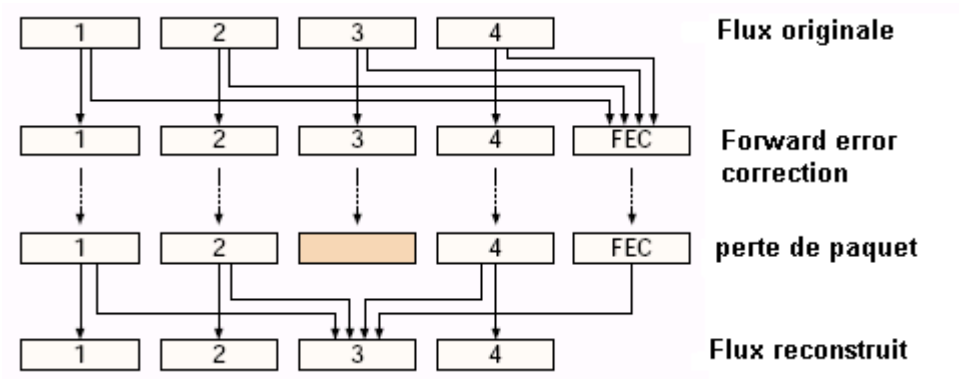


Figure 2.2 :. Exemple de FEC[11]

Plusieurs avantages découlent de cette méthode, nous pouvons citer la faible demande en ressource de calcul et la simplicité de l'implémentation. En contre partie, cette technique impose un retard supplémentaire, une augmentation de la bande passante et une difficile implémentation au niveau du décodeur[11][14].

2.5.1.2- Entrelacement

La technique d'entrelacement ou interleaving est très utile lorsque, les paquets contiennent plusieurs trames et le délai de end to end (bout-en-bout) n'est pas important[12][14]. Avant transmission du flux, les trames sont ré-arrangées de telle manière que celles, initialement, adjacentes se retrouvent séparées dans le flux transmis, puis remises dans leur ordre original au niveau du récepteur.

En conséquence, les effets d'effacement de paquets, sont dispersés. La figure 2.3 illustre un exemple ou chaque paquet contient 4 trames[11][14].

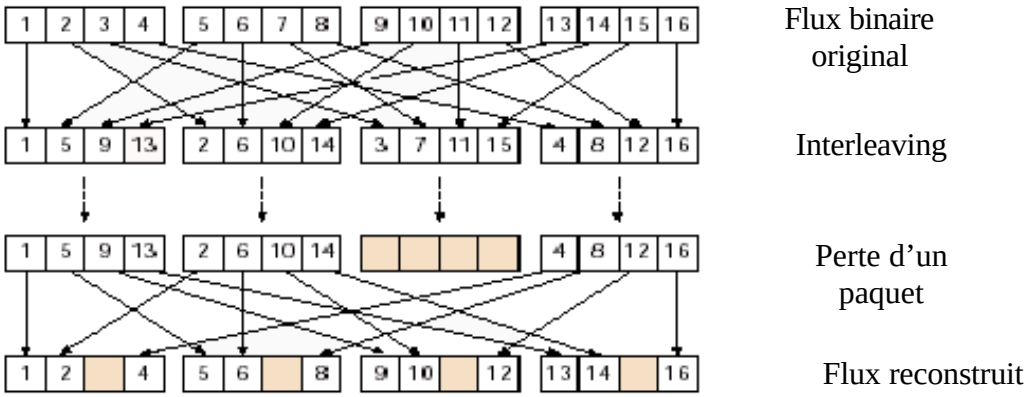


Figure 2.3 : Exemple d'Interleaving [11]

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

2.5.1.3- Retransmission

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

2.5.2- Réparation basée au niveau du récepteur

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

Le masquage d'erreur repose sur le principe de remplacer les paquets perdus par des paquets similaires aux originaux. Ceci reste possible du fait de la similarité à court-terme du flux. La figure 2.4 illustre les différentes techniques de masquage d'erreur.

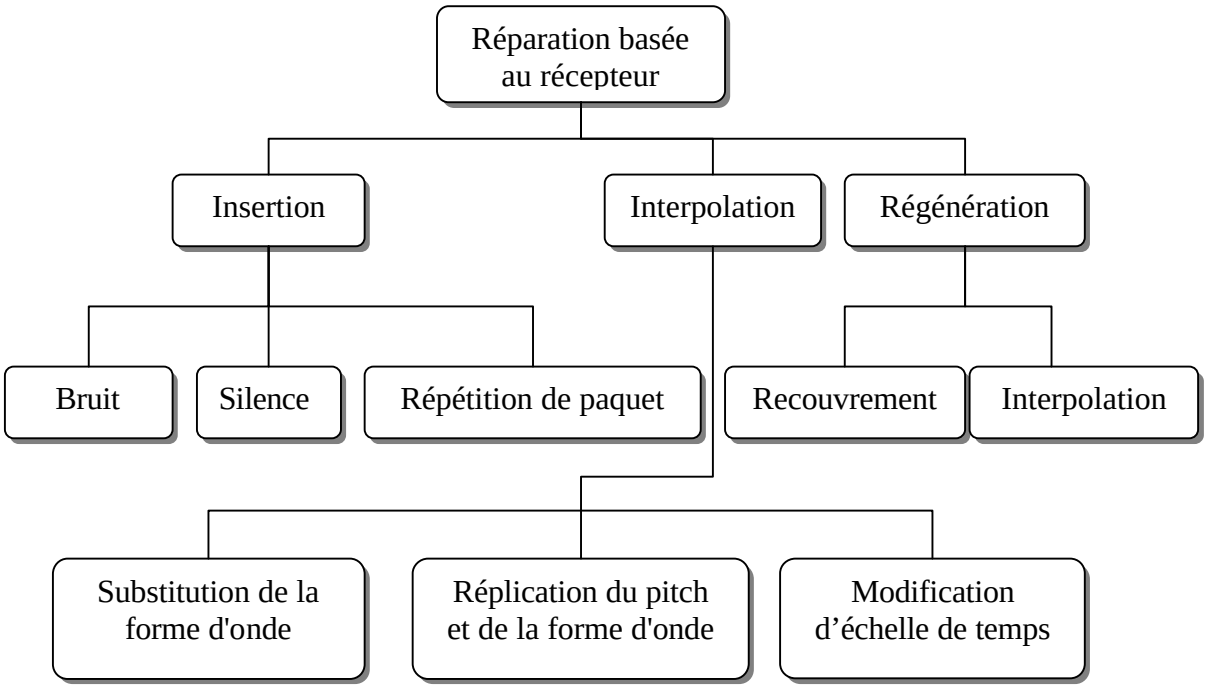


Figure 2.4 : Classification des techniques de masquage d'erreur

2.5.2.1-L'insertion

Cette technique répare les erreurs par l'insertion de paquets-remplaçants. Ces paquets peuvent être soit des silences, du bruit ou carrément la répétition du paquet précédent la perte. Ces techniques sont faciles à implémenter mais, avec l'exception de la répétition, offrent de faibles performances.

2.5.2.2- L'interpolation

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

- ***Substitution de la forme d'onde***

Cette technique met à profit le signal d'avant et, optionnellement, d'après perte pour trouver un signal convenable pour combler le pertes[11][12].

- ***Réplication du pitch et de la forme d'onde***

Cette méthode est une amélioration de la méthode précédente et semble donner de meilleurs résultats, elle utilise, en plus, un algorithme de détection du pitch des deux cotés d'une perte de paquet[11][13].

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

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

2.5.2.3- La régénération

Cette technique exploite les connaissances sur les algorithmes de compression audio afin de retirer les paramètres audio manquants, et ainsi synthétisé les paquets perdus. En conséquence, cette méthode est forcément utilisée dans les codec. De plus, elle est, généralement, d'un point de vu complexité, coûteuse. Cependant, et grâce à une grande quantité d'information, cette méthode a de bonnes performances.

- ***Interpolation de l'état transmis***

Pour les codecs fondés sur le codage par transformation ou par prédiction linéaire, il est possible que le décodeur interpole entre plusieurs états[11].

- ***Modèle***

Le signale parole, d'un coté ou des deux cotés d'une perte, est ajusté à un modèle, qui est utilisé pour régénérer le signal de parole qui comblera la période de pertes.

2.6- Conclusion

Les réseaux IP se montrent robuste et flexible aux pertes se produisant sur des données de type fax, images, bande son stockée. En effet, les données perdues peuvent être ré-émises à partir de la source et ré-introduites à la destination sans que l'intégralité ou que la qualité des fichiers reçus soit altérée. Cependant, dans des applications en temps réel, tel la VoIP, la procédure de ré-émission risque d'être impraticable tant les exigences temporelles sont strictes. Ainsi, lors d'envoi de séquences de paroles, c'est le protocole UDP/IP qui est utilisé. En effet, comme vu plus haut et contrairement au TCP/IP, ce protocole ne requière pas d'interactions initiales avec le destinataire, ni de ré-émission de paquets lors de pertes.

Chapitre 3

Codeur de la norme G.729

3.1- Introduction

Ce chapitre décrit un des codeurs les plus utilisés en codage de signaux vocaux à travers les réseaux d'Internet. 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*).

Le codec standard G.729 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 (Recommandation G.712) puis en l'échantillonnant à 8000 Hz et en le convertissant en signal PCM linéaire à mots de 16 bits, qui est injecté dans le codeur. Inversement, on reconvertira le signal de sortie du décodeur en signal analogique[16][18].

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. Toutes les trames de 10 ms, le signal vocal est analysé pour en extraire les paramètres du modèle de prédiction CELP (coefficients du filtre de prédiction linéaire, index et gains de répertoire codé adaptatif et de répertoire codé fixe). Ces paramètres sont codés et transmis. L'affectation des positions binaires aux paramètres de codage est représentée dans la Table 3.1.

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

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

3.2.1- Codeur

Le principe du codage est schématisé sur la figure 3.1. Le signal d'entrée subit un filtrage passe-haut et une normalisation dans le bloc de prétraitement. la sortie de ce dernier sera utilisée comme entrée pour toutes les analyses suivantes. L'analyse prédictive linéaire est effectuée toutes les trames de 10 ms afin de calculer les coefficients de filtrage prédictif linéaire. Ceux-ci sont convertis en paires de lignes spectrales LSP (*Line Spectrum Pairs*) et numérisés sur 18 éléments binaires *(L0, L1, L2, L3)* par quantification vectorielle VQ (*Vector Quantization*) prédictive en deux étapes. Le signal d'excitation est choisi au moyen d'une procédure de recherche par analyse et synthèse dans laquelle l'erreur entre le signal vocal original et le signal vocal reconstitué est minimisée en fonction d'une mesure de distorsion pondérée par la perception. A cette fin, le signal d'erreur passe par un filtre de pondération perceptive dont les coefficients sont déduits du filtre de prédiction linéaire avant quantification. Les poids de la pondération perceptive sont rendus adaptatifs afin d'améliorer la qualité des signaux d'entrée ayant une réponse en fréquence uniforme.

Les paramètres d'excitation (par répertoire codé fixe et par répertoire codé adaptatif) sont déterminés à chaque sous-trame de 5 ms (soit 40 échantillons). Les coefficients du filtre de prédiction linéaire, quantifiés et non quantifiés, sont utilisés pour la deuxième sous-trame, alors que la première utilise une interpolation des coefficients du filtre de prédiction linéaire (aussi bien quantifiés que non quantifiés). Le délai tonal en boucle ouverte est estimé toutes les trames de 10 ms, sur la base du signal vocal issu du pondérateur perceptif. Les opérations suivantes sont reprises pour chaque sous-trame. Le signal cible $x(n)$ est calculé par filtrage de l'énergie résiduelle du codage prédictif linéaire dans le filtre de synthèse pondérée $W(z)/\hat{A}(z)$. Les états initiaux de ces filtres sont mis à jour par filtrage de l'erreur mesurée entre l'énergie résiduelle du codage prédictif linéaire et l'excitation. Cela équivaut au procédé courant consistant à soustraire – du signal vocal pondéré – la réponse à entrée nulle du filtre de synthèse pondérée. La réponse impulsionnelle $h(n)$ du filtre de synthèse pondérée est calculée. Une analyse tonale en boucle fermée est ensuite effectuée (afin de déterminer le délai et le gain par répertoire codé adaptatif) au moyen du signal cible $x(n)$ et de la réponse impulsionnelle $h(n)$, par recherche autour de la valeur du délai tonal en boucle ouverte. On utilise un délai tonal fractionnaire, de résolution 1/3. Ce délai tonal est codé sur 8 éléments binaires dans la première sous-trame et codé différentiellement sur 5 éléments binaires dans la deuxième sous-trame. Le signal cible $x(n)$ est mis à jour par soustraction de la contribution (filtrée) du répertoire codé adaptatif et ce nouveau signal cible, $x'(n)$, est utilisé lors de l'exploration du répertoire codé fixe afin de déterminer l'excitation optimale. On fait appel à un répertoire algébrique de mots de 17 éléments binaires pour l'excitation par répertoire codé fixe. Les gains des contributions par répertoire codé adaptatif et par répertoire codé fixe sont quantifiés vectoriellement sur 7 éléments binaires (avec application au gain par répertoire codé fixe d'une prédiction par analyse à moyenne mobile). Finalement, les mémoires des filtres sont mises à jour au moyen du signal d'excitation ainsi déterminé.

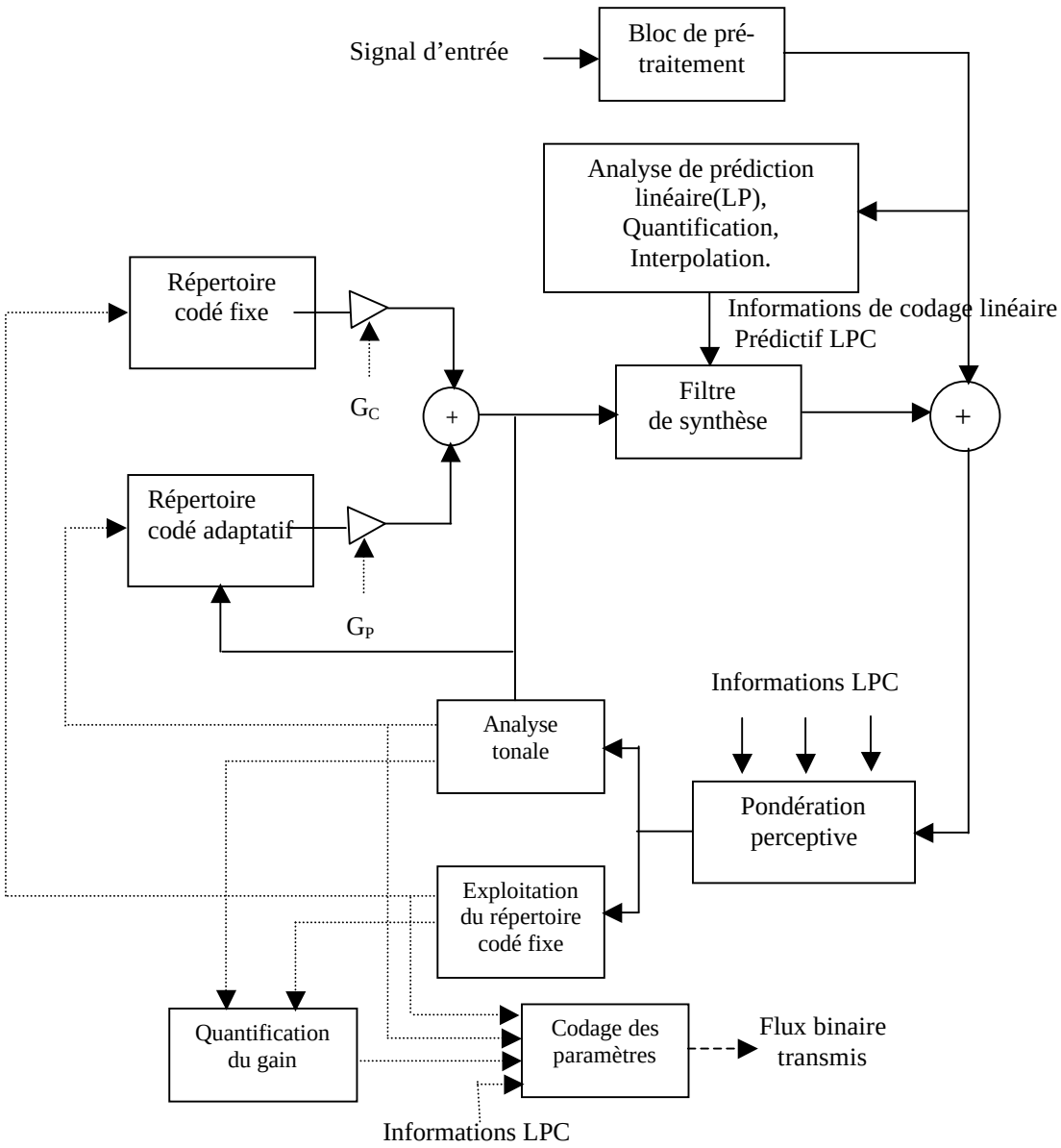


Figure 3.1 : Principe du codeur *CS-ACELP* ***G.729*** [16]

3.2.2- Décodeur

Le principe du décodeur est représenté sur la Figure 3.2. Les index paramétriques sont d'abord extraits du flux binaire reçu. Ces index sont ensuite décodés pour obtenir les paramètres de codage correspondant à une trame vocale de 10 ms. Ces paramètres sont les coefficients convertis en paires de raies spectrales (LSP), les 2 délais tonaux fractionnaires, les 2 vecteurs de répertoire codé fixe et les deux séries de gains par répertoire codé adaptatif et par répertoire codé fixe. Les coefficients en paires LSP sont interpolés et reconvertis en coefficients de filtre de prédiction linéaire pour chaque sous trame de 5 ms, qui passe par les étapes suivantes:

- L'excitation est construite par combinaison des codes vectoriels adaptatifs et fixes, normalisés par leur gain respectif;
- Le signal vocal est reconstitué par filtrage de l'énergie d'excitation dans le filtre de synthèse du codage prédictif linéaire.
- Le signal vocal reconstitué est envoyé dans un bloc de post-traitement, qui comprend un post-filtre adaptatif utilisant la sortie des filtres de synthèse à court et à long terme, suivi d'un filtre passe-haut et d'un échantillonneur-normalisateur.

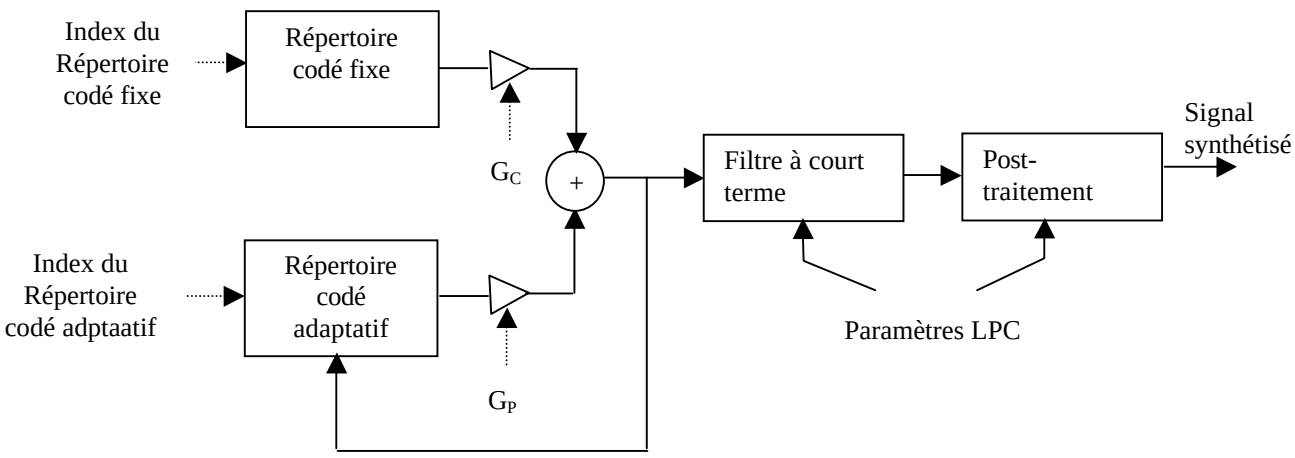


Figure 3.2 : Principe du décodeur CS-ACELP **G.729** [16]

3.2.3- Délai

Ce codeur numérise les signaux audio, en particulier vocaux, sous la forme de trames de 10 ms. Il s'y ajoute un délai d'exploration de 5 ms, ce qui porte le délai algorithmique total à 15 ms. Tous les délais additionnels d'une mise en œuvre concrète de ce codeur sont dus à ce qui suit:

- temps de traitement nécessaire pour les opérations de codage et de décodage;
- temps de transmission dans la liaison de communication;
- délai de multiplexage lors de la combinaison de données audio avec d'autres données.

3.3- Quantification des coefficients LSP

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:

$$\omega_i = \arccos(q_i) \qquad i = 1,...,10 \qquad (3.1)$$

Afin de réduire la bande passante, le codec G.729 utilise une prédiction par moyenne mobile périodique du $4^{ème}$ 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 étapes. La première étape est une quantification vectorielle à 10 dimensions qui utilise le répertoire 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 répertoires à 5 dimensions, L2 et L3 contenant 32 entrées (5 bits) chacun. Cette structure à deux étages s'appelle une structure conjuguée, et représente le *CS* (*Conjugate-Structure*) dans le nom du codec.

3.4- Dissimulation des trames effacées

Une procédure de masquage des erreurs a été incorporée dans le décodeur afin de réduire la dégradation dans le signal vocal reconstitué en raison d'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. 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. Sinon, la trame actuelle est considérée également comme « apériodique ». Une trame effacée hérite sa classe de la trame vocale (reconstituée) précédente. On notera que la classification des éléments voisés est mise à jour en permanence sur la base de ce signal vocal reconstitué[18].

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

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

2) affaiblissement de gains de répertoire codé adaptatif et de répertoire codé fixe.

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

4) production de l'excitation de remplacement.

3.5- Conclusion

Ce chapitre a été consacré à la description générale du codeur standard G.729(CS-ACELP) qui est une norme de codage numérique de la parole qui a été approuvé à l'UIT(Union International de Télécommunication) en Novembre 1995[17]. Cette norme permet de coder la parole avec un débit de 8Kb/s en conservant une qualité de grande fidélité. Ce codeur représente un bon compromis en terme de délai, débit, qualité et robustesse.

Chapitre 4

Résultats et Interprétation

4.1- Introduction

L'objectif de ce chapitre est de présenter les différents algorithmes nécessaires à la simulation de notre codeur G.729. Nous avons implémenté l'algorithme de quantification scalaire Intra-DQ (ou *DSQ : Differential Scalaire Quantization*) à la place de la PSVQ (Inter-trames) adoptée par le codec standard G.729. En outre, la technique de dissimulation interpolative des trames effacées a été introduite afin de réduire les pertes des paquets dans le réseau, ensuite nous avons comparé la distorsion spectrale causée par la dissimulation interpolative des trames effacées. Les résultats expérimentaux des distorsions spectrales *SD(Spectral Distortion)* et distorsions spectrales modifiées *EMBSD*(*Enhanced Modified Bark Spectral Distortion*) sont donnés d'une façon comparative entre le G.729 original et le G.729 modifié(utilisation de la Intra-DQ et la dissimulation interpolative).

Lorsque des paquets de parole sont envoyés en temps réel à travers des réseaux IP, il n'y a aucune garantie de les recevoir dans une manière appropriée, ce qui est dû à la nature "*best effort*" des réseaux. 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: les méthodes dissimulatives basées sur le codeur et celles basées sur le décodeur(figure 4.1).

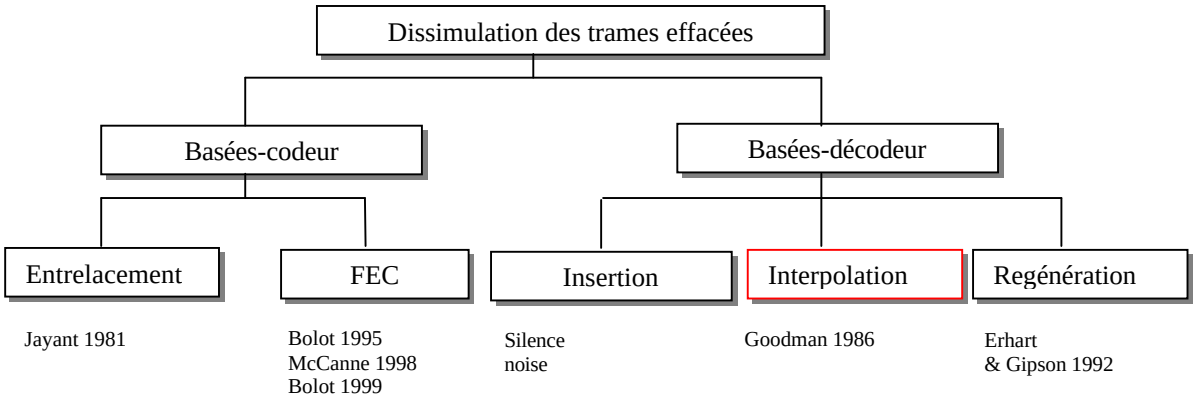


Figure 4.1 : Différentes méthodes implémentées dans la dissimulastion des trames effacées [10]

La dissimulation de l'Erreur par transmission [21] (*FEC*: *Forward Error Concealment*) 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*[11][21][29] 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 dissimulation par le décodeur 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 codeur de l'ITU G.729 possède une procédure de traitement des trames effacées basée sur une méthode de dissimulation prédictive. Cette méthode n'introduit aucun délai supplémentaire, car les paramètres des trames perdues seront récupérés des bonnes trames précédentes. Cependant, ce codeur quantifie les paramètres *LSFs* par une méthode prédictive, donc la procédure de dissimulation utilisée peut causer une propagation de l'erreur aux autres trames. Afin d'améliorer la performance de notre codeur, on a introduit la méthode de dissimulation basée sur le décodeur (interpolative).

4.2- Propagation de l'erreur avec la quantification inter-trame

On sait que la quantification des *LSFs* dans le codec G.729 se fait par la méthode PSVQ, cette dernière met à profit la bonne corrélation qui existe entre les trames successives (corrélation inter-trames) afin d'obtenir une bonne quantification.

Cependant, lors de l'effacement d'une trame, et à cause de la méthode de quantification elle-même (inter-trame), l'erreur commise dans le masquage ne va pas se limiter à la trame elle-même mais se propagera, au moins, aux 3 trames suivantes. La figure (4.2) illustre les distorsions spectrales d'une séquence de parole codée avec et sans effacement de trames. Dans le graphe l'effacement se produit à la trame 8 et à la trame 30, mais on observe que l'erreur se propage aux trames suivantes.

Puisque cette propagation d'erreur est due essentiellement à la façon dont les *LSFs* sont quantifiés (quantification inter-trame), une méthode de quantification agissant sur chaque trame séparément et exploitant la corrélation existante entre chaque élément d'une même trame (quantification intra-trame) empêchera la propagation des erreurs et donnera de résultats meilleurs.

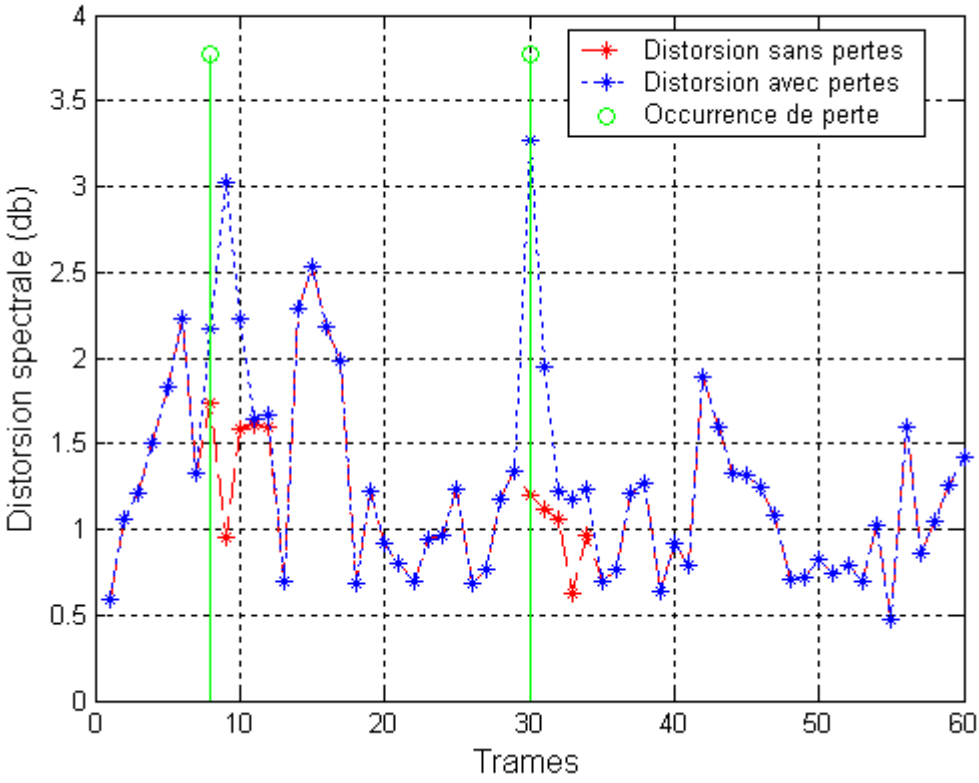


Figure 4.2 : Propagation de l'erreur dans le codage inter-trame (PSVQ)de G.729

4.3- Base de données

Une bonne base de donnée reste la condition sine qua non pour la validation d'un quelconque résultat trouvé. Dans notre travail, nous avons utilisé la base de données *« DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT), Training and Test Data ».*

Le corpus TIMIT, formé de parole lue, a été conçu afin de fournir des données de parole, pour l'acquisition de connaissance acoustique-phonétique et pour le développement et l'évaluation de systèmes automatiques de reconnaissance de parole.

4.3.1- Répartition des locuteurs

630 locuteurs, provenant de régions pratiquant les 8 principaux dialectes des États-Unis, ont participé à l'élaboration du corpus TIMIT. Chacun d'eux lit 10 phrases distinctes, totalisant ainsi, 6300 phrases (table 4.1)

(région)	Male	Femelle	Total
1	31 (63%)	18 (27%)	49 (8%)
2	71 70%)	31 (30%)	102 (16%)
3	79 (67%)	23 (23%)	102 (16%)
4	69 (69%)	31 (31%)	100 (16%)
5	62 (63%)	36 (37%)	98 (16%)
6	30 (65%)	16 (35%)	46 (7%)
7	74 (74%)	26 (26%)	100 (16%)
8	22 (67%)	11 (33%)	33 (5%)
Total	438 (70%)	192 (30%)	630 (100%)

Table 4.1 : Répartition des locuteurs de la base de donnée TIMIT

4.3.2- Test et Training

La base de données a été divisée en 2 parties, une pour l'entraînement (Training) et l'autre pour le test. Les données de test ont un noyau composé de 24 locuteurs : 2 de sexe male et1 de sexe femelle pour chaque région, récitant 192 phrases.

Il est à noter que les séquences de parole de la base de données sont échantillonnées à 16KHz et qu'elles comportent un entête. Donc, pour adapter ces signaux de base aux signaux d'entrés du codec G.729, il a fallu effacer les entêtes, ré-échantillonner à 8KHz, et regrouper les phrases pour faciliter le travail.

La table 4.2 rassemble les caractéristiques :

	Longueurs (Temps)	**Nombre de Trame de 10ms**
Entraînement	3^h08'42 "	1132228
Test complet	1^h08'56"	413623
Noyau du test	0^h09'43 "	58300

Table 4.2. Longueurs de la base de données.

Les données d'entraînement ont été utilisées dans les études statistiques et la conception de tous les dictionnaires, alors que ces mêmes dictionnaires ont été testés par le noyau du test.

4.4-Mesure des distorsions

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

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[22][23]:

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

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 (table 4.3) et conviennent à l'évaluation de la parole dégradée par des erreurs de transmission dans des environnements réels de réseau, tels que des erreurs de bit et des effacements des trames.

Catégories	**qualité de la parole**	**EMBSD Distortion Perceptuelle**
1	mauvaise	8
2	médiocre	6
3	passable	4
4	bonne	2
5	excellente	0

Table 4.3 : Table de conversion temporaire des valeurs du MOS (mesure subjective) à la distorsion perceptuelle (EMBSD)[24]

4.5- Etude statistique

Cette étude statistique comprend des tracés d'histogramme normalisé (figure 4.3), des calculs de l'écart type et de la dynamique (Table 4.4) pour chaque élément du vecteur *LSF*. Enfin une matrice regroupant les coefficients de corrélations linéaires $\bullet(i,j)$ entre chaque paire des éléments du vecteur *LSF* suivant la formule (4.2)

$$r[i,j] = \frac{\mathrm{cov}(LSF[i], LSF[j])}{\sqrt{\mathrm{cov}(LSF[i], LSF[i])} * \sqrt{\mathrm{cov}(LSF[j], LSF[j])}} \qquad (4.2)$$

Où $\mathrm{cov}(LSF[i], LSF[j])$: est la covariance entre les élément i et j du vecteur *LSF* à travers toutes les observations faites.

A noter que les *LSFs* utilisés dans cette étude ont été calculés par le standard ITU G.729 en utilisant les fichiers d'entraînements (Train) de la base de données TIMIT, puis, ont été normalisés par • .

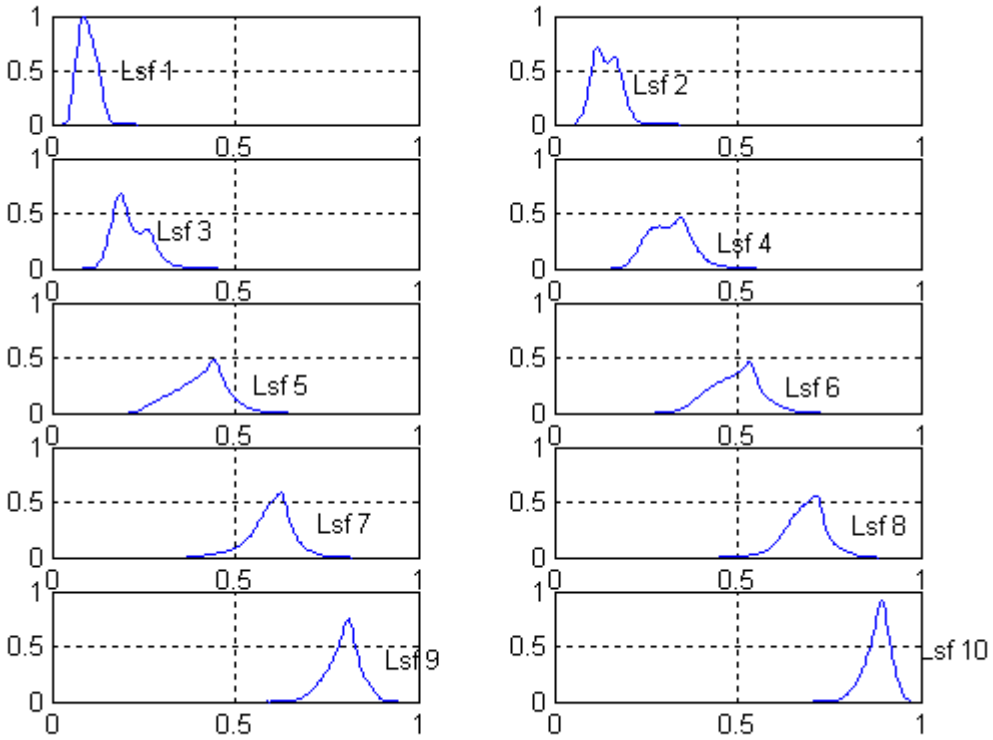


Figure 4.3 : Histogramme normalisé représentant les variations des coefficients *LSFs*

LSF	1	2	3	4	5	6	7	8	9	10
Ecart type *10^{-4}	247	348	487	562	676	656	574	535	435	339
Dynamique *10^{-4}	2174	2954	3752	4038	4472	4605	4592	4385	3598	2748

Table 4.4 : Ecart type et dynamique des coefficients *LSFs*

De la figure 4.3 nous observons que :

- Les tracés des éléments 1, 2, 9 et 10 sont très proches du tracé d'une distribution Gaussienne.
- Les éléments intermédiaires 3, 4 jusqu'à 8 ont une grande dynamique

Pour trouver la partition optimale des vecteurs *LSF*, la corrélation *intra-trame* a été calculée pour les vecteurs 1132228, 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 la table 4.5.

De plus, la matrice de corrélation renvoie d'assez grandes valeurs sur la diagonale au-dessus de la diagonale centrale (surlignée en jaune) indiquant une grande corrélation entre chaque paire d'élément successive.

LSF	1	2	3	4	5	6	7	8	9	10
1	1	0.72363	0.4212	0.18437	0.013871	-0.049523	0.051768	0.077057	0.087349	0.0038356
2	0.72363	1	0.77811	0.41084	0.25659	0.21035	0.27276	0.34433	0.24821	0.20013
3	0.4212	0.77811	1	0.66666	0.41989	0.45923	0.39983	0.49956	0.37673	0.28993
4	0.18437	0.41084	0.66666	1	0.69903	0.54928	0.4342	0.39305	0.4264	0.24968
5	0.013871	0.25659	0.41989	0.69903	1	0.75617	0.51541	0.44706	0.27183	0.27258
6	-0.049523	0.21035	0.45923	0.54928	0.75617	1	0.6986	0.59953	0.40359	0.27466
7	0.051768	0.27276	0.39983	0.4342	0.51541	0.6986	1	0.72727	0.47617	0.39544
8	0.077057	0.34433	0.49956	0.39305	0.44706	0.59953	0.72727	1	0.54996	0.37291
9	0.087349	0.24821	0.37673	0.4264	0.27183	0.40359	0.47617	0.54996	1	0.50533
10	0.0038356	0.20013	0.28993	0.24968	0.27258	0.27466	0.39544	0.37291	0.50533	1

Table 4.5 : La corrélation entre LSF_i et LSF_j de la même trame

Afin de mieux illustrer cette dernière propriété, nous avons calculé les différences entre chaque paire d'éléments du vecteur *LSF*, selon l'équation(4.3) et refait la même étude statistique, c'est-à-dire le tracé d'histogramme normalisé (figure 4.4) et les calculs de l'écart type et de la dynamique(table 4.6).

$$\begin{cases} DLSF[1]=LSF[1] \\ DLSF[i]=LSF[i]-LSF[i-1] \quad \text{pour i=2 à 10} \end{cases} \qquad (4.3)$$

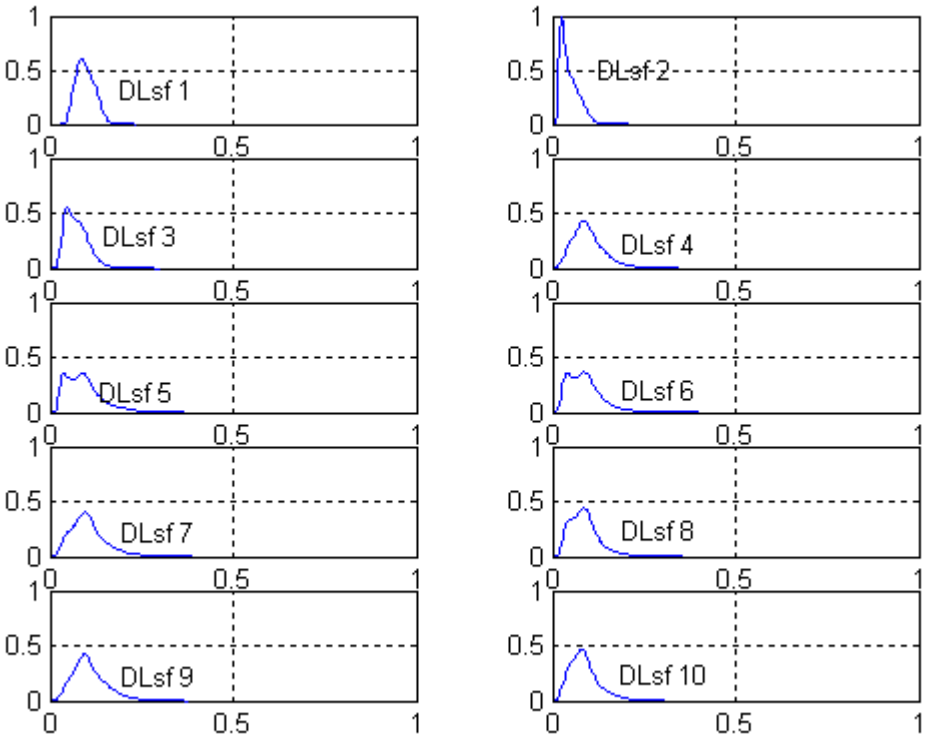


Figure 4.4: Histogramme normalisé des coefficients *DLSF*

LSF	1	2	3	4	5	6	7	8	9	10
Ecart type *10^{-4}	247	240	307	433	491	465	483	411	469	394
Dynamique *10^{-4}	2174	2102	2970	3398	3643	4015	3870	3558	3745	3069

Table 4.6 : Ecart type et dynamique des coefficients DLSF.

On peut conclure que les *DLSFs* ont une dynamique plus petite que celle des *LSFs* et se prêteraient, donc, mieux à la quantification.

La méthode de quantification qu'on a implémenté au G.729 est la quantification différentielle scalaire Intra-DQ des *DLSF* avec la dissimulation interpolative.

4.6- Dissimulation interpolative

Si les futures données de la parole sont disponibles, ou peuvent être générées, alors une approche interpolative pour la dissimulation des trames effacées devient possible. Cela devrait intuitivement produire une meilleure dissimulation que l'approche répétitive simple, en payant un délai supplémentaire[27].

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 futures 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 appliquer la dissimulation Interpolative et améliorer la performance dans le cas des trames effacées, sans aucun coût supplémentaire en terme de délai.

La figure 4.5 illustre l'application de la dissimulation interpolative dans un récepteur VoIP typique.

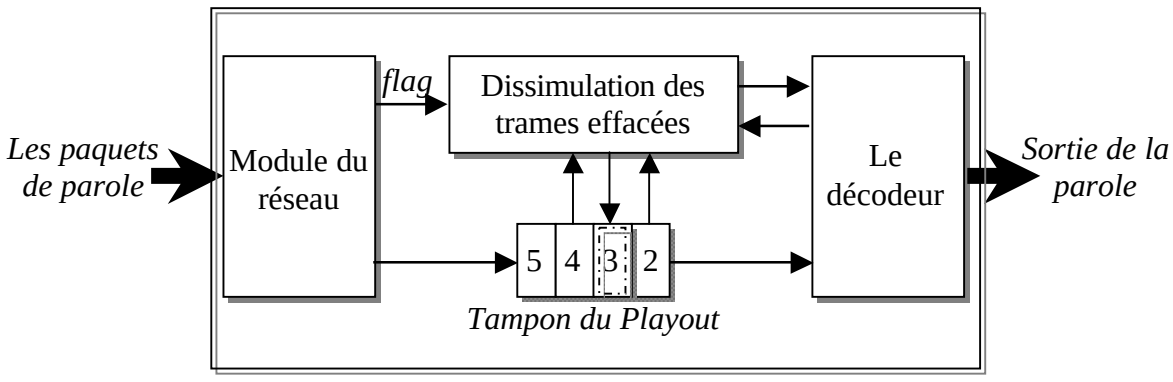


Figure 4.5 : Récepteur VoIP typique : Dissimulation Interpolative[27]

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 de la dissimulation 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).

4.7- Application de la dissimulation Interpolative au G.729

Le schéma de la dissimulation interpolative a été implémenté au codec standard G.729. Ainsi, le décodeur est modifié, alors si une trame est détectée effacée une dissimulation interpolative est appliquée au lieu de la méthode défini par le standard.

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 la dissimulation interpolative au G.729, nous allons chercher une quantification *intra-trame* qui donne des performances égales ou meilleurs que la quantification prédictive (*inter-trame*), utilisée par le G.729.

4.7.1- Espérance de l'erreur quadratique de la dissimulation interpolative

Premièrement, on calcule l'espérance de l'erreur quadratique pour une dissimulation interpolative des *LSFs* 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, $F_n = (f_1, f_2, ..., f_P)$ est le vecteur LSF de la nième trame et $\hat{F}_n$ est le vecteur LSF quantifié ou interpolé correspondant; alors le vecteur LSF perdu interpolé peut être écrit [28]:

$$\hat{F}_{n+x} = \frac{L+x-1}{L+1}\hat{F}_n + \frac{x}{L+1}\hat{F}_{n+L+1} \qquad (4.4)$$

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

$$D_L = \frac{1}{L}\sum_{x=1}^{L}\sum_{p=1}^{P}(f_{n+x,p} - \hat{f}_{n+x,p})^2 \tag{4.5}$$

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

$$\begin{aligned} ED_{\text{int}} = \frac{\Phi(0)}{L}\sum_{x=1}^{L}[1 + \frac{(L+1-x)^2 + x^2}{(L+1)^2} - \frac{2(L+1-x)}{L+1}\mathsf{f}(x) \\ - \frac{2x}{L+1}\mathsf{f}(L+1-x) + \frac{2x(L+1-x)}{(L+1)^2}\mathsf{f}(L+1)] \end{aligned} \tag{4.6}$$

Où $\Phi(.)$ et $\phi(.)$ sont, respectivement, la somme des auto-corrélations, et la somme normalisée des auto-corrélations des vecteurs LSFs.Ils sont définies comme :

$$\begin{aligned} \Phi(\mathsf{t}) &= \sum_{p=1}^{P} E[f_{n,p} f_{n+\mathsf{t},p}] \\ \mathsf{f}(\mathsf{t}) &= \frac{\sum_{p=1}^{P} E[f_{n,p} f_{n+\mathsf{t},p}]}{\sum_{p=1}^{P} E[f_{n,p}^2]} \end{aligned} \tag{4.7}$$

4.7.2- Espérance de l'erreur quadratique de la dissimulation prédictive

Pour la dissimulation prédictive, la trame LSF perdue est récupérée des bonnes trames précédentes reçues codées inter-trame prédictive par un scalaire fixe b et le vecteur *LSF* dissimulé[28].

$$\hat{F}_{n+x} = B^x \hat{F}_n \tag{4.8}$$

Noter que l'erreur de dissimulation peut se propager aux autres trames à plus tard. 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 :

$$\hat{F}_{n+L+1} = \mathsf{b}^{L+1}\hat{F}_n + e_{n+L+1} \tag{4.9}$$

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 :

$$L \times ED_{L,pred} = \Phi(0) \sum_{x=1}^{L} [1 + \beta^{2x} - 2\beta^{x} \rho(x)] \tag{4.10}$$

Pour la partie propagée :

$$D_{prop} = \sum_{p=1}^{P} (f_{n+L+1,p} - \beta^{L+1} f_{n,p} - e_{n+L+1,p})^2 \tag{4.11}$$

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 :

$$ED_{prop} = \Phi(0)[1 + \beta^{2(L+1)} - 2\beta^{L+1} \rho(L+1)] \tag{4.12}$$

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

$$\begin{aligned} ED_{\Pr ed} &= \frac{1}{L+1}(L \times ED_{L,pred} + ED_{prop}) \\ &= \frac{\Phi(0)}{L+1} \sum_{x=1}^{L+1} [1 + \beta^{2x} - 2\beta^{x} \rho(x)] \end{aligned} \tag{4.13}$$

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 :

$$ED_{rep} = \frac{\Phi(0)}{L+1} \sum_{x=1}^{L+1} [2 - 2\rho(x)] \tag{4.14}$$

4.7.3- Quantification *Intra-trame* des LSFs

Nous avons choisit la Intra-DQ (*Quantification différentielle scalaire*)qui a été décrite en [25] comme technique de quantification des *LSFs*.

La procédure de la quantification Intra-DQ est :

1 *. Quantifier LSF(n,1) à* LSF_q *(n,1) ;*

2*. calculer DLSF (n,j)=LSF(n,j) -* LSF_q*(n,j-1) ; pour j=2,10*

3*. quantifier DLSF(n,j) à* $DLSF_q$*(n,j) ;*

4*. si j> 10 stop ; sinon reconstruire LSFq(n,j)=*LSF_q*(n,j-1) +* $DLSF_q$*(n,j), puis aller à 2.*

4.7.4- Allocation de bits

L' allocation de bits par trame pour le G.723.1 d'après[25] est illustrée par la table 4.7

Indice	1	2	3	4	5	6	7	8	9	10	**Total**
Bits	2	3	3	3	3	3	3	3	2	2	**27**

Table 4.7: Allocation de bits pour l'Intra-DQ

Les dix dictionnaires ont été conçus par l'algorithme LBG[26].

4.7.5- Performance du quantificateur

L'Intra-DQ, sera testé par le noyau de test(58300 trames) de la base de données TIMIT. Sachant que la distorsion spectrale moyenne calculée pour les vecteurs de la base de donnée de test pour le codec standard G.729 est :

$$SD_{G.729} = 1.29dB \qquad (4.15)$$

Et que les résultats de cette méthode Intra-DQ ; selon différentes allocations de bits sont donnés par la table 4.8.

Nombre de bits total des *LSFs*	**Allocation de bits pour chaque LSF(LSF_0 - LSF_9)**	**DS.Moy (dB)**	**2–4 dB (%)**	**> 4dB (%)**
27	2-3-3-3-3-3-3-3-2-2	1.44	11.77	0.15
28	3-3-3-3-3-3-3-3-2-2	1.36	9.54	0.13
28	2-3-3-3-3-3-3-3-3-2	1.30	7.47	0.03
29	3-3-3-3-3-3-3-3-3-2	1.22	5.64	0.03

Table 4.8 : Distortion spectrale de l'Intra-DQ

On voit bien d'après les résultats de la table 4.8 que l'allocation de 29 bits par trame donne une distorsion spectrale moyenne meilleure que celle du G.729 original (table 4.9).

	DS.Moy (dB)	**2–4 dB**	**> 4dB**	**EMBSD**
PSVQ	1.29	9.72	0.25	0.753
Intra-DQ	1.22	5.64	0.03	0.748

Table 4.9 : Performance de la méthode Intra-DQ

Les résultats de l'Intra-DQ sont meilleurs que ceux du standard, néanmoins cette méthode requière 29 bits pour la quantification des coefficients LSFs, soit 11 bits de plus par trame par rapport au standard ITU G.729 (18 bits), soit une augmentation de débit de 1,1 Kbits/s.

4.8- Simulation et Résultats

Nous avons trouvé qu'avec une quantification Intra-DQ, on réalise une distorsion minimale pour le codeur G.729, cette quantification qui est 1,1kbits/trame plus que le codage prédictive original du G.729, mais qui présente des propriétés adéquates à l'application de la récupération interpolative des *LSFs*. Nous allons simuler la voix en temps-réel sur des paquets en réseau où chaque paquet contient une trame.

4.8.1- 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 [25][27][32].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 :

$$\begin{aligned} P_1 &= P_0 \cdot p + P_1 \cdot q \\ P_1 + P_0 &= 1 \end{aligned} \tag{4.16}$$

$$\Rightarrow P_0 = \frac{1-q}{p+1-q} \;;\; P_1 = \frac{p}{p+1-q} \tag{4.17}$$

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* (P_1). La figure 4.6 présente le modèle de Markov avec ses probabilités de transition, et la table 4.10 cite les taux de perte utilisés dans notre simulation.

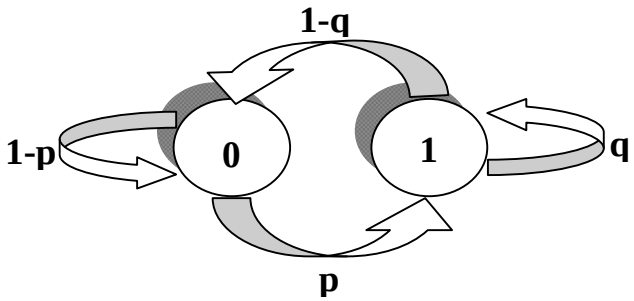


Figure 4.6 : Perte des paquets modélisée par un processus aléatoire de Markov

Taux(%)	p	q
0	0.0	0.00
10	0.1	0.15
20	0.2	0.20
30	0.3	0.35
40	0.4	0.40
50	0.5	0.55

Table 4.10 : Les taux de Pertes simulés.

4.8.2- Procédure de dissimulation implémentée

Le processus complet de dissimulation peut être résumé comme suit [27] :

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 (*pitch lag*);
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 adaptative à zéro, -Utiliser l'information précédente du gain, -Remplacer les signaux d'excitation par une séquence de nombres aléatoires normalisée par le gain atténué.

4.8.3 Résultats

La figure 4.7 montre les performances de la méthode par interpolation appliquée aux paramètres LSF quantifiés avec Intra-DQ, 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 la table 4.11.

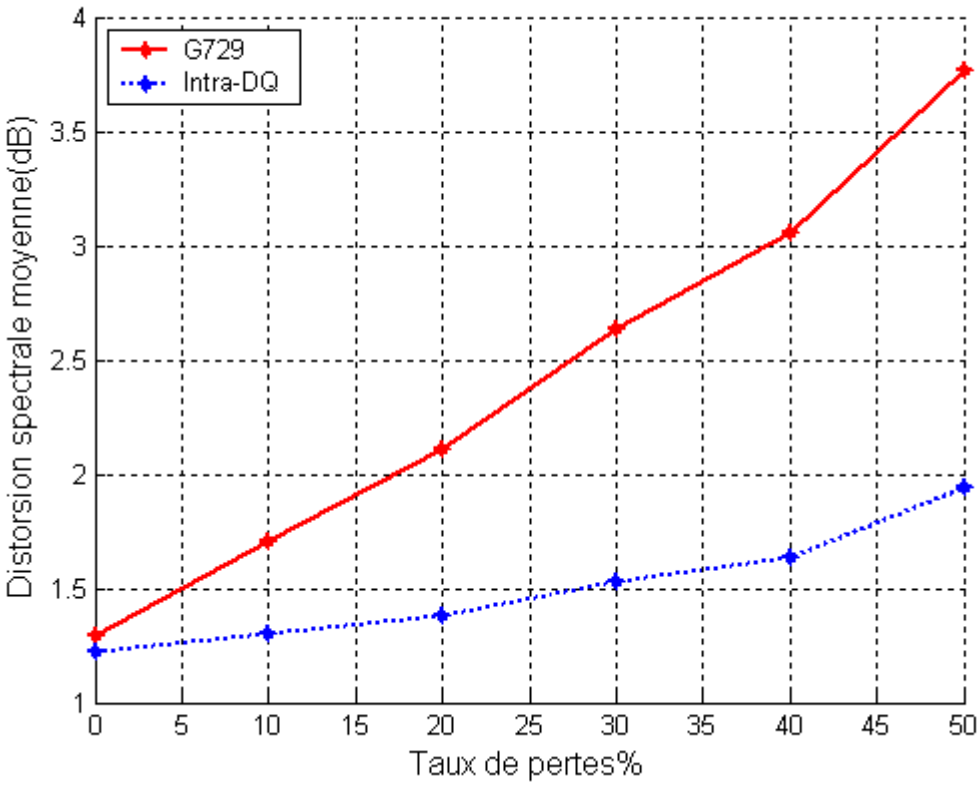


Figure 4.7 : Distorsion spectrale moyenne avec des trames effacées

On voit bien que, avec un débit de 1,1 kbit/s de plus que le standard G.729. Notre méthode de quantification et de récupération des *LSFs* ainsi appliquée, réalise 0.07 à 1.83 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.

Trames perdues (%)	G729			Intra-DQ		
	SD_{moy} (dB)	*Outliers* (%)		SD_{moy} (dB)	*Outliers* (%)	
		2-4 dB	> 4dB		2-4 dB	> 4dB
0	1.29	9.72	0.25	1.22	5.64	0.03
10	1.71	23.30	3.55	1.30	8.35	0.53
20	2.11	33.87	7.65	1.38	10.97	1.14
30	2.64	41.56	15.39	1.53	15.36	2.54
40	3.06	46.42	21.96	1.64	18.67	3.70
50	3.77	45.71	35.09	1.94	25.01	7.51

Table.4.11 : Distorsion spectrale moyenne et *Outliers* avec des trames effacées

La distribution, des distorsions spectrales, représentée sur la figure 4.8, montre que la plupart des trames perdues ont des petites distorsions, avec la quantification Intra-DQ(intra-trame) en la comparant avec la quantification PSVQ (inter-trame) du G.729.

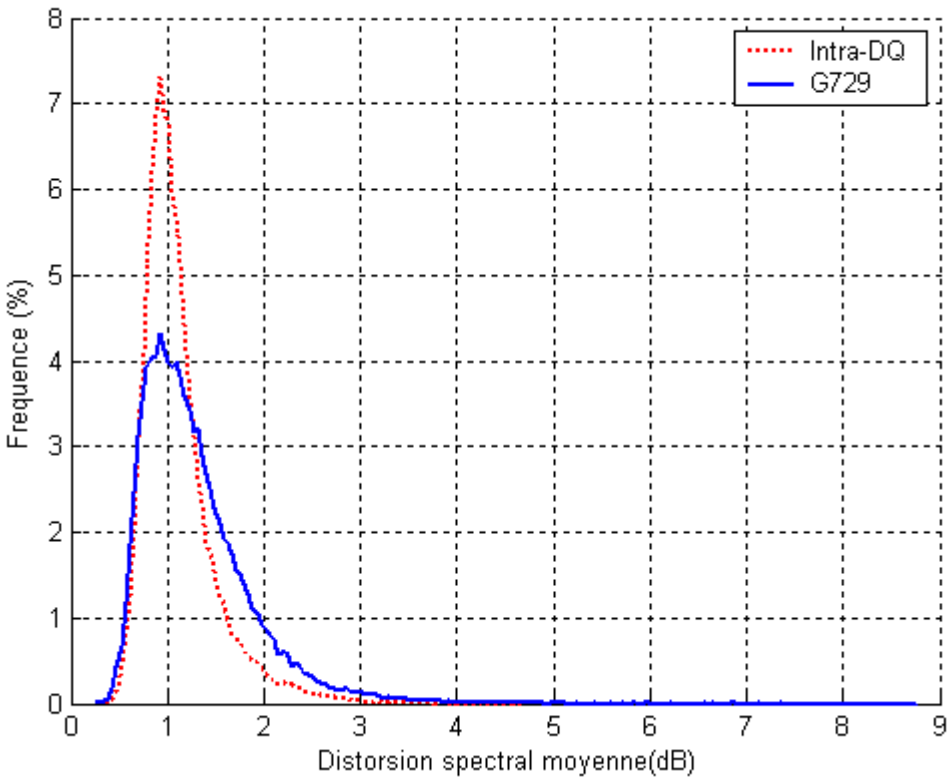


Figure 4.8 : Histogrammes des Distorsions Spectrales (SD)

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 4.9 montre, pour une nouvelle fois, les performances de la dissimulation interpolative appliquée aux paramètres *LSFs* quantifiés avec Intra-DQ, comparées à la méthode prédictive adoptée par le G.729.

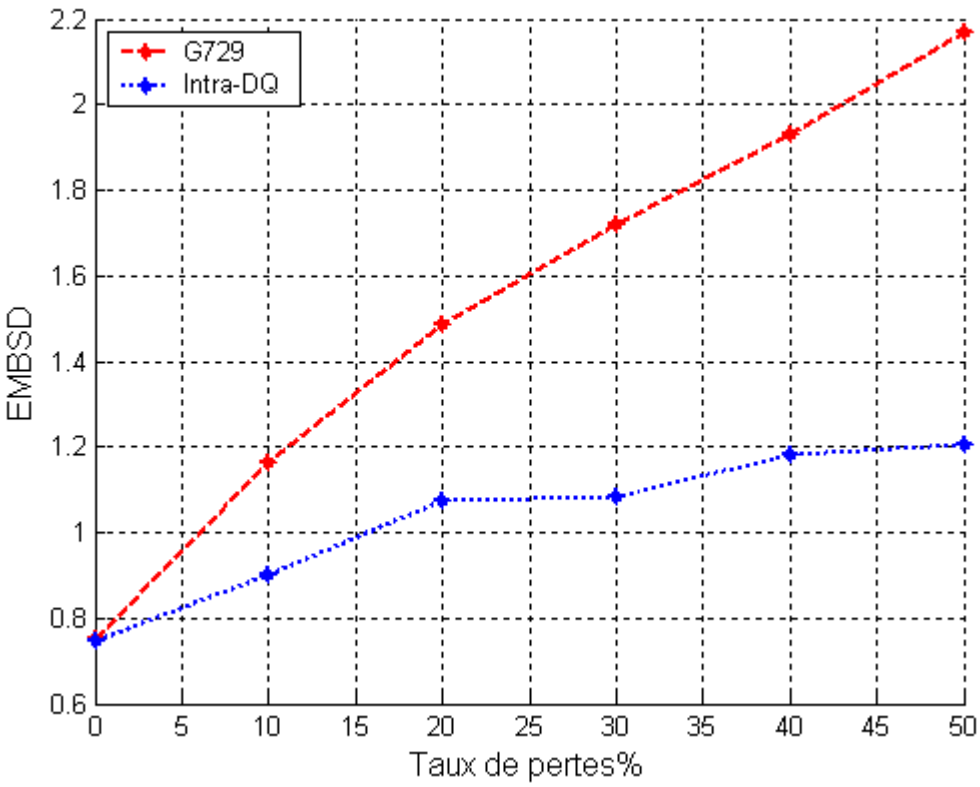


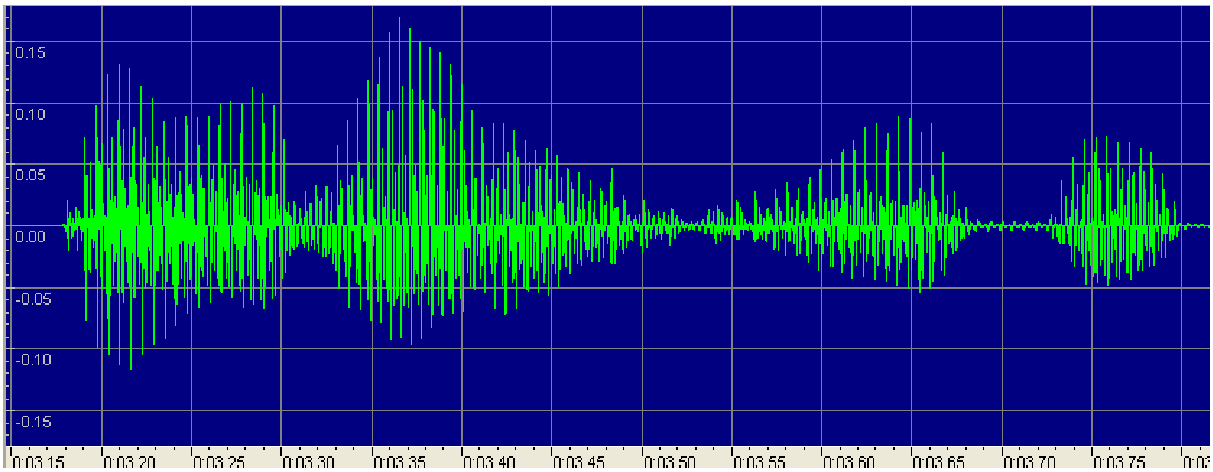
Figure 4.9 EMBSD avec des trames effacées

On voit bien, pour une deuxième fois, qu'avec un débit de 1.1 kb/sec de plus par rapport au standard. Notre méthode de quantification et de récupération des *LSFs* ainsi appliquée, réalise jusqu'à 0.962 de moins de distorsion perceptuelle EMBSD, comparée à la méthode adoptée par le standard G.729. Et si on compare les résultats de la figure 4.9 avec ceux de la table 4.3 on peut conclure que notre méthode donne une qualité «*bonne*» jusqu'au 50% de taux de perte ce qui correspond à une qualité « *passable* » pour le G.729 original.

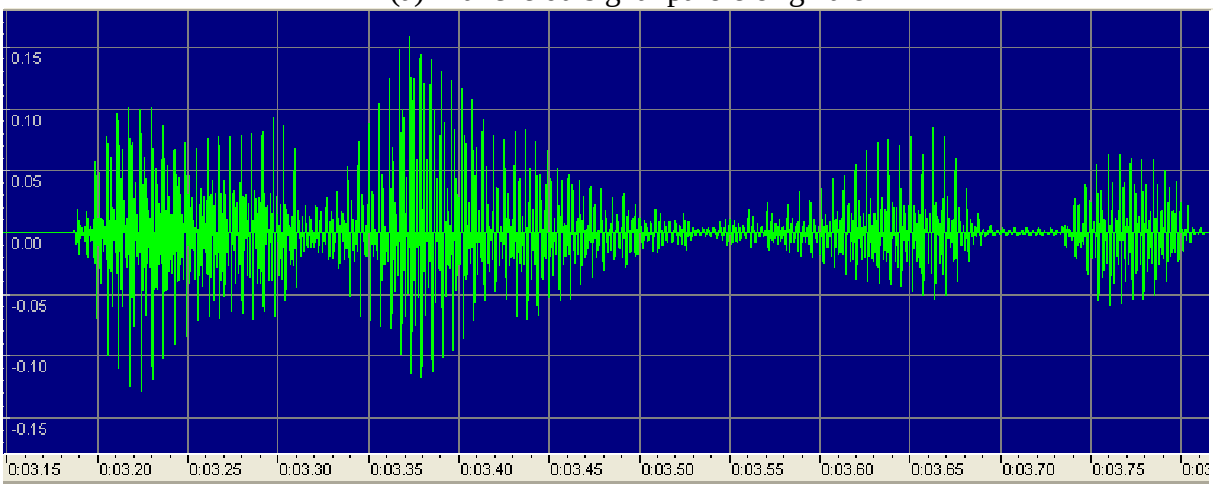
Il est à noter que le délai total d'interpolation est la multiplication des délais des trames effacées. Si, par exemple, on a trois trames effacées, alors le délai sera 3 * 10ms+5ms + RTT/2 où RTT (*Round Trip Time*) est le temps moyen de l'aller-retour des paquets sur le réseau, généralement compris entre 10 et 700 ms pour un réseau typique[28]. 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.

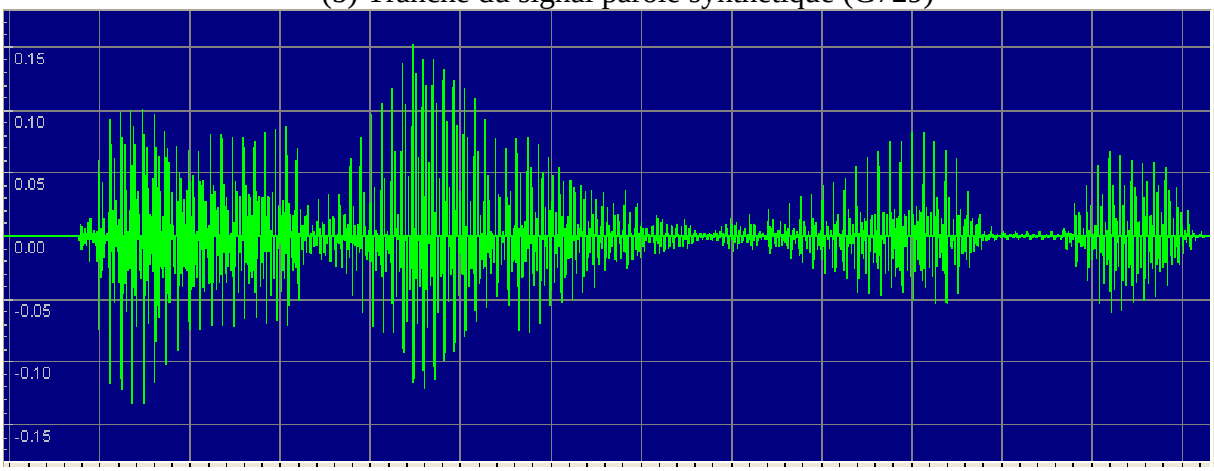
La figure 4.10 nous illustre les audiogrammes des tranches de parole "noyau de test"de la base de données TIMIT. Le logiciel de son utilisé est le GoldWave.



(a) Tranche du signal parole originale



(b) Tranche du signal parole synthétique (G729)



(c) Tranche du signal parole synthétique (Intra-DQ :29 bits)

Figure 4.10: Audiogrammes des signaux de parole

4.9- Conclusion

Avec la quantification intra-DQ des *LSFs*, 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 montrent (tests subjectifs simples) que l'application de la quantification *intra-trame*, et de la récupération des *LSFS* par interpolation (*Interpolative Concealment*), améliorent considérablement la qualité des trames de parole effacées.

L'inconvénient de cette méthode peut être le délai supplémentaire requis pour l'interpolation, et probablement l'extra débit de 1.1Kb/s (13,7% du débit total) ce qui équivaut à un débit de 9.1Kb/s pour la quantification intra-DQ par rapport au G.729 original (8Kb/s).

Conclusion

Après avoir constaté que, lors de pertes de trames, la quantification prédictive des coefficients *LSFS* (PSVQ, Inter-trame) utilisée par le codecG.729, engendre une erreur qui ne se limite pas à la trame perdue mais qui se propage systématiquement aux trames suivantes, affectant ainsi, les performances globales du codec; nous avons essayé d'implémenter une nouvelle méthodes de quantification qui traitent chaque trame séparément et qui exploitent les corrélations Intra-trame, c'est-à-dire la corrélation entre chaque élément d'une même trame, espérant de la sorte, enrayer la propagation d'erreur.

La première étape fut la réalisation d'une petite étude statistique sur les données à quantifier, c'est-à-dire, les coefficients *LSFs*. Afin d'exploiter la corrélation Intra-trame.

La deuxième étape consistait à quantifier les différences *DLSFs* par le biais d'un quantificateur scalaire non uniforme[25]. Cette méthode a renvoyé de bon résultat en égalant la PSVQ dans des conditions de pertes nulles et, surtout, en confirmant la non propagation des erreurs lors de pertes de trames en implémentant l'algorithme de la dissimulation interpolative.

L'inconvénient majeur de la Intra-DQ est l'augmentation du débit de 1,1 Kbits/s. Pour remédier à ce problème on a pensé à la Split-VQ qui a été déjà implémentée par d'autre.

Vu les caractéristiques du G729, on a remarqué d'après les résultats obtenus, que cette norme est meilleure que celle du G.723.1. Ainsi, en implémentant la méthode Intra-DQ avec disimulation interpolative nous avons enrayé la propagation d'erreurs et donc améliorer la qualité du codec G729.

Perspectives futures:

- Confirmer les résultats mathématiques par des tests d'écoute.
- Améliorer les performances de l'interpolation des sous-trames.
- Implémentation du codec G.729 sur un chip (DSP).

Annexe A

Algorithme de Levinson-Durbin :

Les coefficients d'autocorrelation R(k), k=0,1,……..,P sont utilisés pour obtenir les coefficients du filtre *LP* après résolution du sysème 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 récurssion 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 $\bullet_p$ 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:

$$a_m(0) = 1, \qquad (m = 1,2,....., p) \qquad E_0 = R(0) = \sigma_x^2$$

récurssion:

pour: $m = 1,2,...p.$

$$k_m = -\frac{1}{E_{m-1}}\left[R(m) - \sum_{k=1}^{m-1} a_{m-1}(k)R(m-k)\right] \tag{A.2}$$

pour k=1,2,…m-1.

$$a_k(m) = a_k(m-1) - k_m a_{m-k}(m-1) \tag{A.3}$$

$$E_m = E_{m-1}(1-k_m^2) \tag{A.4}$$

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.

Bibliographie

[1] M. Xie et D.Berkani. **"Amélioration des performances des codeurs de parole"** AJOT N°10, 1997.

[2] James H. Y. Loo,**"Intraframe and Interframe Coding of Speech Spectral Parameters"** Department of Electrical Engineering McGill University Montreal, Canada September 1996.

[3] Tamanna Islam; **"Interpolation of Linear Prediction Coefficients for Speech Coding",** Department of Electrical Engineering McGill University Montreal, Canada April 2000.

[4] Nadim Batri, **"Robust Spectral Parameter Coding in Speech Processing"**, Department of Electrical Engineering, McGill University, Montreal, Canada, May 1998.

[5] F. Merazka, D. Berkani, **"Robust Split Vector Quantization of LSP Parameters at Low Bit Rates"** AJSE journal. April 2004; Vol. 29, Number 1B.

[6] F. Merazka, D. Berkani, **"Vector Quantization of LSP Parameters by Split"**, SSST'98, 30th IEEE Southeastern Symposium on System Theory, Morgantown, pp.334 -337. West Virginia, USA. March 1998.

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

[8] M.R Shroeder and B.S. Atal , **"Code Excited Linear Prediction (CELP) high quality speech at very low bit rates",** Proc.Int.conf.Acoust. Speech and signal processing. pp.937-940,1985.

[9] W.B.Kleijn, **"Continuous Representation in Linear Predictive coding",** Proc.Int.conf.Acoust. Speech and signal processing. pp.201-204,1991.

[10] Jean Chiappini, **"Performances de la VoIP sur réseaux wireless"**, Haute École spécialisée Suisse occidentale, 2000.

[11] C. Perkins, O. Hodson and V; Hardman **"A survey of Packet Loss Recovery Techniques for Streaming Audio"**. 1998 IEEE Network.

[12] D. J; Goodman et al., **"Waveform Substitution techniques for recovering missing speech segments in packets voice communications"**, IEEE Trans. Acoustic, Speech, and Sig. Processing, Vol. ASSP-34, no. 6, Dec. 1986, pp. 1440-48.

[13] W. Pariera, **"Modifying LPC parameter dynamics to improve speech coder efficiency,"** Thesis, Departement of Electrical & Computer Engineering, McGill University, Montreal, Canada, Sep 2001.

[14] Yvan Calas**,"Performances des codes correcteurs d'erreur au niveau applicatif dans les réseaux"**,Thèse de Doctorat, Université des Sciences et Techniques du Languedoc, Montpellier II,Décembre 2003.

[15] Henning Sanneck and Nguyen Tuong Long Le,**"Speech Property-Based FEC for Internet Telephony"**, Abstract , San Jose, CA, January 2000.

[16] G.729 **"Codage de la parole à 8 kbit/s par prédiction linéaire avec excitation par séquences codées à structure algébrique conjuguée"**,ITU-1996.

[17] Romain Trilling **"Codage Large Bande De La Parole Par Encapsulation Du Codeur ITU G.729 (CS-ACELP)"** mémoire de maîtrise en sciences appliquées Spécialité : génie informatique Sherbrook(Québec), Canada –Août 1998.

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

[19] F. Itakura,**"Line spectrum representation of linear prediction coefficients of speech signals",** Journal Acoustical Society America, vol. 57, p. 535, 1975. (Abstract).

[20] 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, Dallas , TX USA, 2001.

[21] M. Podolsky, C. Romer and S. McCanne, **"Simulation of FEC-based error control for packet audio on the Internet"** Proceedings - IEEE INFOCOM, vol. 2, pp. 505-515, April 1998.

[22] W. Yang, K. Krishnamachari and R. Yantorno, **"Improvement of the MBSD objective speech quality measure using TDMA data"** Submitted to IEEE Speech Coding Workshop, 1999.

[23] W. Yang, M. Benbouchta and R. Yantorno, **"Performance of the modified bark spectral distortion as an objective speech quality measure".** ICASSP, vol. 1, pp. 541-544, Seattle,1998.

[24] Mohammad M. A. Khan **"Coding of Excitation Signals In a Waveform Interpolation Speech Coder"** thesis department of electrical & computer engineering, McGill University Montreal, Canada, sep 2001.

[25] J.Wang and J.D.Gibson, **"Performance comparison of intraframe and interframe LSP quantization in packet network ".** Proc. 2000 IEEE Workshop on speech Coding, Delevan, WI, USA, Septembre 2000.

[26] LBG algorithm: Y. Linde, A. Buzo and R. M. Gray, **"An algorithm for vector quantizer design"**, IEEE Trans. Comm. Vol. 20, pp. 84-95, 1980.

[27] Juan Carlos De Martin, Takahiro Unno and Vishu Viswanathan **"Improved Frame Erasure Concealment For CELP-Based Coders"** DSPS R&D, Texas Instruments, Dalas, Texas.

[28] 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, Dallas , TX 75275.

[29] Ejaz Mahfuz **"Packet Loss Concealment for Voice Transmission over IP Networks"**. A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Master of Engineering- Canada September 2001.

[30] ITU, G.723.1 : **"Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbits/s"**, ITU 1996.

[31] F. Merazka, D. Berkani **"Split VQ and Predictive Split VQ of LSP Parameters In Packet Networks"** 10th the IEEE Digital Signal Processing Workshop 2nd Signal Processing Education Workshop *October 13-16, 2002*. Georgia, USA.

[32] F. Merazka, D. Berkani, **" Performance Comparison of Split VQ and two-stage VQ-Lattice VQ of LSP parameters In Packet Networks"** 46th IEEE International Midwest Symposium on circuits and systems MWSCAS, Cairo, Egypt, 27-30 Dec. 2003.