REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

Ministère de l'Enseignement Supérieur et
de la Recherche Scientifique
Ecole nationale Polytechnique

وزارة التعليـــــــم العالـــــــــي
و البحث العلمـــــي
المدرسة الوطنية المتعددة التقنيات

Département d'Electronique

Projet de fin d'étude

Thème :

Implémentation d'une méthode de récupération des trames perdues basée sur la répétition en vu de l'amélioration des performances du codec G.729

Proposé et dirigé par :
M^elle^ : F.MERAZKA

Étudié par:
- *RAHIL Ali*
- *SEKFANE Mourad*

Promotion 2005

Laboratoire signal et communication
E.N.P. 10, Avenue Hassen-Badi, El Harrach, ALGER

REMERCIEMENTS

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

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

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

Nous tenons à remercier tous nos amis et camarades pour toute leur sincère amitié le long de cinq années d'études.

DEDICACE

A ma très chère mère Malika

A mon très cher père Athmane

A mon frère Lahcene

A mes sœurs Nabila, Lmia et Siheme

A toute la famille

A tous mes amis

A VOUS TOUS, *MERCI* !

Mourad

DEDICACE

A ma très chère mère Malika

A mon très cher père Mohamed

A mes frères Mourad, M'hamed, Abdelghani et Amine

A mes sœurs Hayet et Amel

A toute la famille, spécialement Mahdi

A tous mes amis

A VOUS TOUS, *MERCI* !

Ali

ملخص

لوحظ في الرّامزة النموذجية ج.729 المقدمة و المعدّلة من طرف الاتحاد الدولي للاتصالات عن بعد، أن بعد ضياع قطع من الكلام، الرّامزة النموذجية ج.729 تصحح هذه القطع الضائعة لكن قيمة المكاسب تصغر(مكسب القاموس التكيفي ومكسب القاموس الثابت). لكن، حتى بعد تصحيح القطعة الضائعة، القطع التالية تتلف بسبب استعمال المكاسب المصغرة .
لقد درسنا و جربنا في هذا البحث الطرق التي تصحح هذا التلف، لقد استعملنا طرقا أساسها تكرار بارا متر قطع الكلام الصحيحة التي وصلت (تصحيح أساسه التكرار وتصحيح أساسه استكمال التنبيه).
الميزة الأساسية لهذه الطرق هي عدم الحاجة لأي وقت زائد.

مفاتيح الكلمات
ترميز الكلام، الكلام عبر شبكة انترنت، إخفاء فقدان القطع، استكمال، تنبيه، تكرار .

RESUME

Dans le codec G.729 de l'ITU (International Telecom Union), nous avons observé qu'après un effacement de trame, la dissimulation standard du codec dissimule les trames perdues avec une atténuation des gains (le gain du dictionnaire adaptatif $g_p^{(n)}$ et le gain du dictionnaire fixe $g_c^{(n)}$). Mais, même après la correction de la trame effacée est fini, les trames suivantes seront détériorés à cause de l'utilisation d'une version atténuée des gains.
Nous avons étudié et tester des méthodes qui corrige cette détérioration, nous avons utilisé des méthodes qui sont basées sur la répétition des paramètres des bonnes trames reçues (Dissimulation basée sur la répétition et Interpolation de l'excitation).
L'avantage majeur de ces méthodes est qu'on n'utilise aucun délai supplémentaire.

Mots clefs :
Codage de la parole, Voix sur IP, gain du dictionnaire adaptatif, gain du dictionnaire fixe, masquage des pertes, interpolation, excitation, répétition.

ABSTRACT

In the codec G.729 of ITU (International Telecom Union), we have observe that after a frame erasure the standard concealment of the codec conceals the lost frames with attenuation of codebook gain (Adaptive Codebook Gain and Fixed Codebook Gain). However, even further the frame erasure is over, the speech signal is further decayed in the subsequent frames. This is because the adaptive codebook is updated with the attenuating excitation signal so the attenuation propagates to the subsequent frames.
We have studied and tested methods that are based on repetition of the parameters of the good frames received on the bad frames (Repetition Based Concealment and Excitation interpolation).
The advantage of these methods of not introducing any extra delay.

Key words:
Speech coding, Voice over IP, adaptive codebook gain, fixed codebook gain, packet loss concealment, interpolation, excitation, repetition.

Sommaire

Liste des figures

Liste des tables

Lexique

ACBK	Adaptatif Code BooK
ADPCM	Adaptive Differential Pulse Code Modulation.
AR	Auto-Regressif.
ARMA	Auto-Regressif Moving Average
CELP	Code Excited Linear Prediction.
CS-ACELP	Conjugate Structure Algebraic Code Excited Linear Prediction
DAM	Diagnostic Acceptability Measure
DPCM	Differential Pulse Code Modulation.
DRT	Diagnostic Rhyme Test
EMBSD	Enhanced Modified Bark spectral Distortion
FCBK	Fixed Code BooK
FEC	Forward Error Correction.
IP	Internet Protocol.
ITU	International Telecommunication Union
LP	Linear Prediction.
LPC	Linear Prediction Coding.
LSP	Line Spectrum Pairs.
MIPS	Million d'operations par seconde
MOS	Mean Opinion Score
PCM	Pulse Code Modulation.
PESQ	Perceptual evaluation of Speech Qualité
PLC	Packet Loss Concealment
RTP	Real time Protocol
RTCP	Real time Control Protocol
SNR	Signal to Noise Ratio
SD	Spectral Distortion.
RSB	Rapport Signal sur Bruit
RSBseg	Rapport Signal sur Bruit segmenté
SQ	Scalar Quantization.
SVQ	Split Vector Quantization.
VoIP	Voice cover IP network.
VQ	Vector Quantization.

Tableau I Table des Abréviations

Introduction

Un changement fondamental et radical s'est produit dans le domaine des télécommunications, à savoir l'émergence de la transmission par paquets tels que *IP* (Internet Protocol).

Les réseaux locaux informatiques dits *LAN* (Local Area Network), basés en grande partie sur les protocoles TCP/IP (Transmission Control Protocol), sont à l'origine de cette évolution de la commutation par paquets, et plus particulièrement de la transmission par le protocole *IP*. Le protocole Internet est reconnu comme le système de transport commun pour les réseaux du futur. L'Internet public avec ses messages électroniques et son service d'information sur la toile mondiale est devenu partie intégrale de la vie de tous les jours.

Ainsi, ces dernières années, le trafic des données a augmenté d'avantage comparé à celui du trafic téléphonique. Avec l'explosion du service Internet, il est devenu manifestement intéressant d'acheminer la parole sur le réseau de données, ce qui est contraire à la tradition établie, qui consiste à acheminer les données sur le réseau téléphonique.

La voix sur *IP* fait appel à deux notions : le *réseau IP* et la voix véhiculée traditionnellement par le *réseau RTC* (Réseau Téléphonique Commuté).

Le codage et le décodage de la parole sont très complexes. Ils peuvent être menés avec un matériel économique. Les deux codecs appelés *"frame-based"* le G.723.1 et le G.729 sont très convenables pour la *VoIP* (Voice over Internet Protocol), car ils fournissent une qualité téléphonique de la parole avec des faibles débits binaires

Les réseaux *IP* étant du type *"Best-Effort"*, il n'y a donc aucune garantie quant à la réception des paquets envoyés.

Les G.723.1 et le G.729 sont basés sur le codage prédictif, 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.

De plus, tel que la méthode de masquage d'erreurs en cas de pertes est implémentée au niveau du G.729, on constate que les paramètres reconstitués ne donnent pas de bons résultats.

> L'objectif de notre travail est d'implémenter de nouvelles techniques de dissimulation des trames perdues dans le réseau qui sont basées sur la répétition des paramètres (Repetition-Based Concealment), pour améliorer les performances de ce codec.

Nous avons organisés notre travail en quatre chapitres :

Le premier chapitre est consacré au codage de la parole : la prédiction linéaire, le modèle de production de la parole humaine et sa distorsion.

Le deuxième chapitre est consacré à la voix sur *IP* (*VoIP*), description du réseau IP et de ces caractéristiques qui sont liées à la transmission de la voix, et les différentes méthodes de recouvrement de pertes.

Le troisième chapitre décrit la norme G.729, le codeur, le décodeur.

Le quatrième chapitre regroupe l'étude des méthodes que nous avons implémentées, les simulations réalisées et l'interprétation des résultats obtenus. En fin une conclusion générale.

Chapitre I

Codage de la parole

Introduction

Le traitement de la parole est aujourd'hui une composante fondamentale des sciences de l'ingénieur. Située au croisement du traitement du signal numérique et du traitement du langage (c'est-à-dire du traitement de données symboliques), cette discipline scientifique a connue depuis les années 60 une expansion fulgurante, liée au développement des moyens et des techniques de télécommunications.

Ce chapitre regroupe des généralités sur les notions fondamentales de la production du signal parole, ses propriétés ainsi que sa perception. Cet aspect est utile à la bonne compréhension de l'évolution des techniques de codage de la parole.

I.1 Le Signal vocal

La parole peut être décrite comme étant le résultat de l'action volontaire et coordonnée d'un certain nombre d'organes. Cette action se déroule sous le contrôle du système nerveux central qui reçoit en permanence des informations par rétroaction auditive et par les sensations kinesthésiques [4].

I.1.1 Mécanisme de phonation

Les principaux organes composant l'appareil phonatoire sont [1]: les poumons, la trachée artère, le pharynx, les cavités buccales et nasales qui sont schématisés par la Figure I.1.

L'appareil respiratoire fournit l'énergie nécessaire à la production de sons, en poussant de l'air à travers la trachée-artère. Au sommet de celle-ci se trouve le ***larynx*** où la pression de l'air est modulée avant d'être appliquée au conduit vocal. Le larynx est un ensemble de muscles et de cartilages mobiles qui entoure une cavité située à la partie supérieure de la trachée.

Les ***cordes vocales*** sont en fait deux lèvres symétriques placées en travers du larynx. Ces lèvres peuvent fermer complètement le larynx, et en s'écartant progressivement, déterminer une ouverture triangulaire appelée ***glotte***. L'air y passe librement pendant la respiration et la voix chuchotée, ainsi que pendant la phonation des sons non voisés.

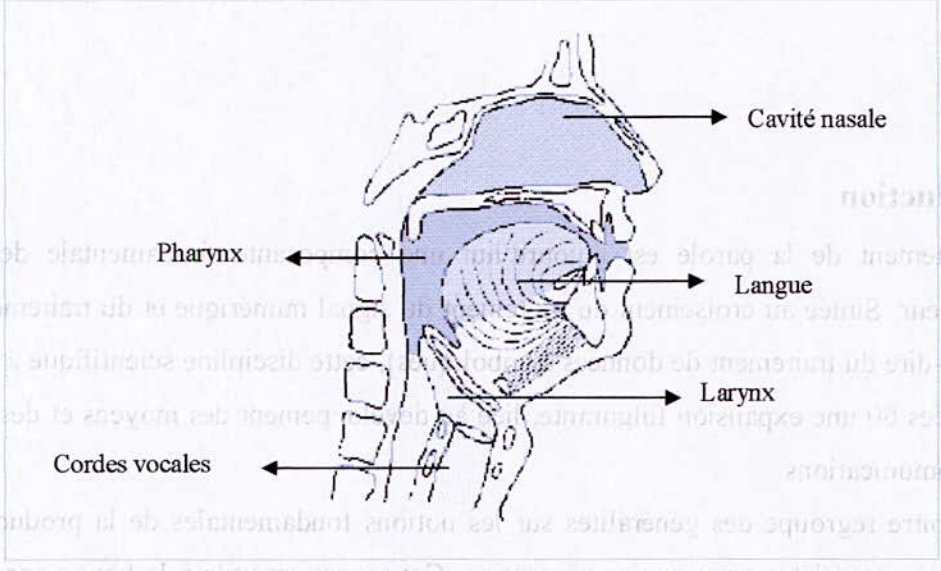

Fig. I.1 Appareil phonatoire

Les sons voisés résultent, au contraire, d'une vibration périodique des cordes vocales. Le larynx est d'abord complètement fermé, ce qui accroît la pression en amont des cordes vocales et force ces dernières à s'ouvrir, ce qui fait tomber la pression en permettant aux cordes vocales de se refermer. Des impulsions périodiques de pression sont ainsi appliquées au conduit vocal composés des cavités pharyngienne et buccale pour la plupart des sons. Lorsque la ***luette*** est en position basse, la cavité nasale vient s'y ajouter en dérivation. Notons pour terminer le rôle prépondérant de la langue dans le processus phonatoire. Sa hauteur détermine la hauteur du pharynx : plus la langue est basse, plus le pharynx est court. Elle détermine aussi le ***lieu d'articulation***, région de rétrécissement maximal du canal buccal, ainsi que l'aperture qui représente l'écartement des organes au point d'articulation. L'intensité du son émis est liée à la pression de l'air en amont du larynx. Sa hauteur est fixée par la fréquence de vibration des cordes vocales, appelée fréquence du fondamental ou pitch. La fréquence du fondamental peut varier [2][3]

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

Un ***son voisé*** est un signal quasi périodique dont le spectre est tracé à la Figure I.2. On y observe les raies qui correspondent aux harmoniques du fondamentale F_0 (pitch).

L'enveloppe de ces raies présente des maximums appelés ***formants*** et qui correspondent aux fréquences propres F_i du conduit vocal (structure formantique). Les trois premiers formants sont essentiels pour caractériser le spectre vocal; les formants d'ordre supérieur ont une influence plus limitée.

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 comme le montre la Figure I.3; son spectre ne présente donc pas de structure de pitch.

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

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

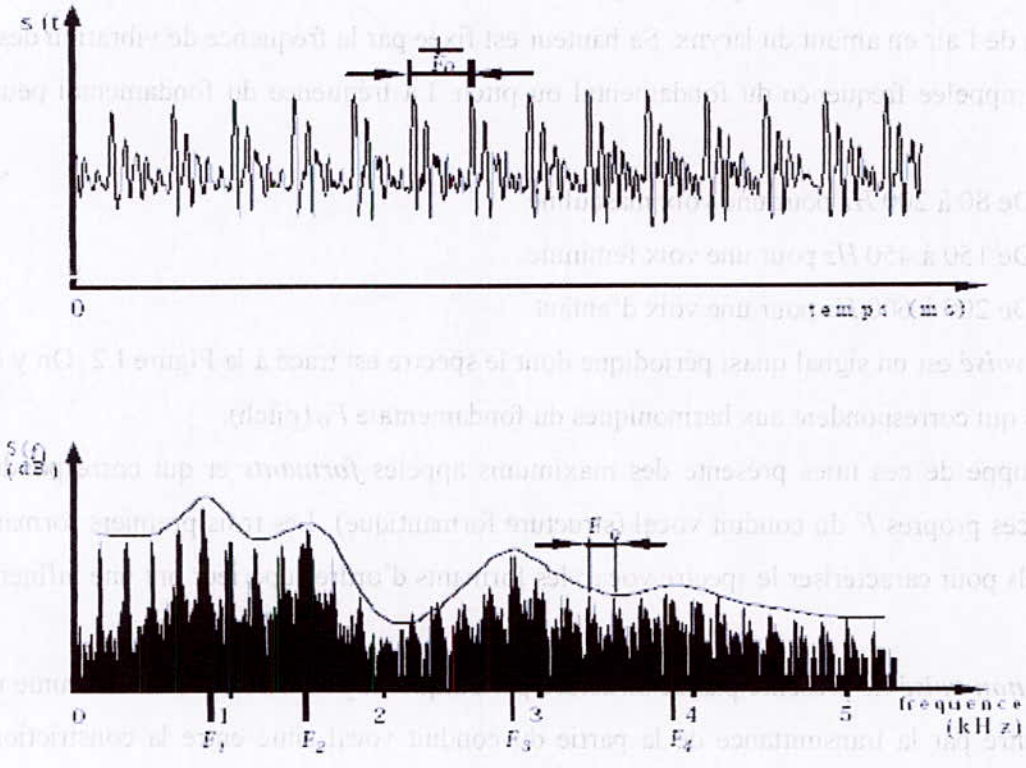

Fig. I.2 Un signal vocal voisé et son spectre [3][4]

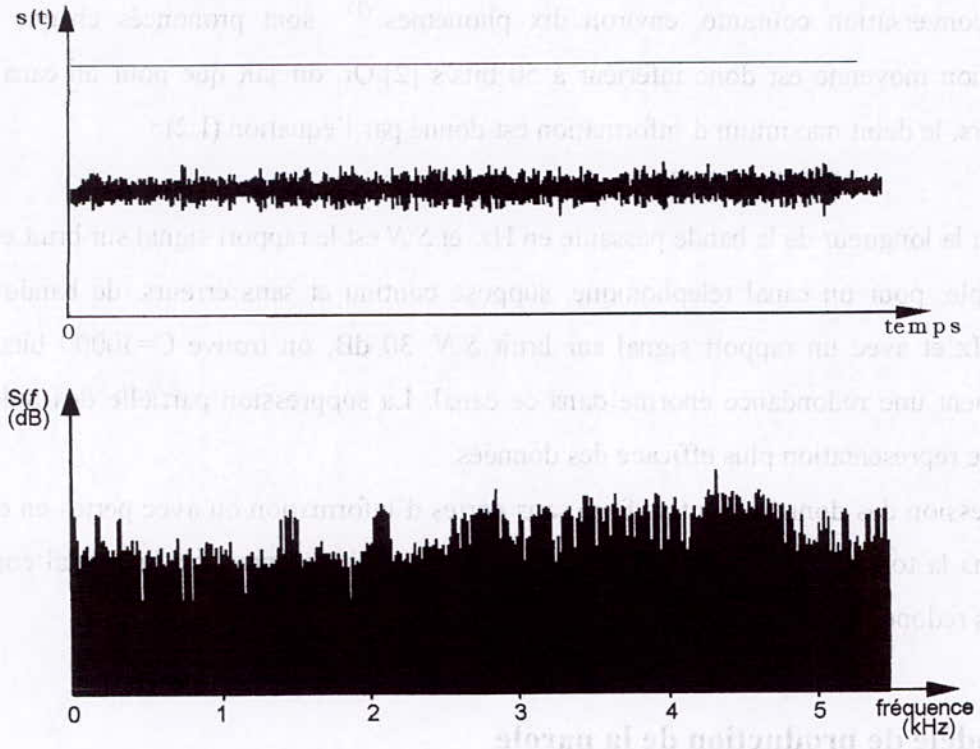

Fig. I.3 Un signal vocal non voisé et son spectre [3][4]

I.1.2 La redondance du signal vocal

Telle que définie par Shannon, la redondance est la partie du signal parole qui, si elle est éliminée, n'affecte pas le contenu du message ou du signal information.

Le signal vocal est caractérisé par une très grande redondance, condition nécessaire pour résister aux perturbations du milieu ambiant, cette redondance sera mise à profit par les techniques de codage de la parole, pour réduire le débit binaire nécessaire au stockage ou à la transmission de la parole, sans, pour autant nuire à son intelligibilité.

On définie l'information associée à un message constitué par des éléments discrets x_i appartenant à un ensemble donné X, et si $p(x_i)$ est la probabilité a priori d'occurrence du symbole x_i , on a donc l'information moyenne associée à l'occurrence du message $X=[x_1,x_2,\ldots,x_I]$ qui vaut :

$$H(X) = -\sum_i p(x_i)\log_2 p(x_i) \qquad \text{(I.1)}$$

C'est l'*entropie* de la source exprimée en bits.

Dans la conversation courante, environ dix phonèmes [1] sont prononcés chaque seconde; l'information moyenne est donc inférieur à 50 bits/s [2].Or, on sait que pour un canal continu sans erreurs, le débit maximum d'information est donné par l'équation (I.2) :

$$C=B\log_2[1+S/N] \quad \text{(I.2)}$$

Avec B est la longueur de la bande passante en Hz, et S/N est le rapport signal sur bruit en dB.

Par exemple, pour un canal téléphonique, supposé continu et sans erreurs, de bande passante B=3000 Hz et avec un rapport signal sur bruit S/N=30 dB, on trouve C=30000 bits/s, il y a apparemment une redondance énorme dans ce canal. La suppression partielle des redondances permet une représentation plus efficace des données.

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

I.1.3 Modèle de production de la parole

L'analyse de la parole est une étape indispensable à toute application de synthèse, de codage ou de reconnaissance.

Le modèle électrique linéaire a été proposé par Fant [3] en 1960, qui spécifie qu'un signal voisé peut être modélisé par le passage d'un train d'impulsions $u(n)$ à travers un filtre numérique récursif de type tous-pôles (*Auto Régressif*). On montre que cette modélisation reste valable dans le cas des sons non voisés, à condition que $u(n)$ soit cette fois un bruit blanc. Le modèle final est illustré à la Figure I.4. Il est souvent appelé modèle auto régressif (AR), parce qu'il correspond dans le domaine temporel à une régression linéaire de la forme :

$$s(n)=G.u(n)+\sum_{i=1}^{p}-a_i s(n-1) \quad \text{(I.3)}$$

Où $u(n)$ est le signal d'excitation et p l'ordre du système.

Chaque échantillon est obtenu en ajoutant un terme d'excitation à une prédiction obtenue par combinaison linéaire des p échantillons précédents.

[1] Phonème : c'est la plus petite unité présente dans la parole et susceptible par sa présence de changer la signification d'un mot [2].

Les coefficients du filtre $\{a_i\}$ sont appelés coefficients de prédiction et le modèle AR est souvent appelé modèle de prédiction linéaire.

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

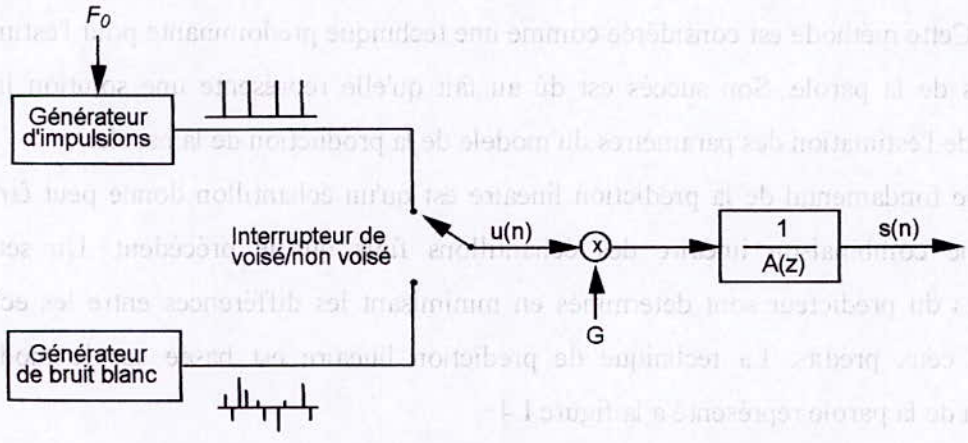

Fig. I.4 Modèle simplifié de production de la parole [3][4]

Les relations d'équivalences entre le modèle physique et le modèle mathématique [4] peuvent être données comme suit :

Conduit vocal	⟺	*H(z)*, le filtre LPC
Le flux d'air	⟺	Le signal d'excitation *u(n)*
Vibration des cordes vocales	⟺	Voisé
Fricatives et plosives	⟺	Non Voisé/Voisé
Volume d'air	⟺	*G*, le gain

Le problème de l'estimation d'un modèle AR, souvent appelée analyse *LPC*, revient à déterminer les coefficients d'un filtre tous-pôles dont on connaît le signal de sortie, mais pas celui de

l'entrée. Il est par conséquent nécessaire d'adopter un critère, afin de faire un choix parmi l'ensemble infini de solutions possibles. Le critère généralement utilisé est celui de la minimisation de l'énergie de l'erreur de prédiction.

I.1.4 Prédiction Linéaire

La prédiction linéaire est assez bien utilisée dans les systèmes de codage et de compression [6][7][8]. Cette méthode est considérée comme une technique prédominante pour l'estimation des paramètres de la parole. Son succès est dû au fait qu'elle représente une solution linéaire au problème de l'estimation des paramètres du modèle de la production de la parole.

Le principe fondamental de la prédiction linéaire est qu'un échantillon donné peut être prédit à partir d'une combinaison linéaire des échantillons finis qui le précèdent. Un seul jeu de coefficients du prédicteur sont déterminés en minimisant les différences entre les échantillons actuels et ceux prédits. La technique de prédiction linéaire est basée sur le modèle de la production de la parole représenté à la figure I.4.

Le signal parole *s(n)* peut être modélisé comme la sortie d'un système *auto régressif à moyenne ajustée* (ARMA) avec une entrée *u(n)* [3][5][9]. Son expression est alors :

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + G\sum_{i=0}^{q} b_i u(n-i), \quad b_0=1, \tag{I.4}$$

Où le gain *G*, les coefficients $\{a_k\}$ et $\{b_i\}$ sont les paramètres du système, et p et q sont les ordres des polynômes. L'équation (I.4) prédit la sortie courante en utilisant une combinaison linéaire des sorties précédentes et les entrées courantes et précédentes.

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

$$H(z) = \frac{B(z)}{A(z)} = \frac{G[1+\sum_{i=1}^{q} b_i z^{-i}]}{1-\sum_{k=1}^{p} a_k z^{-k}} \tag{I.5}$$

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

Si a_k=0 pour $1 \leq k \leq p$, *H(z)* devient un modèle tous-zéros ou modèle à *moyenne ajustée* (*MA*).

Si pour b_i=0, pour $1 \leq i \leq q$, *H(z)* devient un modèle tous-pôles ou modèle *auto régressive* (*AR*), exprimé par :

$$H(z) = \frac{1}{A(z)} \tag{I.6}$$

L'analyse spectrale montre que les pôles correspondent aux résonances du conduit vocal, c'est-à-dire aux *pics* du spectre, les *formants* ; tandis que les zéros correspondent aux antirésonances, c'est-à-dire aux *vallées*.

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 tous-pôles; 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 :

$$\tilde{s}(n) = \sum_{k=1}^{p} a_k s(n-k) \tag{I.7}$$

La différence entre l'échantillon original $s(n)$ et l'échantillon prédit $\tilde{s}(n)$ est appelée *erreur de prédiction* (ou *résidu*) et elle est définie par:

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \tag{I.8}$$

Le problème de l'analyse par prédiction linéaire se réduit donc à trouver un ensemble de coefficients a_k de façon à minimiser l'erreur de prédiction $e(n)$ dans un certain intervalle. Les méthodes d'estimation des coefficients a_k sont nombreuses [10].

Deux grandes approches sont utilisées pour l'analyse par prédiction linéaire LPC court-terme : La méthode d'autocorrélation et la méthode de covariance.

I.1.4.1 Méthode d'Autocorrélation

La méthode d'autcorrélation garantit la stabilité du filtre LP. Les hypothèses 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 est équivaut à multiplier le signal de parole *s(n)* par une fenêtre *w(n)* de longueur finie correspondant à *N* échantillons pour obtenir un segment du signal de parole fenêtré $s_w(n)$ [11].

$$s_w(n) = \begin{cases} w(n).s(n) & pour \quad 0 \le n \le N-1 \\ 0 & ailleurs \end{cases} \tag{I.9}$$

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

$$w(n) = \begin{cases} 0.54 - 0.46\cos\dfrac{2n\pi}{N-1} & pour \quad 0 \le n \le N-1 \\ 0 & ailleurs \end{cases} \tag{I.10}$$

Chaque échantillon peut être prédit approximativement à partir des échantillons précédents. Ceci est valable pour toutes les valeurs du temps; ($-\infty<n<+\infty$).

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

La fonction d'autocorrélation du signal fenêtré $s_w(n)$ est :

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

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

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

$$E = \sum_{n=-\infty}^{\infty} e^2(n) = \sum_{n=-\infty}^{\infty} [s_w(n) - \sum_{k=1}^{p} a_k s_w(n-k)]^2 \tag{I.12}$$

Cette erreur peut être minimisée en annulant les dérivées partielles par rapport aux coefficients du filtre :

$$\frac{\partial E}{\partial a_k} = 0 \qquad 1 \le k \le p \tag{I.13}$$

On obtient p équation linéaire avec p coefficient inconnus a_k :

$$\sum_{k=1}^{p} a_k \sum_{n=-\infty}^{\infty} s_w(n-i)s_w(n-k) = \sum_{n=-\infty}^{\infty} s_w(n-i)s_w(n). \qquad tq : 1 \le i \le p \tag{1.14}$$

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

$$\sum_{k=1}^{p} R(|i-k|)a_k = R(i) \qquad 1 \le i \le p \tag{1.15}$$

La forme matricielle de l'ensemble des équations linéaires (I.14) est représenté par $\boldsymbol{R.a=v}$ et peut être réécrite comme suit :

$$\begin{bmatrix} R(0) & R(1) & \dots & R(p-1) \\ R(1) & R(0) & \dots & R(p-2) \\ & & \dots & \\ . & . & \dots & . \\ . & . & \dots & . \\ R(p-1) & R(p-2) & \dots & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ . \\ . \\ . \\ a_p \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ . \\ . \\ . \\ R(p) \end{bmatrix} \tag{I.16}$$

La matrice d'autocorrélation *pxp* obtenue est symétrique dont tous les éléments de la diagonale sont égaux, c'est une matrice de *Toeplitz*. Ce qui nous permet de trouver les coefficients de prédiction minimisant la moyenne quadratique de l'erreur de prédiction par l'algorithme de *Levinson-Durbin* (Annex A).

I.1.4.2 Méthode de Covariance

Les méthodes d'autocorrélation et de covariance diffèrent dans l'emplacement de la fenêtre d'analyse.

Dans cette méthode c'est le signal erreur qui est fenêtré au lieu du signal parole, de façon à ce que l'énergie à minimiser soit :

$$E = \sum_{n=-\infty}^{\infty} e_w{}^2(n) = \sum_{n=-\infty}^{\infty} e^2(n)w^2(n) \tag{I.17}$$

En annulant les dérivées partielles en utilisant l'équation (I.13) on obtient p équations linéaires :

$$\sum_{k=1}^{p} \Phi(i,k) = \Phi(i,0) \qquad 1 \le i \le p \tag{I.18}$$

Où la fonction de covariance :

$$\Phi(i,k) = \sum_{n=-\infty}^{\infty} w(n)s(n-1)s(n-k) \tag{I.19}$$

On peut exprimer les p équations, sous la forme : $\Phi.a=\Psi$

$$\begin{bmatrix} \Phi(1,1) & \Phi(1,2) & \dots & \Phi(1,p) \\ \Phi(2,1) & \Phi(2,2) & \dots & \Phi(2,p) \\ \Phi(3,1) & \Phi(3,2) & \dots & \Phi(3,p) \\ & & \dots & \\ & & \dots & \\ \Phi(p,1) & \Phi(p,2) & \dots & \Phi(p,p) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \\ \\ a_p \end{bmatrix} = \begin{bmatrix} \Psi(1) \\ \Psi(2) \\ \Psi(3) \\ . \\ . \\ \Psi(p) \end{bmatrix} \tag{I.20}$$

Tel que; $\Psi(i)=\Phi(i,0)$ pour $1 \le i \le p$

La matrice Φ n'est pas une matrice Toeplitz, et ne garantie pas la stabilité du filtre LPC, elle est symétrique et définie positive. Donc, la matrice de covariance peut être décomposée en deux matrices, l'une triangulaire inférieure L et l'autre triangulaire supérieure U.

$$\Phi=L.U \tag{I.21}$$

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

$$\Phi=C.C^T \qquad \text{tq; } C=L \text{ et } C^T=U$$

Le vecteur a est obtenu en résolvant d'abord l'équation (I.22) :

$$L.y=\Psi \tag{I.22}$$

Puis :

$$U.a=y \tag{I.23}$$

I.1.4.3 Considération Pratiques

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

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

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

- 8 kHz pour les signaux téléphoniques.
- 10 kHz pour les applications de reconnaissance.
- 16 kHz pour les applications de synthèse.

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

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

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

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

Enfin, comme vu précédemment dans la méthode d'autocorrélation, pour compenser les effets de bord, on multiplie en général préalablement chaque tranche d'analyse par une fenêtre de pondération $w(n)$, la plus souvent utilisées est celle de *Hamming* (équation (I.10)).

I.1.4.4 Représentation des paramètres de prédiction

Les coefficients de prédiction linéaire (*LP*) sont calculés à base de "bloc par bloc", généralement sur des trames de 5-40ms [12]. Pour une transmission efficace de la parole, les coefficients *LP* sont sujets à une ***quantification*** et une ***interpolation***. L'interpolation rend possible la transmission de l'information sur les coefficients *LP* moins souvent, ainsi réduisant le débit binaire. Cependant, une simple quantification ou une interpolation des coefficients *LP* est une problématique parce que de petits changements dans les coefficients peuvent induire un grand changement dans le spectre de puissance et causer l'instabilité du filtre de synthèse *LP*. Par conséquent, un nombre de représentations des coefficients *LP* été considéré pour essayer de trouver la représentation qui minimise ses limitations.

Les représentations les plus utilisées sont les coefficients de réflexion, les LAR (log-area ratios) [12] et les LSPs (Line Spectrum Pairs) [13].

Cependant la représentation la plus répandue et la plus prisée pour ses performances reste la représentation en paires de rais spectrales LSP.

Elles seront détaillées dans ce qui va suivre.

I.1.4.4.1 Paires de raies spectrales

Connus aussi sous le nom de fréquences de raies spectrales.

La représentation LSP a été introduite par *Itakura* [13].

Les LSPs sont les solutions des deux équations suivantes :

$$\begin{cases} P(z) = A(z) + z^{-(p+1)}A(z) \\ Q(z) = A(z) - z^{-(p+1)}A(z) \end{cases} \tag{I.24}$$

Ce qui nous donne :

$$A(z) = \frac{1}{2}[P(z) + Q(z)] \tag{I.25}$$

Soong et *Juang* [14] ont montrés que si *H(z)* est stable, où *A(z)* est à phase minimale, alors les zéros des polynômes *P(z)* et *Q(z)* sont appels les LSP. Ces polynômes ont les propriétés suivantes [4]:

- Tous les zéros de *P(z)* et *Q(z)* se trouvent sur le cercle unité.
- Les zéros de *P(z)* et *Q(z)* sont entrelacés les uns aux autres, les LSP sont dans un ordre croissant.

Il a été montré [15] que le filtre LPC *A(z)* est à phase minimale si et seulement si les LSP 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

1. comme illustré dans la figure I.5 il y a une relation évidente entre les LSP et le spectre du filtre LPC. Une concentration des LSP dans une certaine bande de fréquences correspond approximativement à une résonance dans cette bande.
2. 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.

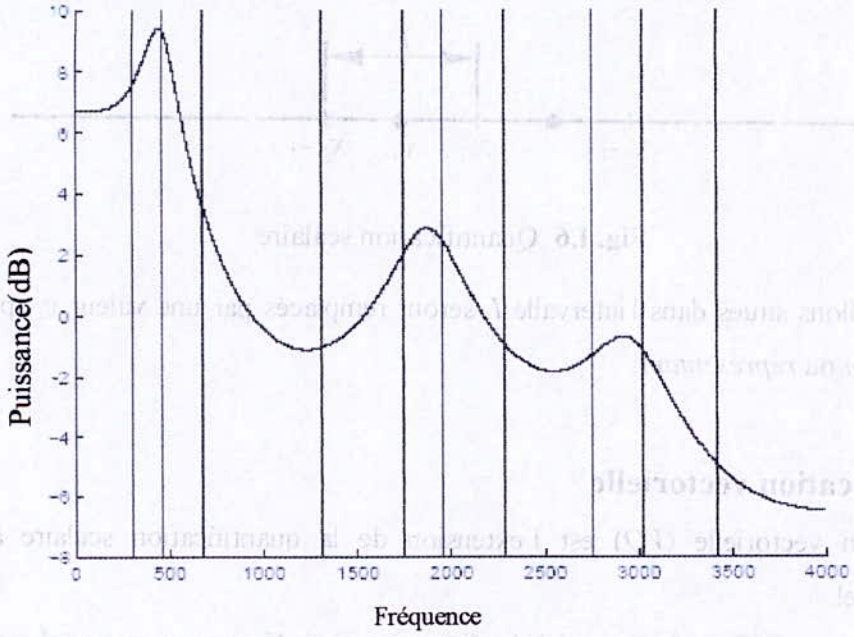

Fig. I.5 Spectre LPC avec LSF superposé

I.2 Principe de la quantification

La quantification est le processus de substitution des échantillons d'un signal analogique par des valeurs arrondies prises parmi un nombre fini de valeurs possibles [4].

La quantification peut être *scalaire* ou *vectorielle* selon que les signaux sont à une ou plusieurs dimensions. La quantification vectorielle peut être de deux types soit statistique ou algébrique.

I.2.1 Quantification scalaire

Dans la quantification scalaire (*QS*), chaque échantillon du signal d'entrée est quantifié séparément des autres échantillons. Comme l'illustre la figure I.6, un échantillon x du signal d'entrée est spécifié par l'indice k s'il se trouve dans l'intervalle suivant :

$$I_k : \{x_k \prec x \leq x_{k+1}\} \qquad k = 1,2,....., N \tag{I.26}$$

Les valeurs x_k et x_{k+1} sont appelées niveaux de décision ou seuils.

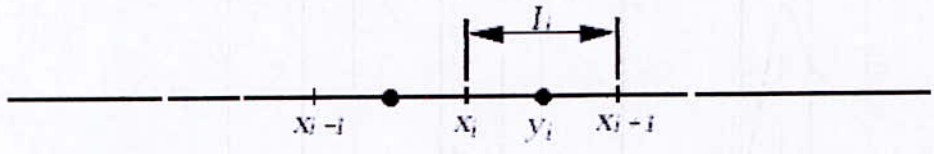

Fig. I.6 Quantification scalaire

Tous les échantillons situés dans l'intervalle I_i seront remplacés par une valeur y_i appelée *niveau de reconstruction* ou *représentant*.

I.2.2 Quantification vectorielle

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

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

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

Contrairement à la quantification scalaire, un quantificateur vectoriel peut fonctionner avec un débit fractionnaire (R < 1) [5].

I.3 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 [32]

- 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.

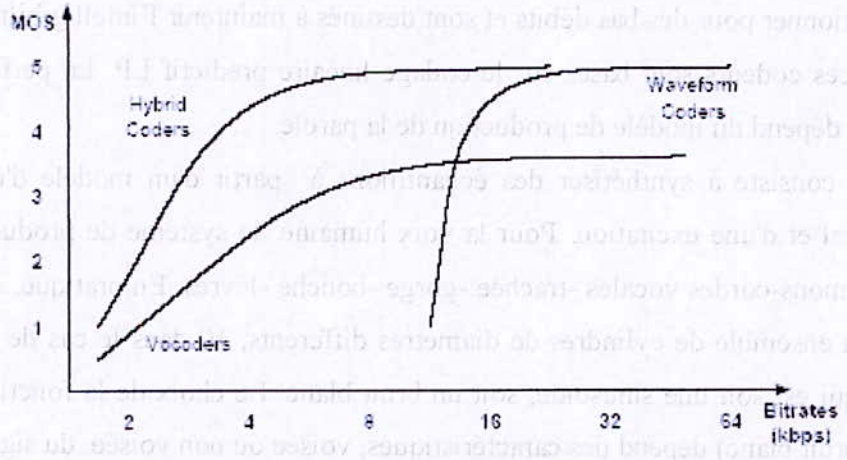

Fig. I.7 Comparaison de la qualité de codage de parole [32]

I.3.1 Le codage de forme d'onde

Les codeurs de formes d'ondes sont relativement simples à mettre en œuvre, ils produisent une qualité acceptable jusqu'à des débits de 16 Kbits/s. En deçà, la qualité du signal reconstruit se dégrade rapidement.

L'algorithme de codage le plus simple est celui qui revient seulement à échantillonner un signal analogique et à quantifier les échantillons (c'est à dire à les convertir des valeurs réelles en valeurs de précision finie) ; ce codage est appelé PCM (*Pulse Coded Modulation*).
Le codage PCM est à la base d'une famille de codages différentiels qui est basé sur l'observation que des échantillons successifs d'une source audio sont fortement corrélés. Il semble donc judicieux d'encoder non pas les échantillons eux même mais la différence entre des échantillons successifs. On peut citer:

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

I.3.2 Le codage par synthèse

Connu aussi sous le nom de codage de source ou vocodeurs (voice coders), ces codeurs sont destinés à fonctionner pour des bas débits et sont destinés à maintenir l'intelligibilité de la parole. La plupart de ces codeurs sont basés sur le codage linéaire prédictif LP. La performance de ce type de codage dépend du modèle de production de la parole.
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.

I.3.3 Le Codage Hybride

La qualité des codeurs de formes d'ondes chute rapidement pour des débits inférieurs à 16 kbits/s, et comme les vocodeurs apportent une amélioration négligeable dans la qualité à des débits supérieurs à 4 kbits/s, Les codeurs hybrides sont alors utilisés pour combler ce vide, donnant ainsi une bonne qualité de la parole à des débits moyens. Cependant, ces codeurs ont tendance à nécessiter un nombre d'opérations plus élevé. Virtuellement, tous les codeurs hybrides reposent

sur l'analyse LPC pour l'obtention des paramètres du modèle de synthèse. Les techniques de formes d'ondes utilisées pour coder le signal d'excitation et les modèles de production du pitch peuvent être incorporés pour améliorer les performances.

A partir des années 80, l'intérêt pour les codeurs CELP (Code-Exited Linear Prediction) ne cesse d'augmenter. Ces codeurs sont basés sur les algorithmes de codage de la parole les plus actuellement utilisés dans la téléphonie sans fil. Dans les codeurs CELP, l'analyse LP est utilisée pour obtenir le signal d'excitation. La modélisation du pitch est utilisée pour coder efficacement le signal d'excitation. Le standard G.729 de l'ITU est un codeur CELP qui produit une qualité téléphonique (toll quality) de la parole à 8 kbits/s [5].

I.4 Qualité des codeurs

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

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

I.4.1 Mesure de distorsion subjective

L'évaluation subjective est obtenue par des tests d'écoutes; dans ces tests, la qualité de la parole est mesurée par l'intelligibilité spécifiquement définie par le pourcentage de mots ou phonèmes correctement écoutés et avec une sonorité naturelle (naturalness).

Il existe trois types de mesures subjectives [4] de la qualité généralement utilisées.

- Le test DRT (Diagnostic Rhyme Test)
- Le test DAM (Diagnostic Acceptability Measure)
- Le test MOS (Mean Opinion Score)

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

Tableau I.1 Qualité avec la mesure MOS.

I.4.2 Mesure de distorsion objective

Le système auditif de l'être humain est l'estimateur le plus adéquat de la qualité et des performances d'un codeur de la parole. Il permet de préciser l'intelligibilité et la sonorité naturelle des sons. Bien que, Les tests d'écoute subjectifs donnent une bonne évaluation pour les codeurs de la parole, ils peuvent exiger beaucoup de temps et sont non conformé. Les mesures objectives peuvent donner une estimation immédiate de la qualité perceptuelle de la parole [16].

Les mesures objectives de distorsions peuvent être calculées aussi bien dans le domaine temporel que fréquentiel [4].

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

Les mesures de distorsions sont classifiées comme suit [2] [4] :

- Domaine temporel (RSB et RSBseg)
- Domaine fréquentiel (distorsion spectrale)

I.4.2.1 Domaine temporel

- Rapport Signal sur Bruit :

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

$$RSB = 10\log\left(\frac{\sum_{n=0}^{Nt-1} S(n)^2}{\sum_{n=0}^{Nt-1}\left[S(n)-\widetilde{S}(n)\right]^2}\right) \quad (dB) \tag{I.26}$$

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

- Rapport Signal sur Bruit segmenté :

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

$$RSB_{seg} = \frac{1}{N_F}\sum_{i=0}^{N_F-1} 10\log\left(\frac{\sum_{j=0}^{Ns-1} S(N_s i+j)^2}{\sum_{j=0}^{Ns-1}\left[S(N_s i+j)-\widetilde{S}(N_s i+j)\right]^2}\right) \quad (dB) \tag{I.27}$$

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

I.4.2.2 Domaine fréquentiel

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

$$DS_i = \sqrt{\frac{1}{F_e}\int_0^{F_e}\left[10Log_{10}\frac{S_i(f)}{\widetilde{S}_i(f)}\right]^2 df} \qquad (dB) \tag{I.28}$$

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

$$S_i(f) = \frac{1}{A_i(e^{j2\pi f/F_e})} \tag{I.29}$$

$$\widetilde{S}_i(f) = \frac{1}{\widetilde{A}_i(e^{j2\pi f/F_e})} \tag{I.30}$$

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

$$DS_i = \sqrt{\frac{1}{n_1 - n_0}\sum_{k=n_0}^{n_1-1}\left[10\log\frac{S_i(e^{j2\pi k/N})}{\widetilde{S}_i(e^{j2\pi k/N})}\right]^2} \qquad (dB) \tag{I.31}$$

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

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

I.4.3 Mesure de distance euclidienne LSP pondérée

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

$$d_{LSF} = \sum_{i=1}^{p} [c_i w_i (\omega_i - \tilde{\omega}_i)]^2 \quad \text{(I.32)}$$

Ou c_i et w_i sont les poids du i^{eme} coefficients LSP ω_i, et p est l'ordre du filtre LP. Pour un filtre d'ordre 10, les poids fixes c_i sont donnés par :

$$c_i = \begin{cases} 1.0 & pour \quad 1 \le i \le 8 \\ 0.8 & pour \quad i = 9 \\ 0.4 & pour \quad i = 10 \end{cases} \quad \text{(I.33)}$$

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

$$w_i = \left[S(e^{j\omega})\right]^r \quad \text{(I.34)}$$

Où r est une constante empirique qui contrôle le degré de la pondération, empiriquement r=0.15. Une pondération plus simple a été proposée par [26], elle a la forme suivante :

$$w_i = \frac{1}{\omega_i - \omega_{i-1}} + \frac{1}{\omega_{i+1} - \omega_i} \quad ou \quad \omega_0 = 0 et \quad \omega_{p+1} = \pi \quad \text{(I.35)}$$

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

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

Conclusion

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

Chapitre II

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

Introduction

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

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

La voix sur *IP* fait appel à deux notions : le *réseau IP* et la voix véhiculée traditionnellement par le *réseau RTC*

II.1 Architecture des réseaux

Un réseau est un ensemble d'ordinateurs (ou de périphériques) autonomes connectés entre eux et qui sont situés dans un certain domaine géographique.

Suivant la distance qui sépare les ordinateurs, on distingue plusieurs catégories de réseaux :

- Les LAN (Local Area Network)
- Les MAN (Metropolitan Area Network)
- Les WAN (Wide Area Network)

Une normalisation de l'architecture logicielle s'impose. Deux grandes familles d'architectures se disputent le marché :

La première provient de l'ISO et s'appelle OSI (Open System Interconnection).

La deuxième est TCP/IP.

Une 3ème Architecture plus récente est UIT - T (Union Internationale de télécommunications). Il s'agit de l'adaptation du modèle OSI pour prendre en compte les réseaux hauts - débit (réseau ATM).

II.1.1 Modèle de référence OSI

Pour faire circuler l'information sur un réseau on peut utiliser principalement deux stratégies.

- L'information est envoyée de façon complète.
- L'information est fragmentée en petits morceaux (*paquets*), chaque paquet est envoyé séparément sur le réseau, les paquets sont ensuite réassemblés sur la machine destinataire.

Dans la seconde stratégie on parle de réseaux à *commutations de paquets.*

La première stratégie n'est pas utilisée car les risques d'erreurs et les problèmes sous-jacents sont trop complexes à résoudre.

Le modèle OSI est un modèle à 7 couches (figure II.1) qui décrit le fonctionnement d'un réseau à commutations de paquets. Chacune des couches de ce modèle représente une catégorie de problème que l'on rencontre dans un réseau.

Découper les problèmes en couche présente des avantages. Lorsqu'on met en place un réseau, il suffit de trouver une solution pour chacune des couches.

L'utilisation de couches permet également de changer de solution technique pour une couche sans pour autant être obligé de tout repenser.

Chaque couche garantit à la couche qui lui est supérieur, que le travail qui lui a été confié a été réalisé sans erreur.

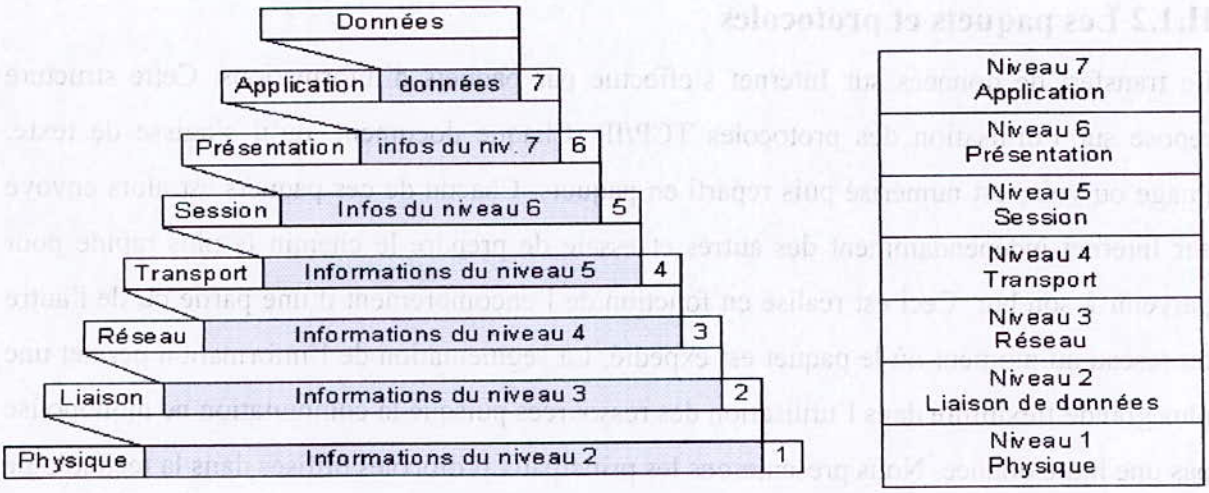

Fig. II.1 Modèle de référence OSI

- La 1ère couche (couche Matériel)

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

- La 2ème couche (couche Liaison)

Dans cette couche on cherche à savoir comment deux *stations* sur le même support physique vont être identifiées. Pour ce faire, on peut par exemple assigner à chaque station une adresse.

- La 3ème couche (couche Réseau)

Le rôle de cette couche est de trouver un chemin pour acheminer un paquet entre 2 machines qui ne sont pas sur le même support physique.

- La 4ème couche (couche Transport)

La couche transport doit normalement permettre à la machine source de communiquer directement avec la machine destinatrice. On parle de communication de bout en bout (*end to end).*

- La 5ème couche (couche Session)

Cette couche a pour rôle de transmettre cette fois les informations de programmes à programmes.

- La 6ème couche (couche Présentation)

A ce niveau on doit se préoccuper de la manière dont les données sont échangées entre les applications.

- La 7ème couche (couche Application)

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

II.1.2 Les paquets et protocoles

Le transfert de données sur Internet s'effectue par paquets d'informations. Cette structure repose sur l'utilisation des protocoles TCP/IP. Chaque document, qu'il s'agisse de texte, image ou voix, est numérisé puis réparti en paquets. Chacun de ces paquets est alors envoyé sur Internet indépendamment des autres et essaie de prendre le chemin le plus rapide pour parvenir à son but. Ceci est réalisé en fonction de l'encombrement d'une partie ou de l'autre du réseau au moment où le paquet est expédié. La segmentation de l'information permet une plus grande flexibilité dans l'utilisation des ressources puisque la commutation ne monopolise pas une ligne donnée. Nous présenterons les principaux protocoles utilisés dans la technologie de la voix sur Internet.

II.1.2.1 protocoles TCP/IP

Le protocole *IP* 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.

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 réacheminer 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.

II.1.2.2 Protocole 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. UDP correspond au niveau transport de l'architecture du modèle OSI, mais c'est un protocole particulièrement simple. Son avantage

est un temps d'exécution court qui permet de tenir compte des contraintes de temps réel ou de limitation d'espace mémoire sur un processeur, contraintes qui ne permettent pas l'implantation de protocoles beaucoup plus lourd comme TCP. De plus, les mécanismes de TCP prévoient une réduction automatique du débit accordé à l'émetteur en cas de congestion du réseau et une remontée lente vers le débit nominal.

II.1.2.3 Protocole RTP

Le protocole de Transport en Temps Réel (RTP) est le standard proposé pour faciliter la synchronisation et la récupération des variations de délai et de perte de paquets. RTP peut également véhiculé par des paquets multicast afin d'acheminer des conversations vers les destinataires multiples; en faite, il a été conçu directement pour un environnement multipoint. RTP aura donc à sa charge aussi bien la gestion du temps réel que l'administration de la session multipoint. Le protocole est encapsulé dans un datagramme UDP et permet d'enrichir ce dernier afin de le rapprocher de TCP offrant des services tels que le séquencement, la gestion de la gigue et la fourniture d'une horloge.

Le rôle principal de RTP consiste à mettre en œuvre des numéros de séquence de paquets *IP* pour reconstituer les informations de voix ou vidéo même si le réseau change d'ordre des paquets, ce qui est susceptible de se produire dans la mesure où le fonctionnement d'Internet ne garantit pas que deux paquets successifs empruntent le même chemin. Cela permet, par exemple pour des applications vidéo, de décoder et placer au bon endroit sur l'écran chaque paquet sans attendre ses prédécesseurs et pour des applications de voix de reconstituer les échantillons de parole.

II.1.2.4 Protocole RTCP

Le protocole de contrôle en temps réel (RTCP) est un protocole d'accompagnement pour l'analyse et la gestion de flux en temps réels. Il coordonne les rapports réalisés par l'expéditeur et le récepteur qui sont livrés périodiquement. Les hôtes RTP peuvent utiliser les rapports RTCP pour obtenir des réactions sur la qualité de livraison des paquets RTP, et modifier les paramètres de transmissions en conséquence. Pour des types d'applications particulières, RTCP peut être complété par un autre protocole de transport. Le protocole transmet périodiquement des paquets de contrôle aux participants, en utilisant les mêmes moyens de diffusion que RTP, mais avec un numéro de port différent. RTCP permet de

recevoir des informations de retour des participants, grâce aux messages «*send report*» et «*receiver report*».

II.1.3 Format de l'en-tête IP

La fonction ou le rôle du Protocole Internet est d'acheminer les datagrammes à travers un ensemble de réseaux interconnectés. Ceci est réalisé en transférant les datagrammes d'un module Internet à l'autre jusqu'à atteindre la destination. Les modules Internet sont des programmes exécutés dans des hôtes et des routeurs du réseau Internet. Les datagrammes sont transférés d'un module Internet à l'autre sur un segment particulier de réseau selon l'interprétation d'une adresse Internet. De ce fait, un des plus importants mécanismes du protocole Internet est la gestion de cette adresse Internet.

Lors de l'acheminement d'un datagramme d'un module Internet vers un autre, les datagrammes peuvent avoir éventuellement à traverser une section de réseau qui admet une taille maximale de paquet inférieure à celle du datagramme. Pour surmonter ce problème, un mécanisme de fragmentation est géré par le protocole Internet.

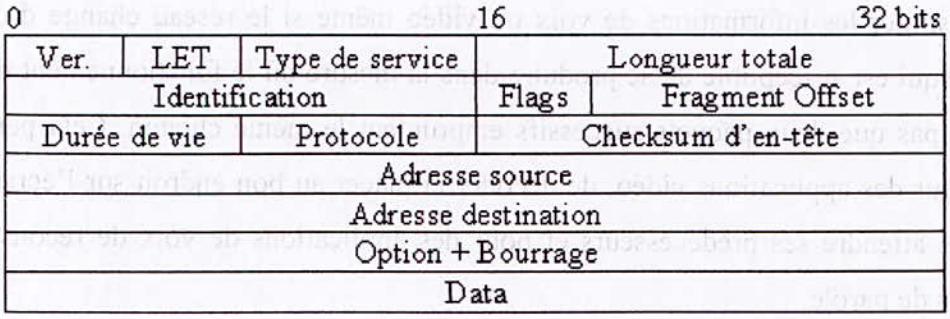

Fig. II.2 format d'en-tête IP

- **Version** sur 4 bits : définie le numéro de la version du protocole IP
- **Longueur d'En-tête (LET)** sur 4 bits : code la longueur de l'en-tête.
- **Type de Service** sur 8 bits : donne une indication sur la qualité de service souhaitée, qui reste cependant un paramètre "abstrait". Ce paramètre est utilisé pour "guider" le choix des paramètres des services actuels lorsqu'un datagramme transite dans un réseau particulier. ce champ est malheureusement pour *VoIP*, peu pris en compte par les routeurs actuels.
- **Longueur Totale** sur 16 bits : donne la longueur du datagramme entier y compris l'en-tête et données (mesurée en octets).

- **Identification** sur 16 bits : identifie les fragments d'un même datagramme.
- **Flags** sur 3 bits : Divers commutateurs de contrôle.

  Bit 0 : réservé, doit être laissé à zéro

  Bit 1: (AF) 0 = Fragmentation possible, 1 = Non fractionnable.

  Bit 2: (DF) 0 = Dernier fragment, 1 = Fragment intermédiaire.
- **Fragment Offset** sur13 bits : indique le décalage du premier octet du fragment par rapport au datagramme complet
- **Durée de vie** sur 8 bits : limite le temps pendant lequel un datagramme reste dans le réseau. Si ce champ prend la valeur zéro, le datagramme doit être détruit. Chaque module IP (routeur) doit retirer au moins une unité de temps à ce champ. Ce mécanisme est motivé par la nécessité de détruire les datagrammes qui n'ont pu être acheminés, en limitant la durée de vie même du datagramme et ainsi éviter tout problème de surcharge réseau.
- **Protocole** sur 8 bits : indique quel protocole de niveau supérieur est utilisé dans la section « données » du datagramme IP (UDP, TCP,...).
- **Checksum d'en-tête** sur 16 bits : calculé sur l'en-tête uniquement. Comme certains champs de l'en-tête sont modifiés (ex : durée de vie) pendant leur transit à travers le réseau, ce Checksum doit être recalculé et vérifié en chaque point du réseau où l'en-tête est réinterprétée.
- **Adresse source** sur 32 bits : l'adresse Internet de la source.
- **Adresse destination** sur 32 bits : l'adresse Internet du destinataire.
- **Options** (variable) : les datagrammes peuvent contenir des options.

II.2 La Voix sur les Réseaux IP

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

Avec un appel *VoIP*, la partie de l'établissement de l'appel doit être simulé c'est-à-dire la tonalité, les signaux de sonneries et les signaux d'occupation. En plus, l'appel lui-même (c'est-à-dire la conversation) a besoin d'être converti de son format analogique à un format

numérique, découpé en paquets et envoyé à travers le réseau, rassemblé de nouveau, et reconverti du format numérique au format analogique.

Les Codecs (Coder and Decoder) à chaque point font la conversion de l'analogique au numérique et vice vers ça [17][18].

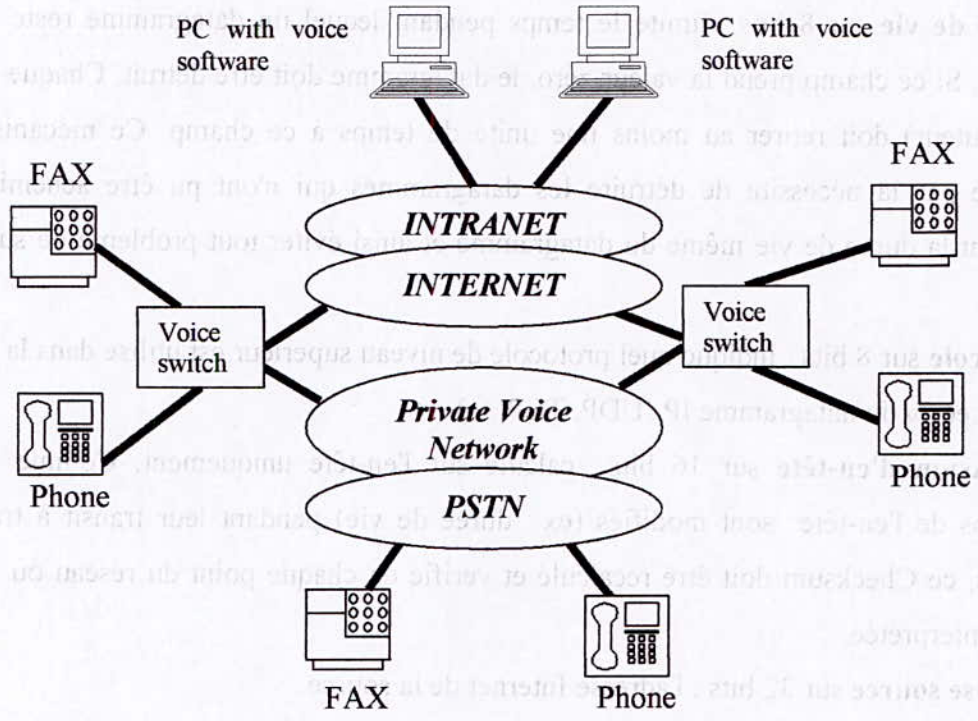

Fig. II.3 Infrastructure du système VoIP

Le modèle du service Internet actuel est horizontal, offrant des classes et un système de délivrance basé sur le meilleur effort (best-effort). Le plus grand problème que rencontre la transmission de la voix à travers les réseaux par paquets est d'assurer une qualité de service comparable à celle obtenue par les réseaux téléphoniques traditionnels.

Ils existent plusieurs facteurs qui déterminent la qualité de service délivrée par le réseau [19]. On peut citer les codecs, la bande passante, le retard, la gigue et les pertes de paquets dans le réseau.

II.2.1 Aperçu des techniques de codage de la voix dans le cadre des transmissions

Dans le domaine du traitement de la parole, on peut distinguer différents types de codage.

Grossièrement, on peut dire qu'il y a les codeurs qui effectuent une quantification scalaire et ceux qui fonctionnent avec une quantification vectorielle, ces derniers étant les plus performants en matière de débit.

Le signal de départ est un signal analogique qui va être échantillonné à une fréquence deux fois supérieure à la fréquence maximale du signal. Ensuite, ces échantillons vont être quantifiés.

II.2.2 Types de la téléphonie sur Internet

Nous pouvons, en une première approche, distinguer trois types de téléphonie sur Internet selon le terminal utilisé par chacun des deux correspondants.

- La communication téléphonique entre deux ordinateurs connectés sur Internet.
- La communication téléphonique entre un ordinateur et un téléphone ou entre deux ordinateurs au moyen de passerelles jouant le rôle d'interface entre le réseau téléphonique et Internet.
- La communication téléphonique entre deux téléphones à l'aide des passerelles et à travers l'Internet et on se servant des serveurs de contrôle.

II.2.3 Les composants VoIP

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

- Les codecs
- Les protocoles *VoIP*
- Les serveurs de téléphonie *IP* et les PBX (Private Branche eXchange)
- Les routeurs et les gateways *VoIP*
- Les gatekeepers
- Les téléphones *IP* et les softphones

II.2.4 La qualité de service

La notion de qualité de service appliquée aux communications prévoit l'établissement de deux listes soumises au réseau lors d'une demande de connexion. La première liste adresse les paramètres de qualité que l'on souhaite typiquement obtenir et maintenir.

La seconde liste fixe les valeurs minimales acceptables pour cette qualité de service, c'est la tolérance. Si les valeurs minimales ne peuvent être fournies par l'un des réseaux traversés ou par l'entité distante, la demande de connexion est refusée.

Une telle démarche est impossible sur Internet, et cela est dû à un certain nombre de contraintes que nous exposerons ci après.

II.2.4.1 Les niveaux de qualité

Pour faciliter l'analyse, on peut repérer trois niveaux de qualité (figure II.4) :

- Niveau Q1

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

- Niveau Q2

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

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

- Niveau Q3

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

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

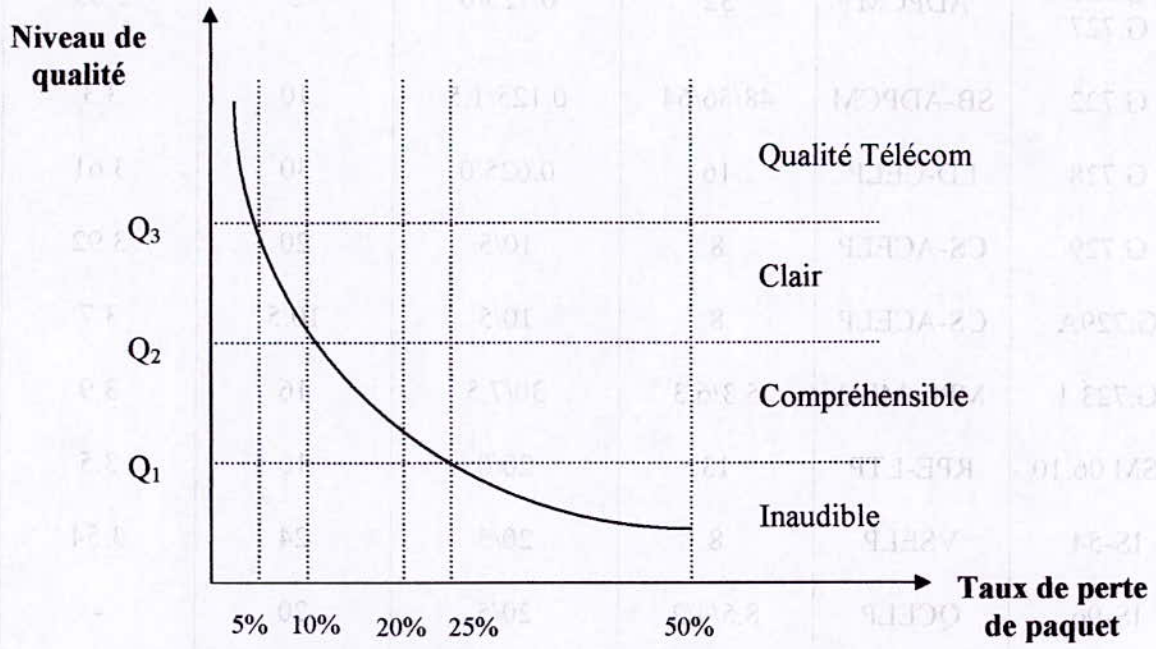

Fig. II.4 Qualité de service en fonction des pertes de paquets

II.2.4.2 Les Facteurs Affectant la Qualité de Service

II.2.4.2.1 Les Codecs

Les services de la téléphonie par Internet doivent opérer dans un environnement dont les contraintes sont les suivantes : largeur de la bande, retard, perte et coût. Récemment, les codecs de l'ITU, G.711, G.723.1, G.729 et G.729A [20][21] ont été conçu pour travailler avec ces contraintes. Ils sont conçus pour différentes applications, ils représentent de bons candidats pour les transmissions *VoIP*. Le tableau II.1 montre les performances pour différents codecs.

Standards	Algorithme	Débit (Kbits/s)	Taille trame (ms/lookahead)	Complexité (MIPS)	Qualité (MOS)
G.711	LOG PCM	64	0.125/0	0.01	4.1
G.726 G.727	ADPCM	32	0.125/0	2	3.85
G.722	SB-ADPCM	48/56/64	0.125/1.5	10	3.3
G.728	LD-CELP	16	0.625/0	30	3.61
G.729	CS-ACELP	8	10/5	20	3.92
G.729A	CS-ACELP	8	10/5	10.5	3.7
G.723.1	MPC-MLQ	5.3/6.3	30/7.5	16	3.9
GSM 06.10	RPE-LTP	13	20/0	10	3.5
IS-54	VSELP	8	20/5	24	3.54
IS-96	QCELP	8.5/4/2	20/5	20	-
FS-1016	CELP	4.8	-	30	3.0
FS-1025	CELP	2.4	-	15	2.4

Tableau II.1 Les Codecs et leurs performance

II.2.4.2.2 Le Retard

Le retard cause deux problèmes, l'écho et le chevauchement. L'écho est causé par la réflexion du signal vocal émet sur le terminal téléphonique du récepteur. L'écho devient un problème considérable dés lors que le délai de l'allé retour devient supérieur à 50 millisecondes.Le chevauchement devient perceptible si le retard engendré dans un seul sens devient supérieur à 250 ms.

Pour obtenir le décalage total entre l'émetteur et le récepteur, il faut sommer trois délais : le temps de traitement par les terminaux d'extrémité, le temps d'acheminement et enfin le temps d'accumulation. En cas de congestion persistante, ces trois temps divergent ensemble.

II.2.4.2.3 La Gigue

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

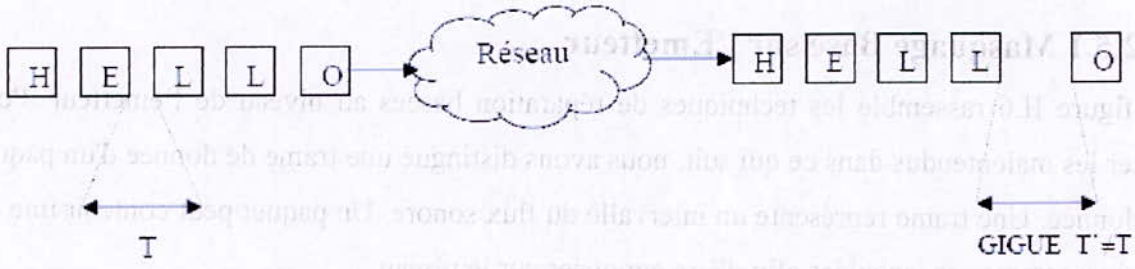

Fig. II.5 Schéma illustrant la GIGUE

II.2.4.2.4 La bande passante

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

II.2.4.2.5 Les Pertes de Paquets

La voix supporte bien les pertes de paquets par rapport à d'autres applications. On considère que le taux de pertes doit être inférieur à 20 %. La retransmission des paquets erronés ou perdus est inutile car elle induirait un temps de latence trop important. Dans un réseau *IP*, lorsque le débit offert à une liaison excède durablement le débit maximal de cette liaison, la mémoire tampon correspondante élimine un certain nombre de paquets. En revanche, il existe dans les terminaux un mécanisme d'ajustement de la fenêtre aux pertes de paquets constatées. Ce mécanisme ralentit l'émission de paquets en cas d'encombrement du réseau. Il s'agit là

d'un mécanisme d'auto-limitation mis en jeu par les terminaux ; d'où la nécessité d'employé des techniques de masquage des paquets perdus.

II.2.5 Les Techniques de Masquage des Paquets perdus

Vu l'impacte très néfaste qu'ont les pertes de paquets sur la qualité des transmissions des flux sonores, Des algorithmes de masquage des pertes PLC (Packet Loss Concealement) sont utilisés au niveau de l'émetteur ou du récepteur afin de combler les pertes de paquets.
Ces techniques peuvent être divisées en deux classes basées respectivement sur l'émetteur (sender-based) et le récepteur (receiver-based), comme indiqué sur la figure II.6 [22][23].

II.2.5.1 Masquage Basé sur l'Emetteur

La figure II.6 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.

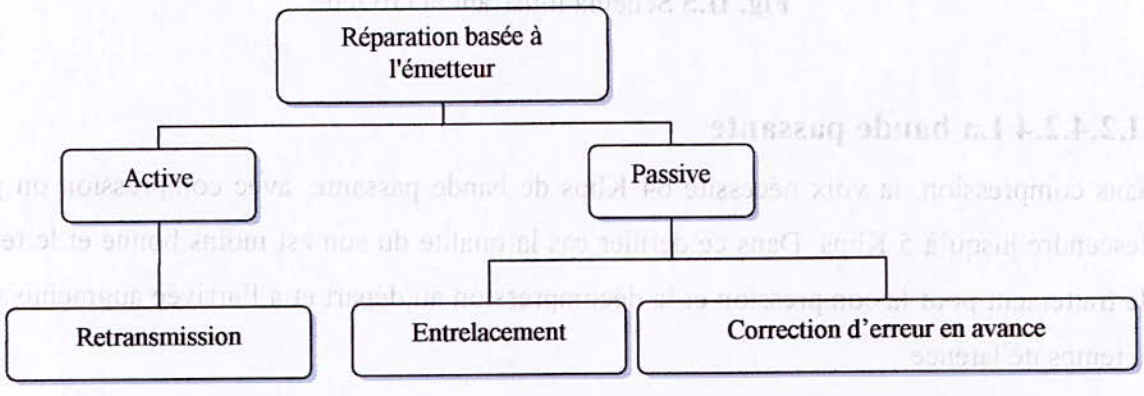

Fig. II.6 Classification des techniques de réparations basées à l'émetteur

II.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 II.7.

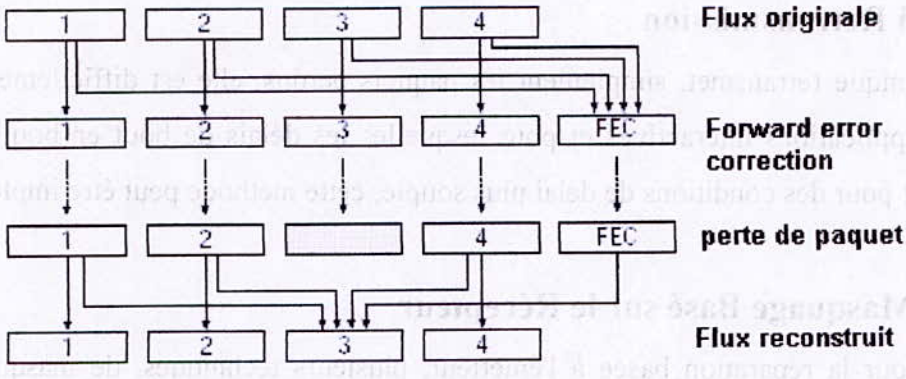

Fig.II.7 Exemple de FEC [8]

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 [24].

II.2.5.1.2 Entrelacement

La technique d'entrelacement ou interleaving est très utile lorsque, les paquets contiennent plusieurs trames et le délai de bout-en-bout « *end to end* » n'est pas important [25]. 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 II.8 illustre un exemple ou chaque paquet contient 4 trames.

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.

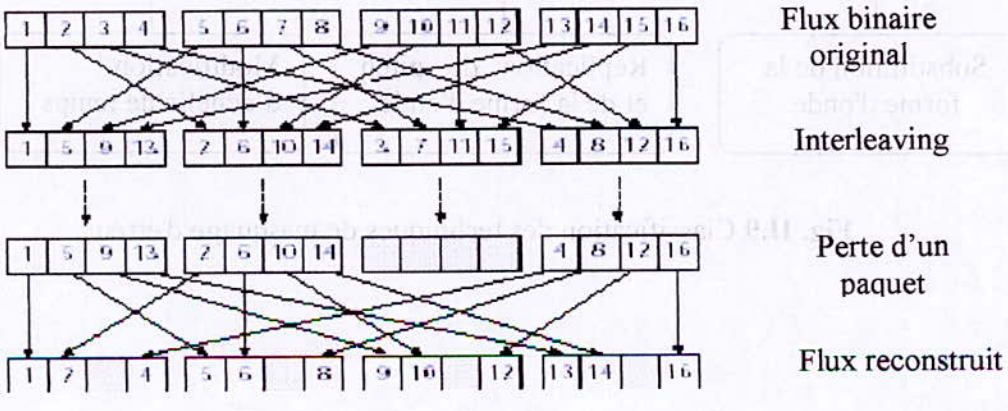

Fig. II.8 Exemple d'Interleaving [24]

II.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.

II.3.5.2 Masquage Basé sur le Récepteur

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

Le masquage d'erreur repose sur le principe de remplacer les paquets perdus par des paquets similaires aux originaux. Ceci reste possible du fait de la similarité à court terme du flux. La figure II.8 illustre les différentes techniques de masquage d'erreur.

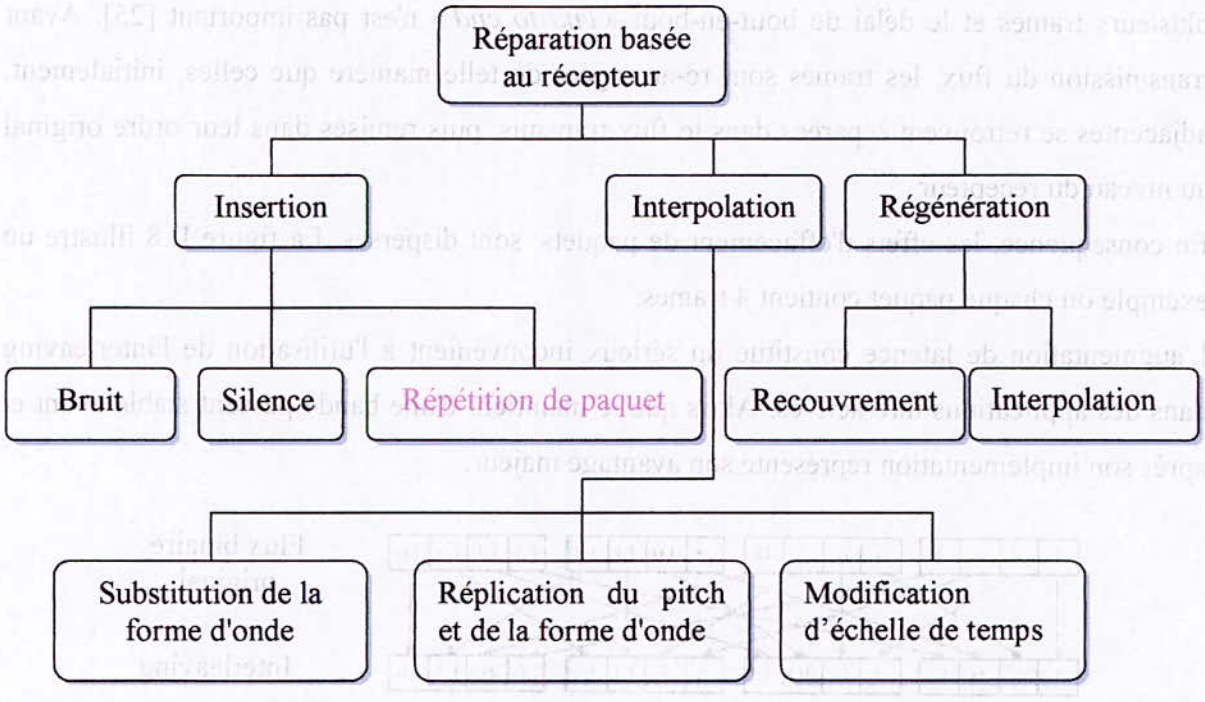

Fig. II.9 Classification des techniques de masquage d'erreur

II.2.5.2.1 L'Insertion

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

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

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

- Substitution par un bruit : Puisque la substitution de pertes par un silence présente de mauvaises performances, une autre méthode a été introduite. Elle consiste à remplacer la trame perdue par un bruit de fond. En plus, une fois comparée au silence, l'utilisation du bruit blanc a donné une qualité subjective meilleure et une intelligibilité améliorée [25].
- Répétition : Avec cette technique, les paquets perdus sont remplacés par la bonne donnée récupérée juste avant la perte.

II.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 les pertes [24].

- 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 [24].

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

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

II.2.5.2.3 La Régénération

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

Conclusion

Dans ce chapitre Nous avons vus l'architecture des réseaux ; les différentes couches et les protocoles utilisés et nous avons donnés aussi un aperçu sur la transmission de la voix via les réseaux *IP*.

Nous avons abordé aussi la qualité et les problèmes affectant cette qualité de service, à savoir les pertes de trames lors de la transmission et les différentes méthodes existant pour masquer ces pertes. Ces techniques peuvent être appliquées au niveau de l'émetteur ou du récepteur. Chaque technique présente une certaine complexité et requière des exigences liées à la méthode de masquage.

Chapitre III

Codeur de la norme G.729

Introduction

La Recommandation G.729 [20] décrit un algorithme pour le codage de signaux vocaux à 8 kbit/s au moyen de la prédiction linéaire à excitation par séquences codées à structure algébrique conjuguée (CS-ACELP) (*conjugate-structure-algebraic-code-excited-linear-prediction*).

Ce codeur est conçu pour fonctionner avec un signal numérique 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 [20][27].

III.1 Description général du Codec G.729

Le codeur de prédiction CS-ACELP est fondé sur le modèle de codage prédictif linéaire à excitation par code (CELP). Le codeur opère sur des trames vocales de 10 ms correspondantes à 80 échantillons à raison de 8000 échantillons par seconde. Toutes les trames de 10 ms, le signal est analysé afin d'en extraire les paramètres du modèle de prédiction CELP (coefficients du filtre de prédiction linéaire, index et gains de dictionnaire adaptatif et de dictionnaire fixe). Ces paramètres devront être codés et transmis. L'affectation des positions binaires aux paramètres de codage pour une trame est représentée dans la Table III.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 dictionnaire adaptatif	P1, P2	8	5	13
Parité du délai tonal	P0	1		1
Index de dictionnaire fixe	C1, C2	13	13	26
Signe de dictionnaire fixe	*S*1, *S*2	4	4	8
Gains de dictionnaire (étape 1)	*GA*1, *GA*2	3	3	6
Gains de dictionnaire (étape 2)	*GB*1, *GB*2	4	4	8
Total				80

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

III.1.1 Codeur

Le principe du codage est schématisé sur la figure III.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 dictionnaire fixe et par dictionnaire 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 dictionnaire 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 dictionnaire adaptatif et ce nouveau signal cible, $x'(n)$, est utilisé lors de l'exploration du dictionnaire fixe afin de déterminer l'excitation optimale. On fait appel à un répertoire algébrique de mots de 17 éléments binaires pour l'excitation par dictionnaire fixe. Les gains des contributions par dictionnaire adaptatif et par dictionnaire fixe sont quantifiés vectoriellement sur 7 éléments binaires (avec application au gain par dictionnaire fixe d'une prédiction par analyse à moyenne mobile). Finalement, les mémoires des filtres sont mises à jour au moyen du signal d'excitation ainsi déterminé.

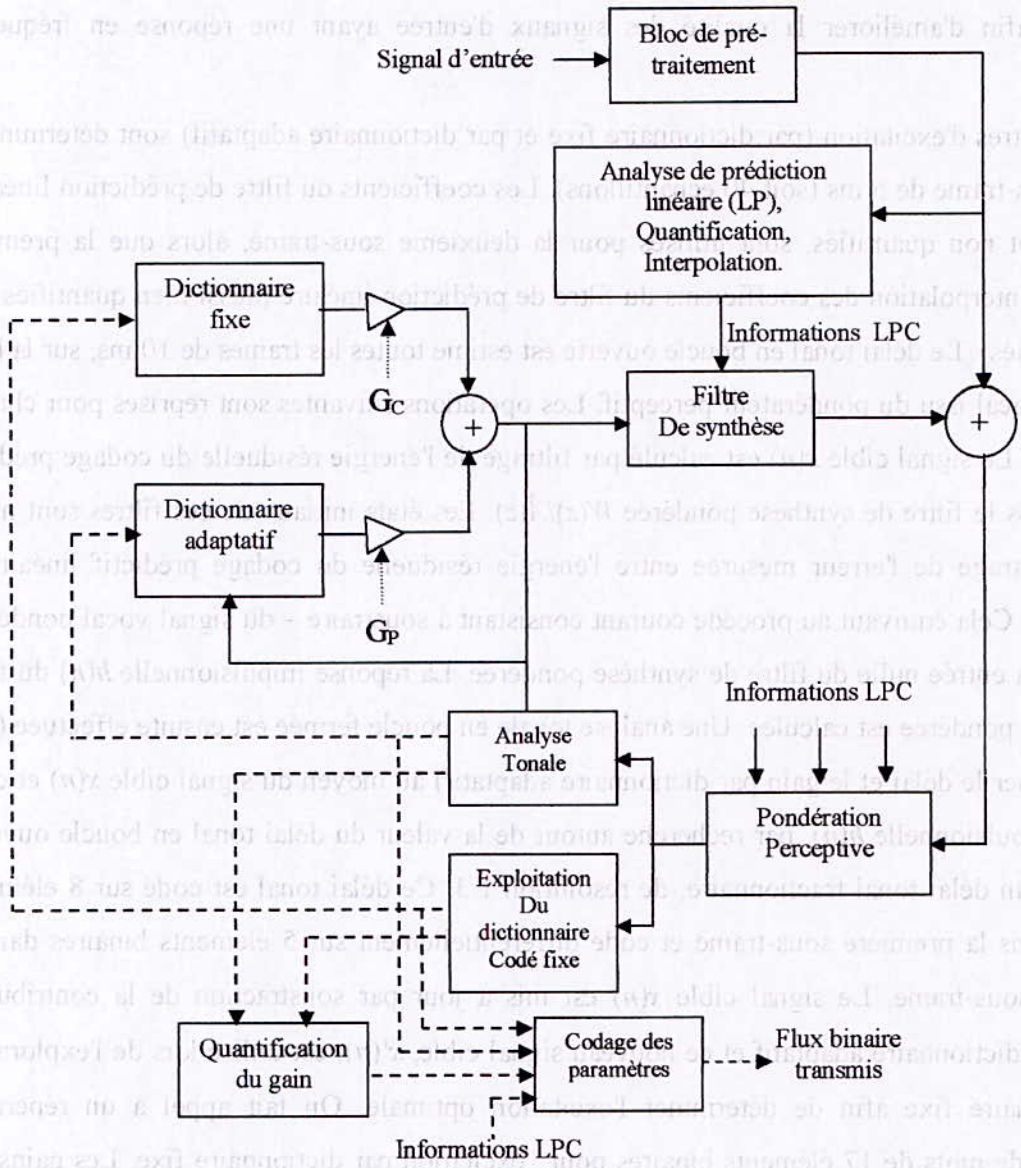

Fig. III.1 Principe du codeur *CS-ACELP G.729* [20]

III.1.2 décodeur

Le principe du décodeur est représenté sur la Figure III.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 tonals fractionnaires, les 2 vecteurs de dictionnaire fixe et les deux séries de gains par dictionnaire adaptatif et par dictionnaire 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 postfiltre adaptatif utilisant la sortie des filtres de synthèse à court et à long terme, suivi d'un filtre passe-haut et d'un échantillonneur-normalisateur.

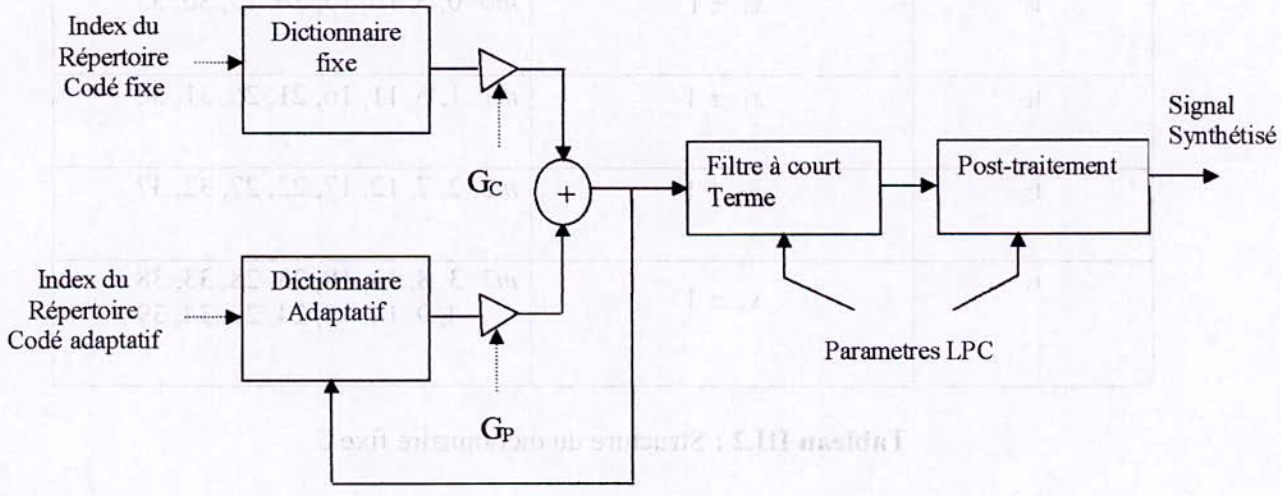

Fig. III.2 Principe du décodeur CS-ACELP *G.729* [20]

III.1.3 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.

III.1.4 Dictionnaire Fixe (Fixed codebook)

Le dictionnaire de séquences fixes est fondé sur une structure algébrique utilisant un modèle de permutation entrelacée d'impulsions (ISPP) (*interleaved single-pulse permutation*) de Dirac [20][27]. Dans ce dictionnaire, chaque vecteur contient 4 impulsions non nulles, dont chacune peut avoir l'amplitude +1 ou –1, avec les positions indiquées dans le Tableau III.2.

Impulsion	Signe	Positions
i_0	s_0: ± 1	$m0$: 0, 5, 10, 15, 20, 25, 30, 35
i_1	s_1: ± 1	$m1$: 1, 6, 11, 16, 21, 26, 31, 36
i_2	s_2: ± 1	$m2$: 2, 7, 12, 17, 22, 27, 32, 37
i_3	s_3: ± 1	$m3$: 3, 8, 13, 18, 23, 28, 33, 38 4, 9, 14, 19, 24, 29, 34, 39

Tableau III.2 : Structure du dictionnaire fixe $\mathcal{C}$

On construit le vecteur de dictionnaire fixe $c(n)$ en prenant un vecteur zéro de dimension 40 et en mettant les 4 impulsions unités (de Dirac) aux positions trouvées, multipliées avec leur signe correspondant.

III.2 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 [20] trouve le prédicateur à long terme pour lequel le gain de prédiction est supérieur à 3 dB. Pour cela, on fixe un seuil de 0,5 pour le carré de la corrélation normalisée. Pour le processus de masquage d'erreur, une trame de 10 ms est déclarée «périodique» si au moins une sous-trame de 5 ms possède un gain de prédiction à long terme supérieur à 3 dB, et dans ce cas seul le dictionnaire adaptatif est utilisé et la contribution du dictionnaire fixe est mise à zéro, Le délai tonal est fondé sur la partie entière du délai tonal contenu dans la trame précédente. Ce délai est répété pour chaque trame successive, Sinon, la trame actuelle est considérée également comme « apériodique » et la contribution du dictionnaire adaptatif est mise à zéro, la contribution du dictionnaire fixe est construite par sélection aléatoire d'un index de dictionnaire et d'un index de signe.

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

- répétition des paramètres du filtre de synthèse (les LSF).
- affaiblissement des gains de dictionnaire adaptatif et de dictionnaire fixe.
- affaiblissement de l'énergie mémorisée par le prédicteur de gain.
- production de l'excitation de remplacement.

III.2.1 Affaiblissement de gains de dictionnaire adaptatif et de dictionnaire fixe

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

$$g_c^{(m)} = 0{,}98 g_c^{(m-1)} \quad \text{(III.1)}$$

Où m est l'index de sous-trame.

Le gain de dictionnaire adaptatif est fondé sur une version affaiblie du precedent gain de dictionnaire adaptatif. Il est donné par:

$$g_p^{(m)} = 0{,}9 g_p^{(m-1)} \quad \text{avec la limite } g_p^{(m)} < 0{,}9 \quad \text{(III.2)}$$

III.2.2 Production de l'excitation de remplacement

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

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

Conclusion

Dans ce chapitre nous avons décrit le Codec *G.729* (CS-ACELP) le plus utilisé en *VoIP*, qui est une norme de codage numérique de la parole qui a été approuvé à l'UIT (Union International de Télécommunication) en Novembre 1995[27]. 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 IV

Simulations et Résultats

Introduction

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

La méthode de masquage d'erreurs introduite dans le G.729 ne donne pas de bons résultats, pour cela dans cette partie de simulation on va décrire et travailler sur des nouvelles techniques de dissimulation des trames perdues basée sur la répétition (figure IV.1).

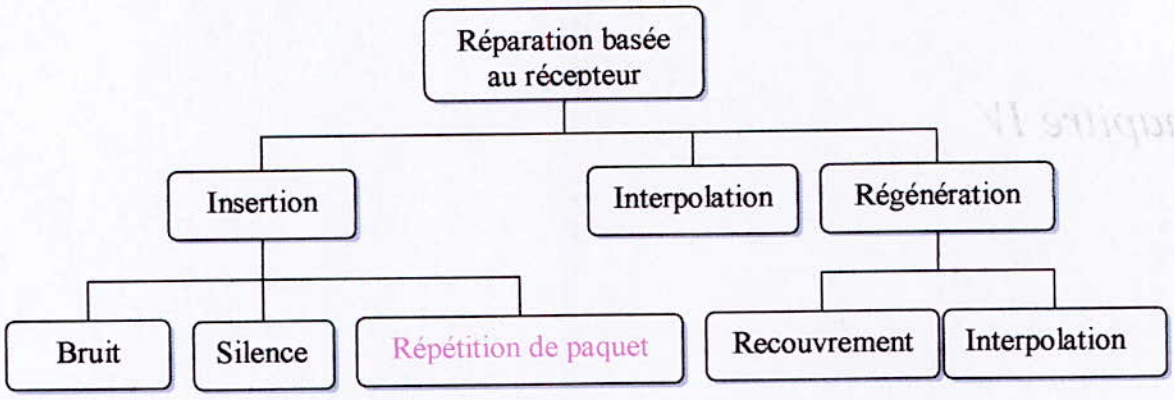

Fig. IV.1 Classification des techniques de masquage d'erreur

IV.1 Masquage des pertes dans le standard ITU G.729

Le codeur G.729 de l'ITU possède une procédure de traitement des trames effacées basée sur une méthode de masquage prédictive. Ce type de méthode n'introduit aucun délai supplémentaire car les paramètres des trames perdues seront récupérés à partir des bonnes trames précédentes.
Donc en cas d'effacement celles-ci se propagent aux trames suivantes. La figure IV.2 illustre ce phénomène.

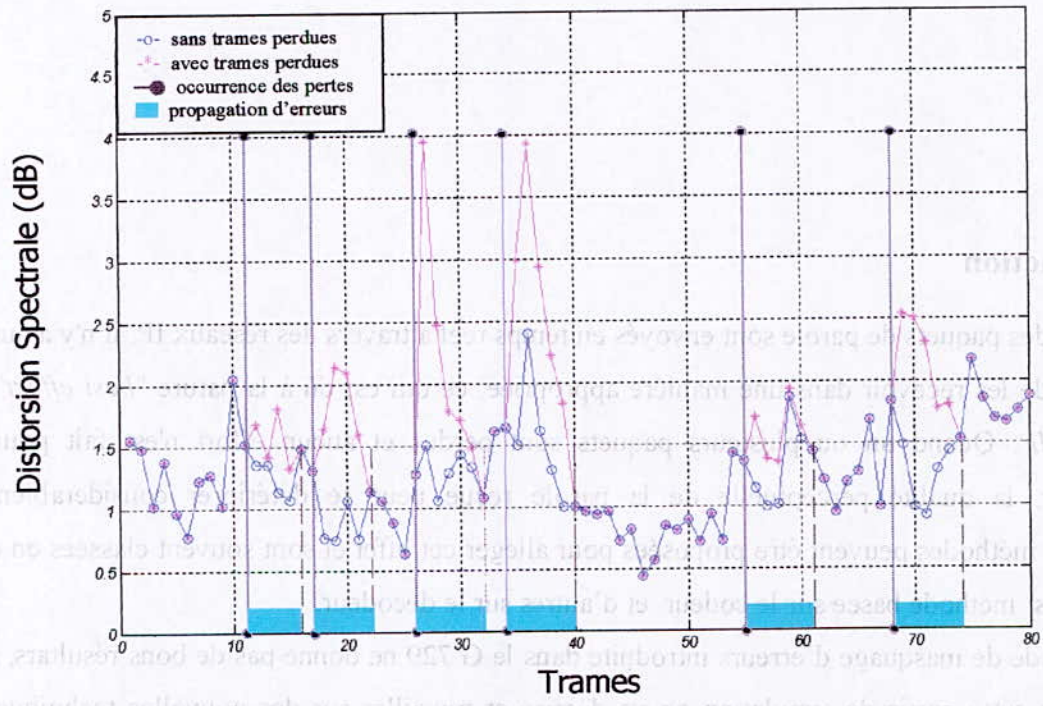

Fig.IV.2 Propagation de l'erreur de la distorsion spectrale dans le G.729 avec pertes de 10%

IV.2 Dissimulation basée sur la répétition

La figure IV.3 montre le schéma fonctionnel du décodeur G.729 contenant la méthode de dissimulation basée sur la répétition. Cette méthode ne requiert pas d'information future pour dissimuler les trames effacées.

Cette méthode comporte trois étapes; La première étape est le nouvel algorithme d'assourdissement (*Muting*) qui mute le signal d'excitation directement avec le facteur d'assourdissement *(Muting factor)* $g_e^{(n)}$ pour élaborer le signal graduellement, au lieu d'atténuer les gains des dictionnaires comme s'est le cas pour le standard G.729 en cas de dissimulation de pertes. La deuxième étape est le l'ajout d'une gigue aléatoire au pitch (*Pitch delay jittering)* qui ajoute une gigue aléatoire au pitch délai quand il y'a une succession de trames perdues et la troisième étape est l'expansion de la largeur de la bande LPC (*LPC bandwidth expansion*) on cas ou il y'a aussi une succession de trames perdus.

La méthode proposée ne s'applique pas seulement pour la reconstruction des trames effacées mais elle s'applique aussi pour les bonnes trames reçues après effacement.

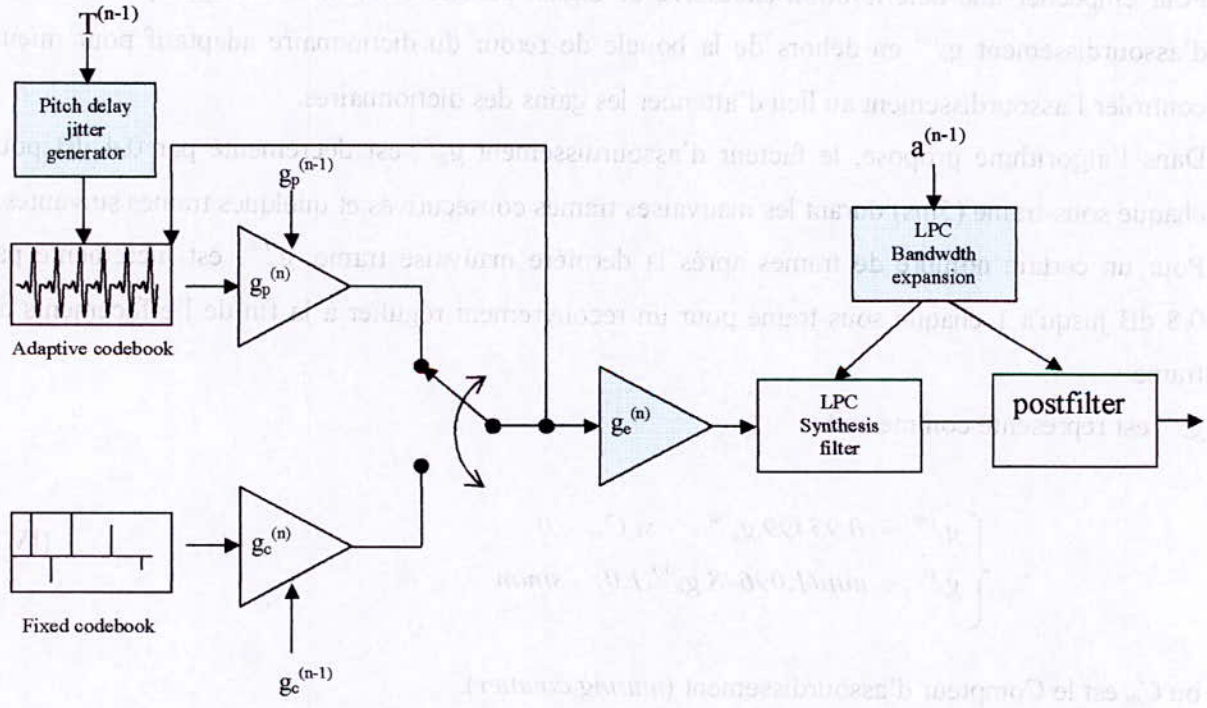

Fig.IV.3 Le décodeur G.729 avec l'algorithme proposé

IV.2.1 Assourdissement du signal d'excitation

Dans la figure IV.3, $g_p^{(n)}$ et $g_c^{(n)}$ sont respectivement les gains des dictionnaires adaptatifs et fixes de la trame courante. Le G.729 utilise une version atténuée des gains $g_p^{(n-1)}$ et $g_c^{(n-1)}$ comme étant les gains courantes des dictionnaires $g_p^{(n)}$ et $g_c^{(n)}$ pour une mauvaise trame tel que :

$$\begin{cases} g_p^{(n)} = 0.9 g_p^{(n-1)} \quad avec \quad g_p^{(n)} < 0.9 \\ g_c^{(n)} = 0.98 g_c^{(n-1)} \end{cases} \tag{IV.1}$$

(n) : indice de la sous-trame

Le gain du dictionnaire adaptatif dans la dernière bonne trame est atténué durant la période d'effacement de trames. Même après l'effacement de trame est fini, le signal parole est détérioré dans les trames suivantes. Cela est dû au dictionnaire adaptatif qui est mis à jour avec une version atténuée du gain, donc l'atténuation se propage aux trames suivantes.

Pour empêcher une détérioration excessive du signal parole, cette méthode applique le facteur d'assourdissement $g_e^{(n)}$ en dehors de la boucle de retour du dictionnaire adaptatif pour mieux contrôler l'assourdissement au lieu d'atténuer les gains des dictionnaires.

Dans l'algorithme proposé, le facteur d'assourdissement $g_e^{(n)}$ est décrémenté par 0.4 dB pour chaque sous-trame (5ms) durant les mauvaises trames consécutives et quelques trames suivantes. Pour un certain nombre de trames après la dernière mauvaise trame, $g_e^{(n)}$ est incrémenté par 0.8 dB jusqu'à 1 chaque sous-trame pour un recouvrement régulier à la fin de l'effacements de trame.

$g_e^{(n)}$ est représenté comme suit :

$$\begin{cases} g_e^{(n)} = 0.95499\, g_e^{(n)} & si\ C_m > 0 \\ g_e^{(n)} = min(1.09648\, g_e^{(n)}, 1.0) & sinon \end{cases} \tag{IV.2}$$

où C_m est le Compteur d'assourdissement (*muting counter*).

C_m est mis à 4 si on a des trames mauvaises successives et décrémenté de 1 si $g_p^{(n)} < 1.0$ et la trame courante est bonne.

Les gains des dictionnaires $g_p^{(n)}$ et $g_c^{(n)}$ sont simplement répétés dans les mauvaise trames au lieu d'êtres atténues. Cependant, la limite maximum g_{pmax} est mise à $g_p^{(n)}$ pour empêcher une montée imprévisible de l'énergie du signal d'excitation.

g_{pmax} est défini comme suit :

$$g_{pmax}=max(1.2-0.1(C_b-1),0.8) \quad \text{(IV.3)}$$

où C_b est le nombre de mauvaises trames successives.

IV.2.2 Ajout d'une gigue au délai du pitch

Dans la dissimulation du standard G.729, le délai du pitch dans une mauvaise trame est le pitch précèdent incrémenté par un (1.0) pour imiter l'évolution du pitch pour un signal parole.

La méthode standard du G.729 évite de reproduire un signal excessivement périodique dans une mauvaise trame, mais il peut accumuler l'erreur d'estimation du délai du pitch dans les mauvaises trames consécutives. L'algorithme proposé par [33] répète le délai du pitch dans les mauvaises trames, mais une gigue aléatoire de 3% est ajouté au délai du pitch répété dans les mauvaises trames consécutives pour éviter une reproduction excessive de la périodicité dans le signal sans accumulation de l'erreur d'estimation.

IV.2.3 Expansion de la bande passante LPC

Dans le décodeur G.729, les paramètres LPC dans la dernière bonne trame sont répètes dans les mauvaises trames. Cependant, il peut résulter dans une qualité synthétique du signal parole si le spectre LPC dans la dernière bonne trame contient un formant aigu. Pour empêcher ce problème, l'algorithme proposé étend la bande passante progressivement dans les mauvaises trames consécutives si et seulement si le minimum de la bande passante des LSF dans le dernière bonne trame est inférieur à 100Hz. le facteur d'expansion de la bande passante $\gamma^{(n)}$ est mis à jour. Comme suit :

$$\gamma^{(n)}=\max(0.95\gamma^{(n-1)},0.8) \quad \text{(IV.4)}$$

où $\gamma^{(n)}$ et $\gamma^{(n-1)}$ sont les facteurs d'expansion de la bande passante LPC courant et précèdent. Le facteur $\gamma^{(n)}$ est appliqué aux paramètres LPC dans les dernières mauvaises trames, où le décodeur

reçoit les bonnes trames après une mauvaise trame (effacement de trame), $\gamma^{(n)}$ est progressivement incrémenté pour un recouvrement régulier d'un effacement de trame :

$$\gamma^{(n)} = \min(1.05\,\gamma^{(n-1)}, 1.0) \qquad \text{(IV.5)}$$

IV.3 Interpolation de l'excitation

Traditionnellement, les signaux d'excitation sont interpolés en se basant sur la décision voisé/non voisé (V/UV) de la trame précédente et seulement un des deux contributions des dictionnaires adaptatif ou fixe est récupéré.

La procédure est :

- Obtenir la décision de voisement de la trame précédente ;
- Si la trame précédente est voisée, mettre la contribution du dictionnaire fixe en zéro et utiliser l'information du pitch, appliquer le filtre du pitch atténué pour obtenir l'excitation courante.
- Si la trame précédente est non voisée, mettre la contribution du dictionnaire adaptatif à zéro, utiliser l'information du gain précédent, remplacer les signaux d'excitation par une séquence de nombres aléatoires normalise par le gain atténué.

Cet algorithme fonctionne pour les trames non voises mieux que pour les trames voisées. Pour les trames voises, nous avons observé qu'il y a une certaine structure de périodicité dans les signaux d'excitation dans le dictionnaire fixe. Car le remplacement de la contribution du dictionnaire fixe par des zéros n'exploite pas une structure pareille. De plus nous avons observés que cette structure peut être représenté par un « échantillon » en terme de la position des impulsions et le gain qui sont relies entre eux. Par conséquent, nous pouvons reconstruire la trame du dictionnaire fixe par recherche de plus proche échantillon pareil reflété par son trames voisines et remplacer les signaux perdus par la trame correspondante dans l'échantillon pareil.

IV.4 Simulations et résultats

IV.4.1 Base de données utilisées

Une bonne base donnée reste la condition sine qua non pour la validation d'un quelconque résultat trouvé. Dans nos travaux, nous avons utilisé une base de données mondialement

reconnue, à savoir, « the DARPA TIMIT Acoustic-Phonetic Continous Speech Corpus (TIMIT), Training and test data ».

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

- Répartition des orateurs :

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

Dialect	Male	Female	Total
1	31 (63%)	18 (27%)	49 (08%)
2	71 (70%)	31 (30%)	102 (16%)
3	79 (67%)	23 (23%)	102 (16%)
4	69 (69%)	31 (31%)	100 (16%)
5	62 (62%)	36 (37%)	98 (07%)
6	30 (65%)	16 (35%)	46 (16%)
7	74 (74%)	26 (26%)	100 (16%)
8	22 (67%)	11 (33%)	33 (05%)
Total	438 (70%)	192 (30%)	630 (100%)

Tableau IV.1 Répartition des orateurs

IV.4.2 Le Modèle du Réseau

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

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

Cette probabilité est également connue comme *la probabilité conditionnelle de perte (CLP).*

Soient P_0 et P_1 les probabilités pour rester dans l'état 0 et l'état 1 respectivement. Nous avons

$$P_0 = P_0.p + P_1.q \tag{IV.6}$$

$$p_0 = \frac{1-q}{p+1-q} \qquad p_1 = \frac{p}{p+1-q} \tag{IV.7}$$

La probabilité pour qu'un paquet soit abandonné sans connaître si le paquet précédent est livré ou abandonné, c'est à dire. *La probabilité de perte sans conditions* (*ULP*) est exactement la probabilité pour que le modèle du réseau soit dans l'état 1 (P1). La figure IV.4 présente le modèle de Markov avec ses probabilités de transition et le tableau IV.2 cite les taux de perte utilisés dans notre simulation.

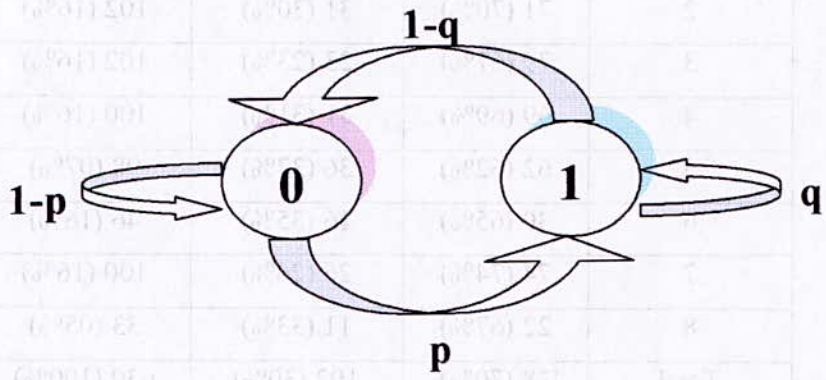

Fig.IV.4 Pertes de paquets modélisées par un processus aléatoire de Markov

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

Tableau IV.2 Les taux de pertes simulés

IV.4.3 Procédure de masquage implémenté

Le processus complet de masquage peut être résumé comme suit :

Si une trame est déclarée perdue :
1. Répétition des paramètres du filtre de synthèse, avec expansion de la bande passante LPC si on détecte un formant aigu ;
2. affaiblissement des gains du dictionnaire adaptatif et du dictionnaire fixe par le facteur *le facteur d'assourdissement* ;
3. En se basant sur la bonne trame précédente, prendre une décision sur le type de la trame (voisée ou non voisée);
4. Si la trame précédente est voisée : Mettre la contribution du dictionnaire fixe à zéro; sinon on met la contribution du dictionnaire adaptatif à zéro ;
5. répéter le pitch de la dernière bonne trame en lui ajoutant un pourcentage d'une gigue aléatoire.

IV.4.4 Résultats de l'Implémentation de la méthode de dissimulation basée sur la répétition au Standard G.729

IV.4.4.1 Assourdissement du signal d'excitation

Nous allons dans ce qui suit voir l'influence du facteur d'assourdissement sur les performance du G.729.

Pour cela nous simulerons dans un premier lieu une voix féminine en lui appliquant un pourcentage de pertes de 10%.

Les résultats obtenus sont représentés par la figure IV.5-(a,b et c).

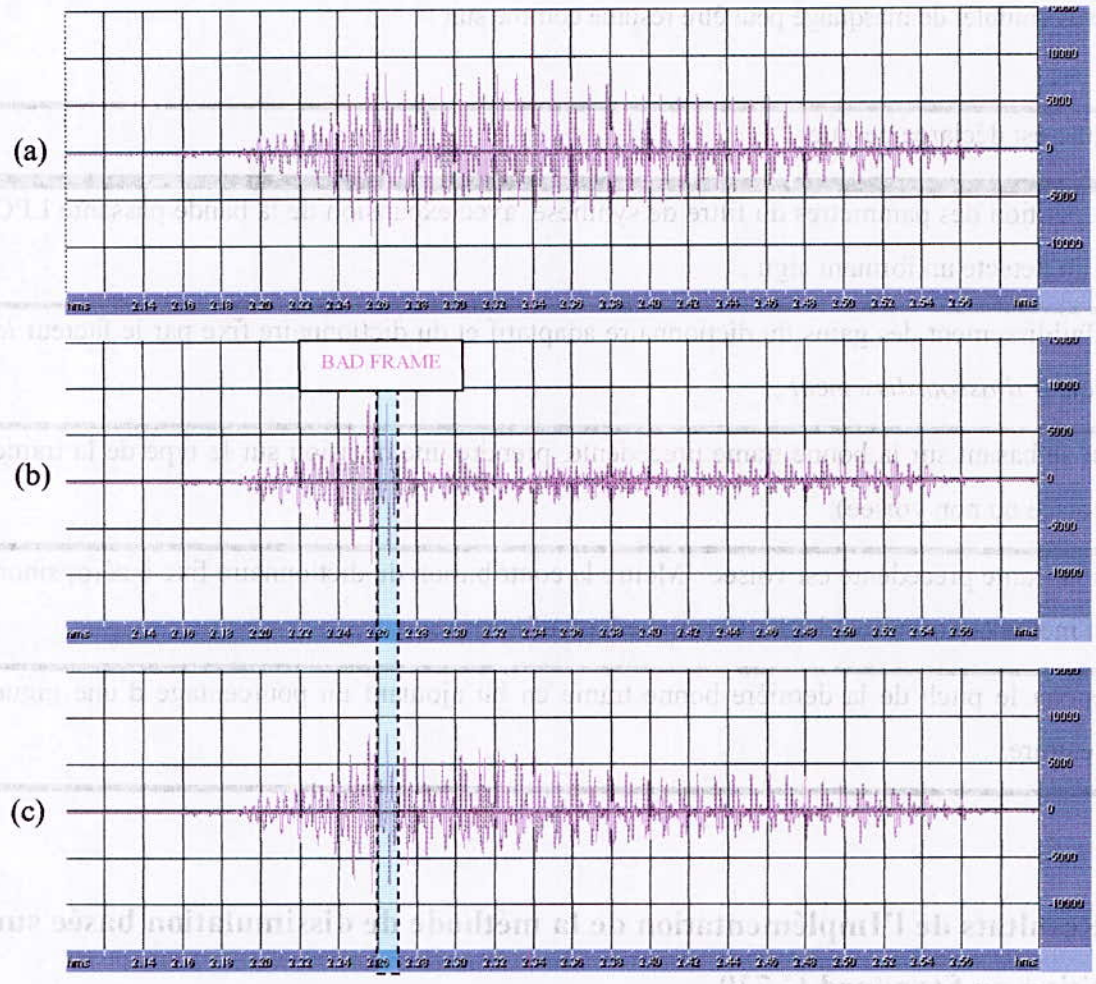

Fig.IV.5 signal parole synthétisé avec la méthode proposée et la méthode standard

(a)- signal parole synthétisé sans pertes

(b)- signal parole synthétisé avec pertes de 10% avec le standard G.729

(c)- signal parole synthétisé avec perte de 10% utilisant l'algorithme

On remarque que la forme d'onde du signal parole synthétisé avec la méthode du standard (figure IV.5-(b)) avec pertes de 10% est détériorée après l'effacement de trames présenté par le champ vert.

La forme d'onde obtenue par l'implémentation de l'assourdissement du signal d'excitation présente une amélioration en terme d'amplitude après l'effacement de trames.

Pour l'évaluation des performances de cette méthode nous avons appliqué la distorsion spectrale donnée par l'équation :

$$DS_i = \sqrt{\frac{1}{n_1 - n_0}\sum_{k=n_0}^{n_1-1}\left[10\log\frac{S_i(e^{j2\pi k/N})}{\widetilde{S}_i(e^{j2\pi k/N})}\right]^2} \quad (dB) \qquad \text{(IV.8)}$$

Les résultats obtenus sont représentés par les figures et les tableaux suivants :

Taux de pertes (%)	G.729			Assourdissement du signal d'excitation		
	Av.SD (dB)	Outliers (%)		Av.SD (dB)	Outliers (%)	
		2-4 dB	> 4dB		2-4dB	>4 dB
0	1.26	9.82	0.10	1.26	9.82	0.10
10	1.89	27.90	2.95	1.65	21.71	2.36
20	2.25	40.67	6.48	1.98	34.68	5.40
30	2.56	48.62	11.10	2.38	43.22	9.82
40	2.88	48.04	17.78	2.75	42.83	16.11

Tableau IV.3 Distorsion spectrale moyenne pour du facteur d'assourdissement pou une voix féminine

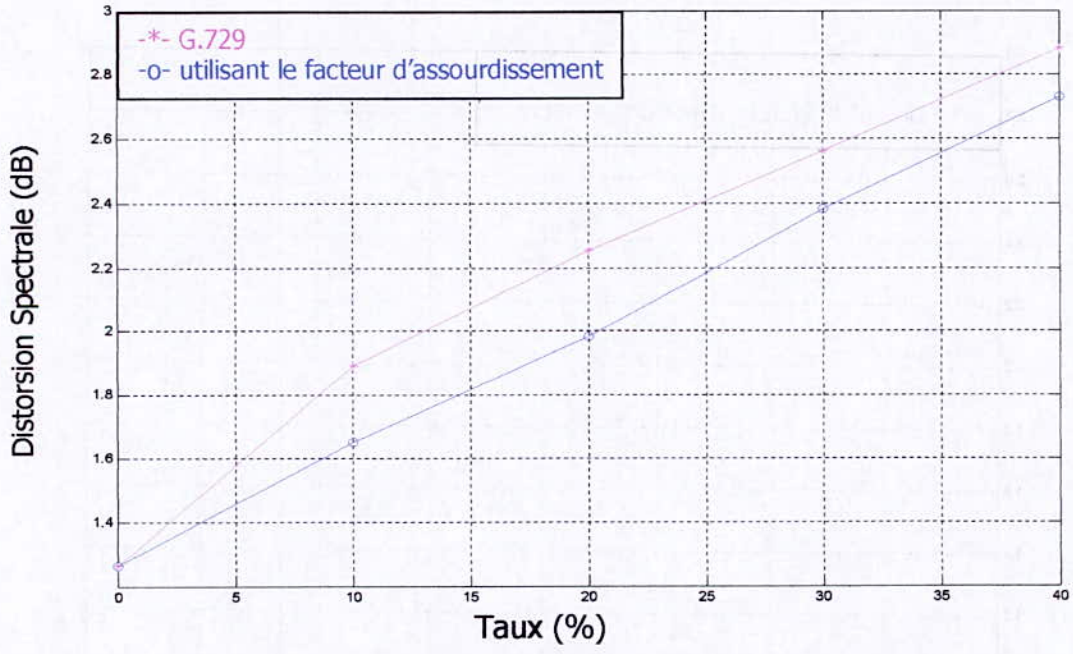

Fig.IV.6 Distorsion spectrale pour l'assourdissement du signal d'excitation pour une voix féminine

D'après le tableau IV.3 et la figure IV.6 on remarque que quelque soit le taux de perte utilisé dans la simulation, on obtient une distorsion moyenne plus faible pour l'algorithme d'assourdissement, la plage d'amélioration de la distorsion varie entre 0.13 dB et 0.27 dB pour une voix féminine.

L'application de l'assourdissement du signal d'excitation nous a permet d'obtenir les résultats suivants :

Taux de pertes (%)	G.729			Assourdissement du signal d'excitation		
	Av.SD (dB)	Outliers (%)		Av.SD (dB)	Outliers (%)	
		2-4 dB	> 4dB		2-4dB	>4 dB
0	1.13	5.77	0	1.13	5.77	0
10	1.60	17.03	1.88	1.47	15.30	1.56
20	1.91	28.14	4.62	1.79	25.54	4.33
30	2.13	36.36	6.64	2.03	32.76	6.64
40	2.46	40.12	10.68	2.36	37.66	9.81

Tableau IV.4 Distorsion spectrale moyenne pour le facteur d'assourdissement pour une voix masculine

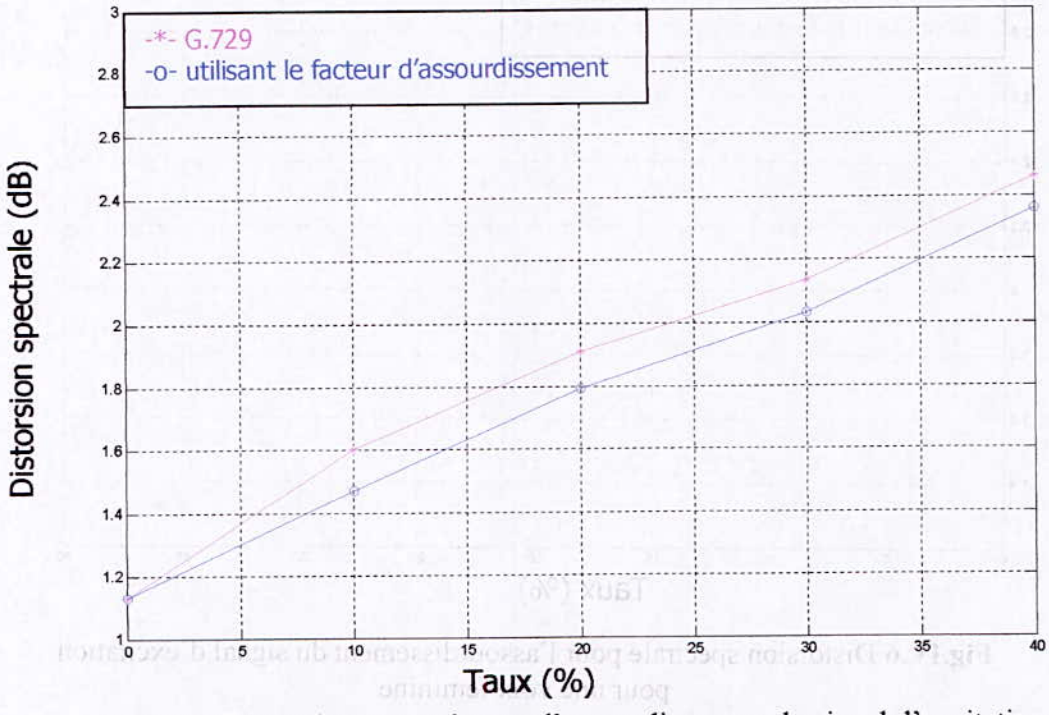

Fig.IV.7 Distorsion spectrale pour l'assourdissement du signal d'excitation pour une voix masculine

Le tableau IV.4 montre que la plage d'amélioration de la distorsion varie entre 0.10 dB et 0.13 dB pour une voix masculine.

D'après les figures et les tableaux précédents, on remarque que l'implémentation de l'algorithme d'assourdissement du signal d'excitation permet d'avoir une amélioration de la distorsion spectrale, car le facteur d'assourdissement (*muting factor*) nous a permet d'éviter la mise à jours du dictionnaire adaptatif par des versions atténuées du gain.

IV.4.4.2 Ajout de 3% d'une gigue aléatoire au délai du pitch

Nous allons maintenant appliquer l'ajout de 3% d'une gigue aléatoire, l'évaluation des performances est faite également par la distorsion spectrale. Les résultats sont représentés dans le tableau IV.5.

Taux de pertes (%)	G.729			Ajout de 3% d'une gigue aléatoire au délai du pitch		
	Av.SD (dB)	Outliers (%)		Av.SD (dB)	Outliers (%)	
		2-4 dB	> 4dB		2-4dB	>4 dB
0	1.26	9.82	0.10	1.26	9.82	0.10
10	1.89	27.90	2.95	1.72	22.69	2.46
20	2.25	40.67	6.48	2.07	35.95	5.40
30	2.56	48.62	11.10	2.41	43.81	10.12
40	2.88	48.04	17.78	2.76	43.91	16.21

Tableau IV.5 Distorsion spectrale pour l'ajout d'une gigue aléatoire au délai du pitch pour une voix féminine

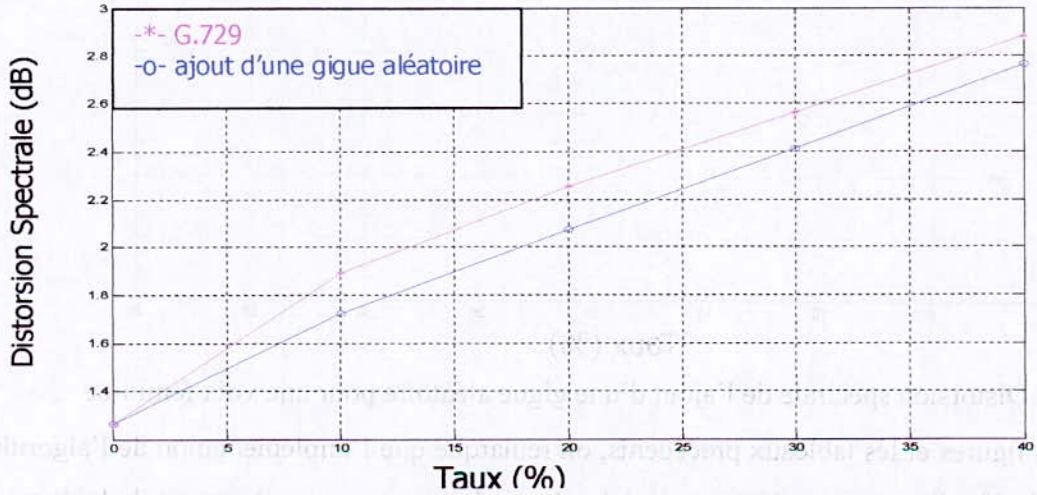

Fig.IV.8 Distorsion spectrale de l'ajout d'une gigue aléatoire pour une voix féminine

D'après le tableau IV.5 et la figure IV.8 on remarque que quelque soit le taux de perte utilisé dans la simulation, on obtient une distorsion moyenne plus faible pour l'algorithme de l'ajout de 3% d'une gigue aléatoire au délai du pitch, la plage d'amélioration de la distorsion varie entre 0.14 dB et 0.18 dB pour une voix féminine.

L'implémentation de cet algorithme en utilisant une voix masculine donne le tableau de distorsion suivant :

Taux de pertes (%)	G.729			Ajout d'une gigue aléatoire de délai du pitch		
	Av.SD (dB)	Outliers (%)		Av.SD (dB)	Outliers (%)	
		2-4 dB	> 4dB		2-4dB	>4 dB
0	1.13	5.77	0	1.13	5.77	0
10	1.60	17.03	1.88	1.51	16.16	1.59
20	1.91	28.14	4.62	1.83	26.55	4.33
30	2.13	36.36	6.64	2.06	33.33	6.35
40	2.46	40.12	10.68	2.39	38.24	10.10

Tableau IV.6 Distorsion spectrale pour l'ajout d'une gigue aléatoire au délai du pitch pour une voix masculine

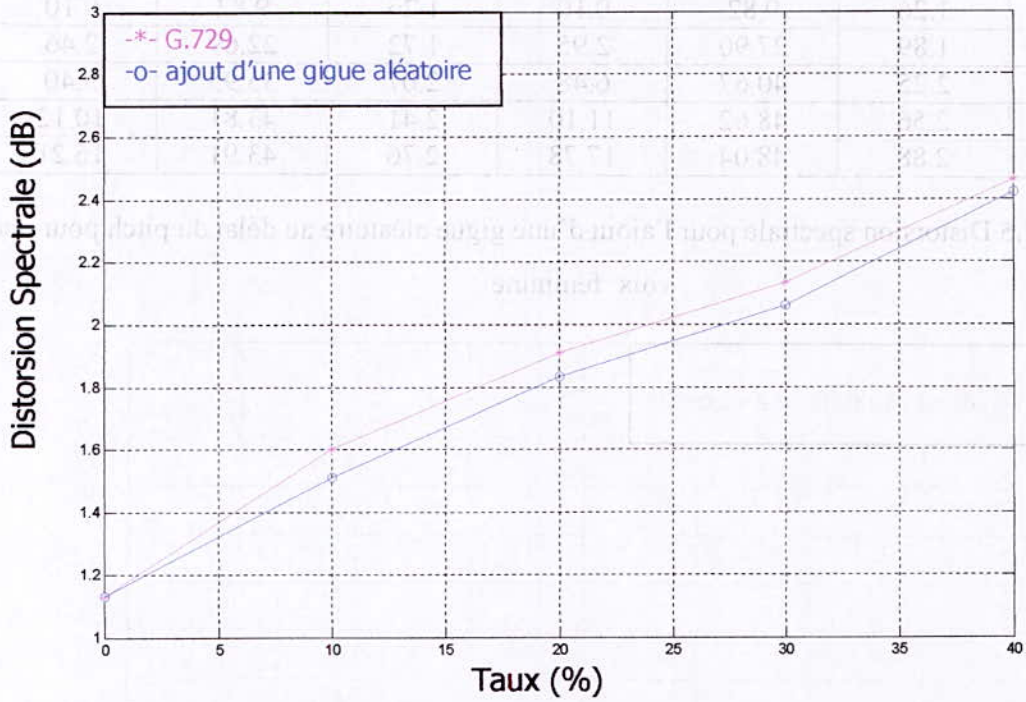

Fig.IV.9 Distorsion spectrale de l'ajout d'une gigue aléatoire pour une voix féminine

D'après les figures et les tableaux précédents, on remarque que l'implémentation de l'algorithme de l'ajout de 3% d'une gigue aléatoire au délai du pitch donne une amélioration de la distorsion

spectrale, cela est dû au faite que cette méthode permet d'incrémenté le pitch sansaccumulation des erreurs dans les trames consécutives, la plage d'amélioration de la distorsion varie entre 0.07 dB et 0.09 dB pour une voix masculine.

IV.4.4.3 Expansion de la bande passante LPC

L'implémentation de cet algorithme nous a permet d'obtenir les résultats suivants :

Taux de pertes (%)	G.729			expansion de la bande passante LPC		
	Av.SD (dB)	Outliers (%)		Av.SD (dB)	Outliers (%)	
		2-4 dB	> 4dB		2-4dB	>4 dB
0	1.26	9.82	0.10	1.26	9.81	0.10
10	1.89	27.90	2.95	1.76	23.38	2.55
20	2.25	40.67	6.48	2.11	37.13	5.70
30	2.56	48.62	11.10	2.44	45.19	10.02
40	2.88	48.04	17.78	2.83	45.28	16.40

Tableau IV.7 Distorsion spectrale pour l'expansion de la bande passante LPC pour une voix féminine

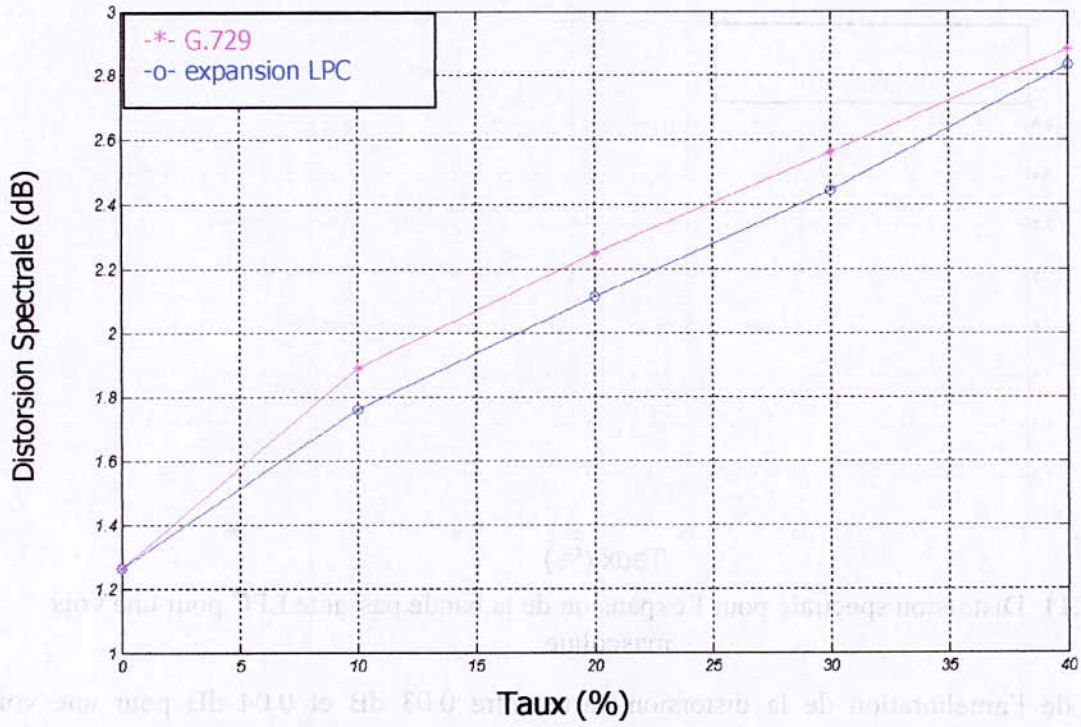

Fig.IV.10 Distorsion spectrale pour l'expansion de la bande passante LPC pour une voix féminine

Les valeurs de la distorsion spectrale pour l'algorithme de l'expansion de la bande passante LPC sont faible que ceux obtenus avec l'algorithme de dissimulation du standard G.729, car ce dernier répète les paramètres LPC de la dernière bonne reçue dans les mauvaises trames trame ce qui provoque une répétions des formants aigus, la méthode proposée permet d'éviter la répétition de ces formants par une expansion de la bande passante, la plage d'amélioration de la distorsion varie entre 0.05 dB et 0.14 dB pour une voix féminine.

Taux de pertes (%)	G.729			expansion LPC		
	Av.SD (dB)	Outliers (%)		Av.SD (dB)	Outliers (%)	
		2-4 dB	> 4dB		2-4dB	>4 dB
0	1.13	5.77	0	1.13	5.77	0
10	1.60	17.03	1.88	1.56	16.88	1.73
20	1.91	28.14	4.62	1.88	27.71	4.47
30	2.13	36.36	6.64	2.10	34.78	6.49
40	2.46	40.12	10.68	2.43	39.68	10.25

Tableau IV.8 Distorsion spectrale pour l'expansion de la bande passante LPC pour une voix masculine

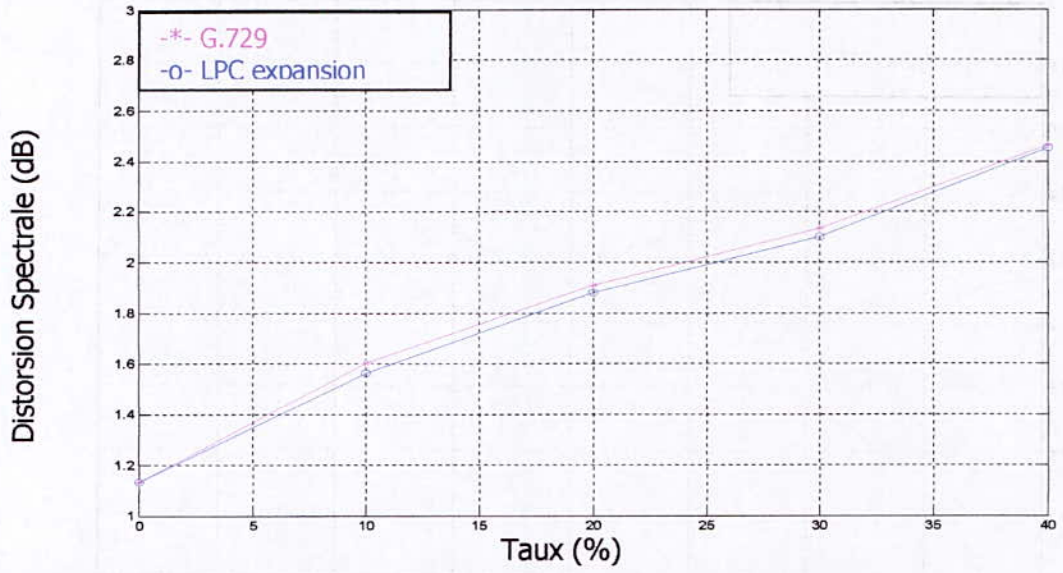

Fig.IV.11 Distorsion spectrale pour l'expansion de la bande passante LPC pour une voix masculine

La plage de l'amélioration de la distorsion varie entre 0.03 dB et 0.04 dB pour une voix masculine.

D'après les figures et les tableaux précédents, on remarque que l'implémentation de l'algorithme d'expansion de la bande passante LPC, apporte une amélioration remarquable pour les femmes par rapport hommes. Cela est dû aux caractéristiques de la voix féminine qui contient plus de formants aigus qu'une voix masculine.

IV.4.4.4 dissimulation basée sur la répétition

Après avoir tester l'efficacité de chaque caractéristique seul, nous avons appliqué les trois ensemble (assourdissement du signal d'excitation, ajout de 3% d'une gigue aléatoire et l'expansion de la bande passante LPC) en vu d'une meilleur amélioration des performances pour cela nous avons pris des signaux de parole pour une voix féminine et masculine, les performances obtenus sont représentés sur les tableaux et les figures suivants :

Taux de pertes (%)	G.729			Repetition-Based Concealment		
	Av.SD (dB)	Outliers (%)		Av.SD (dB)	Outliers (%)	
		2-4 dB	> 4dB		2-4dB	>4 dB
0	1.26	9.82	0.10	1.26	9.81	0.10
10	1.89	27.90	2.95	1.62	20.24	2.36
20	2.25	40.67	6.48	1.95	33.01	5.40
30	2.56	48.62	11.10	2.36	41.75	9.72
40	2.88	48.04	17.78	2.72	42.24	15.91

Tableau IV.9 Distorsion spectrale de la méthode de dissimulation basée sur la répétition pour une voix féminine

Taux de pertes (%)	G.729			Repetition-Based Concealment		
	Av.SD (dB)	Outliers (%)		Av.SD (dB)	Outliers (%)	
		2-4 dB	> 4dB		2-4dB	>4 dB
0	1.13	5.77	0	1.13	5.77	0
10	1.60	17.03	1.88	1.44	14.86	1.44
20	1.91	28.14	4.62	1.76	24.39	4.18
30	2.13	36.36	6.64	2.01	32.76	6.35
40	2.46	40.12	10.68	2.34	37.66	9.67

Tableau IV.10 Distorsion spectrale de la méthode de dissimulation basée sur la répétition pour une voix masculine

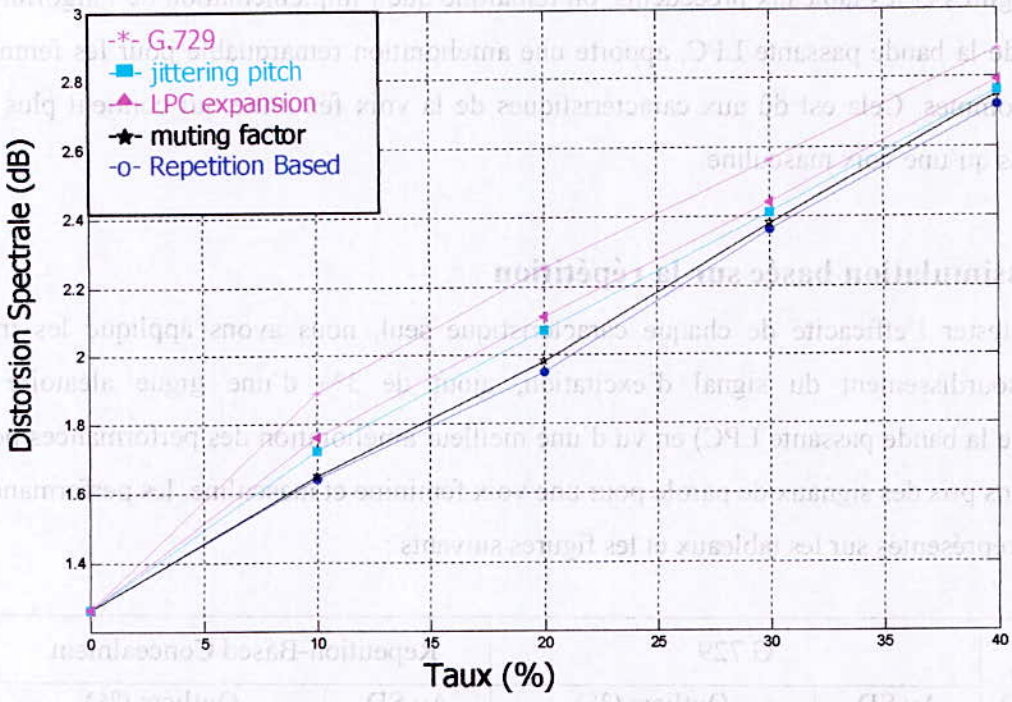

Fig.IV.12 Distorsion spectrale de la méthode de dissimulation basée sur la répétition pour une voix féminine

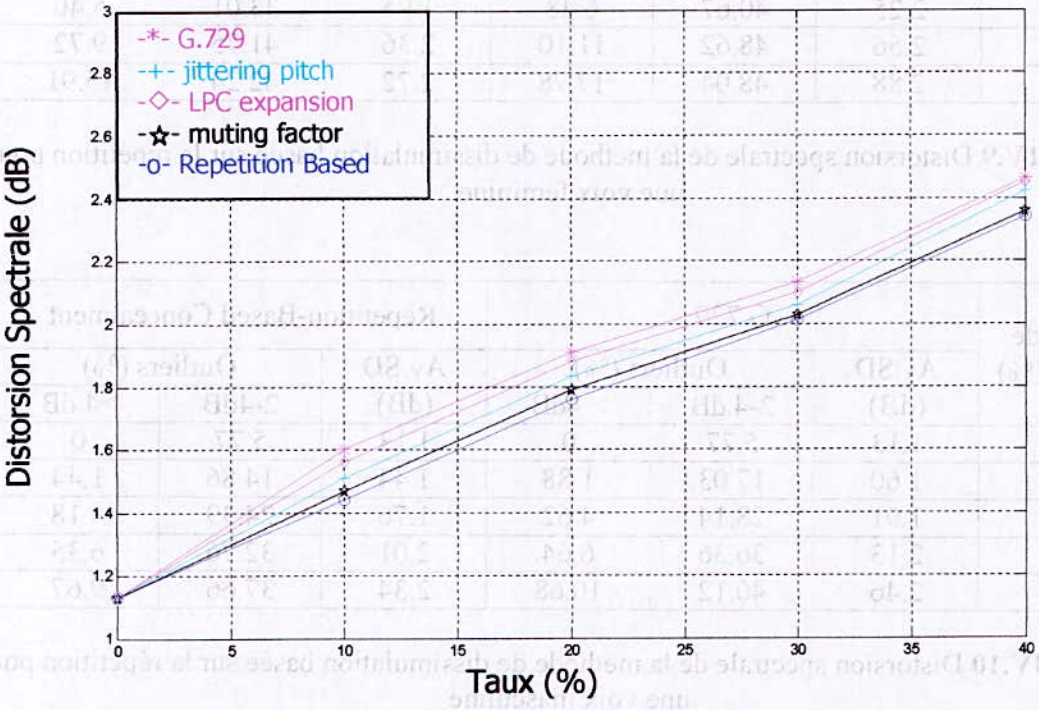

Fig.IV.13 Distorsion spectrale de la méthode de dissimulation basée sur la répétition pour une voix masculine

La plage de l'amélioration de la distorsion varie entre 0.16 dB et 0.27 dB pour une voix féminine alors qu'elle varie de 0.12 dB jusqu'à 0.16 dB pour une voix masculine

D'après les plages d'amélioration de la distorsion, on constate que la méthode de la dissimulation basée sur la répétition apporte une amélioration en terme de distorsion spectrale par rapport au standard G.729, elle est plus efficace pour les voix féminines que pour les voix masculines, à cause des caractéristique de la voix féminine.

Pour montrer encore l'efficacité de cette méthode on a utilisé une mesure de distorsion appelée perceptuelle (voir paragraphe I.4.3) tel que l'EMBSD qui a donnée les résultats suivants :

Taux (%)	EMBSD (méthode standard)	EMBSD (méthode proposée)
0	0.420	0.420
10	0.742	0.570
20	1.707	1.672
30	3.101	2.685
40	4.420	3.521

Tableau IV.11 L'EMBSD pour une voix féminine

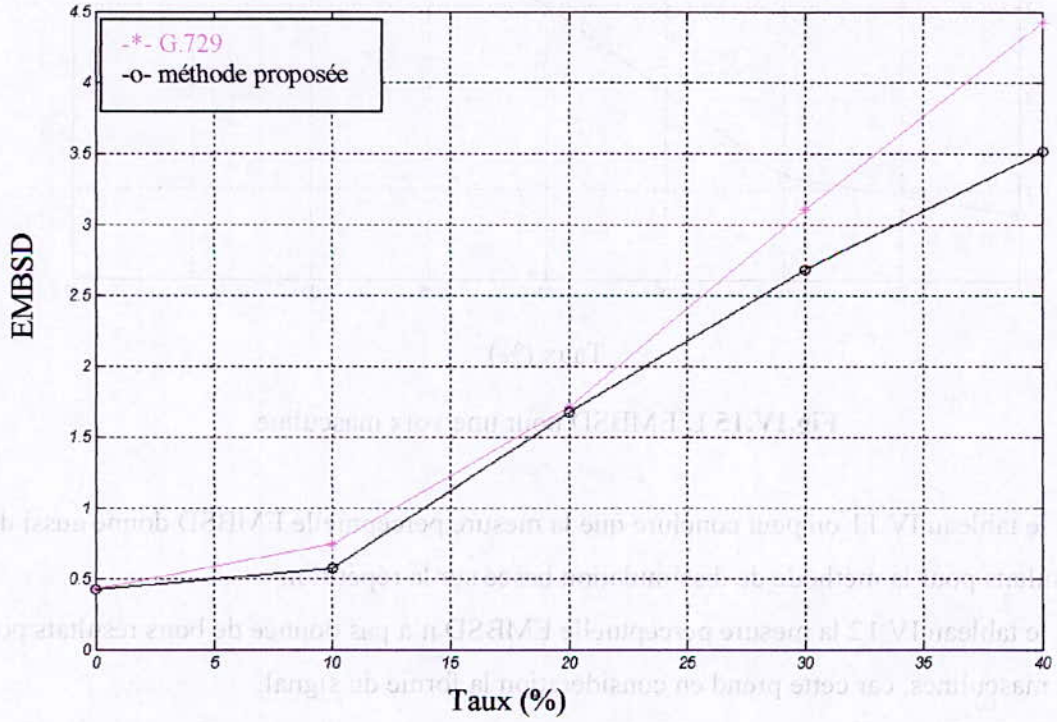

Fig.IV.14 L'EMBSD pour une voix féminine

Taux (%)	EMBSD (méthode standard)	EMBSD (méthode proposée)
0	1.023	1.023
10	1.119	1.071
20	1.910	1.992
30	3.242	2.705
40	3.144	3.600

Tableau IV.12 L'EMBSD pour une voix masculine

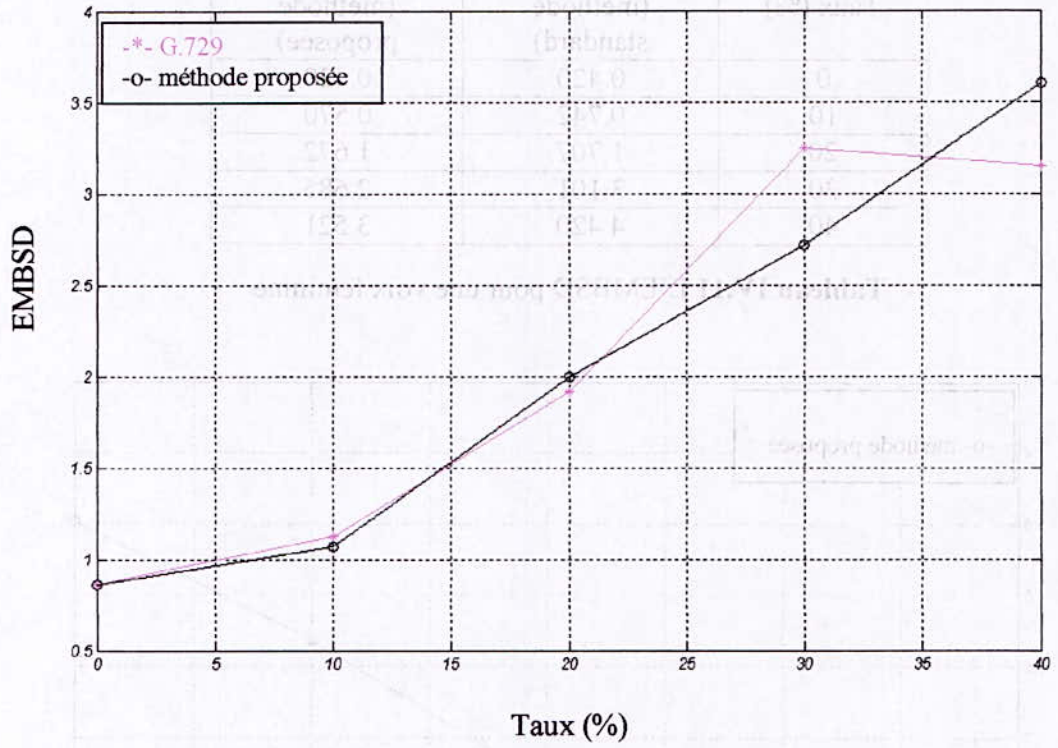

Fig.IV.15 L'EMBSD pour une voix masculine

D'après le tableau IV.11 on peut conclure que la mesure perceptuelle EMBSD donne aussi de bons résultats pour la méthode de dissimulation basée sur la répétition.

D'après le tableau IV.12 la mesure perceptuelle EMBSD n'a pas donnée de bons résultats pour les voix masculines, car cette prend en considération la forme du signal.

Pou mieux voir l'efficacité de la méthode, la figure suivante représente une visualisation des signaux synthétisés sans pertes et avec pertes, en comparaison de la dissimulation standard et la méthode proposée implémentée.

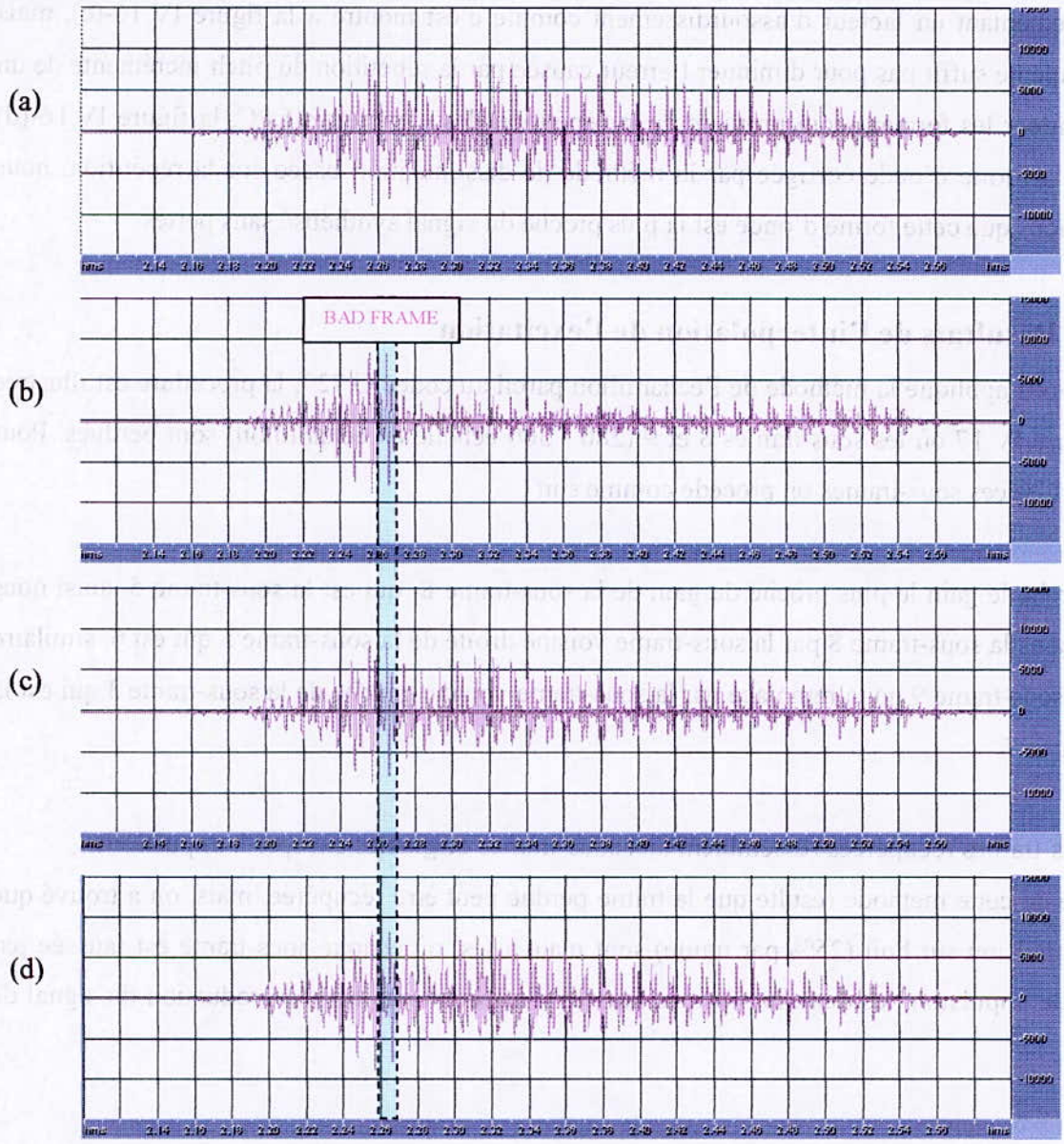

Fig.IV.16 Signal parole synthétisé avec la méthode proposée et la méthode standard

(a)- signal parole synthétisé sans pertes
(b)- signal parole synthétisé avec pertes de 10% avec le standard
(c)- signal parole synthétisé avec l'assourdissement de l'excitation
(d)- signal parole synthétise avec la méthode basée sur la répétition

La figure IV.16-(b) montre la forme d'onde du signal parole avec pertes de 10%, on remarque bien qu'après un effacement de trame le signal se détériore à cause de la mis à jour des dictionnaires avec une version atténués des gains, nous avons réussi à corriger cette détérioration en implémentant un facteur d'assourdissement comme c'est montré à la figure IV.16-(c), mais cette étape ne suffit pas pour diminuer l'erreur causée par la répétition du pitch incrémenté de un et à éliminer les formants qui provient de la répétition des paramètres LPC, la figure IV.16-(d) montre la forme d'onde corrigée par la méthode de dissimulation basée sur la répétition, nous remarquons que cette forme d'onde est la plus proche du signal synthétisé sans pertes.

IV.4.5 Résultats de l'interpolation de l'excitation

Nous avons appliqué la méthode de l'échantillon pareil au codec G.729, la procédure est illustrée à la figure IV.17 où les sous-trames 8 et 9 (280 - 360 échelle en échantillon) sont perdues. Pour reconstruire ces sous-trames on procède comme suit :

On cherche le gain le plus proche du gain de la sous-trame 8 qui est la sous-trame 5, ainsi nous remplaçons la sous-trame 8 par la sous-trame voisine droite de la sous-trame 5 qui est 6, similaire pour la sous-trame 9 on la remplace par la sous-trame voisine gauche de la sous-trame 3 qui est la sous-trame 2.

Les sous-trames récupérées ressemblent aux sous-trames originales excepté 2 impulsions.
Le teste de cette méthode resulte que la trame perdue peut être récupérée, mais, on a trouvé que deux impulsion sur huit (25% par trame) sont mauvaises, où chaque sous-trame est faussée par une seule impulsion, les deux impulsions fausses n'affectent pas trop la production du signal de sortie.

Les résultats de cette méthode sont illustrés dans les schémas suivants :

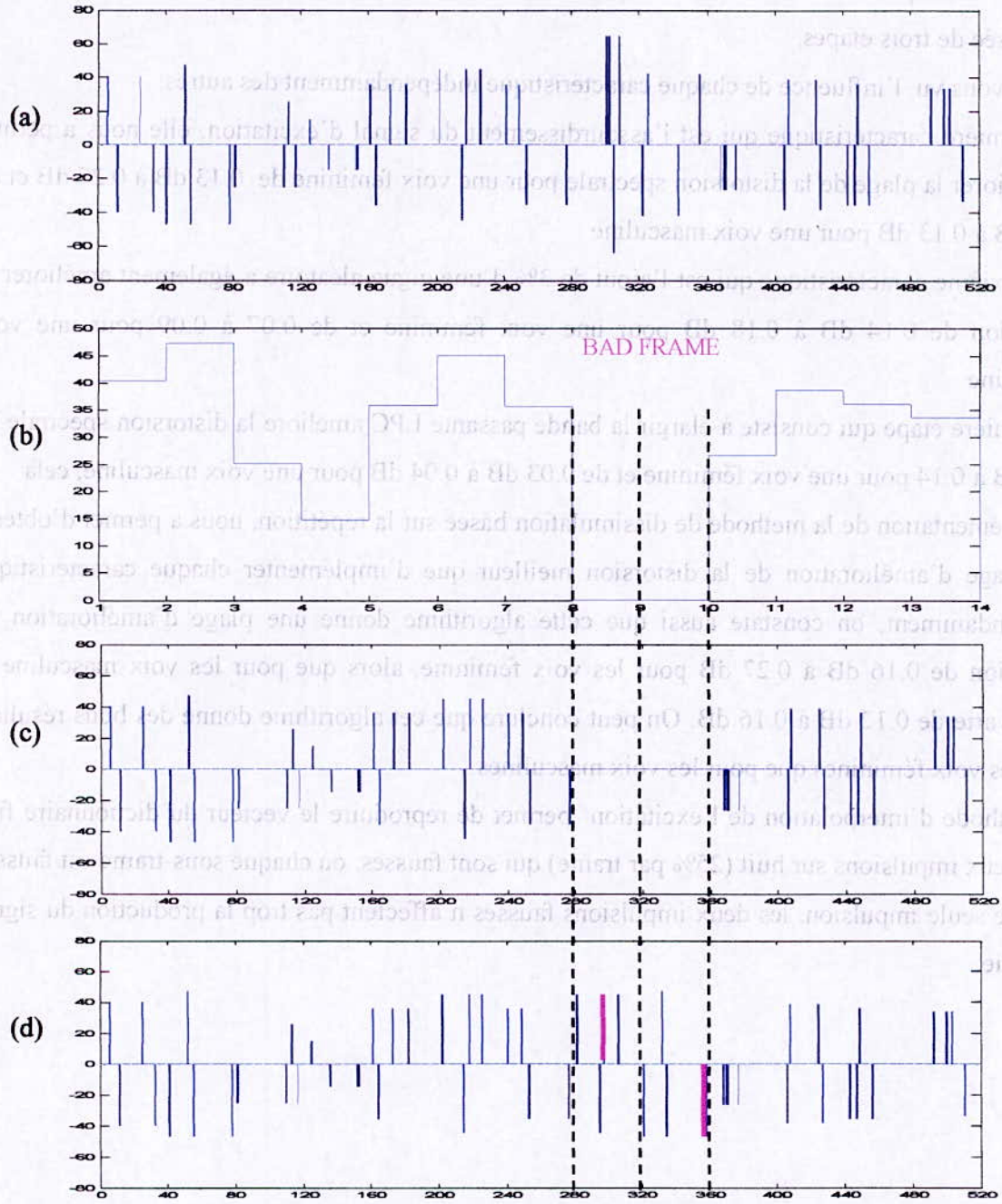

Fig.IV.17 Interpolation de l'excitation
(a)-FCBK original
(b)-Le gain du FCBK
(c)-FCBK avec trame effacée
(d)-FCBK corrigé

Conclusion

Durant ce chapitre nous avons étudié une méthode de dissimulation basée sur la répétition, composée de trois étapes.

Nous avons vu l'influence de chaque caractéristique indépendamment des autres.

La première caractéristique qui est l'assourdissement du signal d'excitation, elle nous a permet d'améliorer la plage de la distorsion spectrale pour une voix féminine de 0.13 dB à 0.27 dB et de 0.10 dB à 0.13 dB pour une voix masculine.

La deuxième caractéristique qui est l'ajout de 3% d'une gigue aléatoire a également améliorer la distorsion de 0.14 dB à 0.18 dB pour une voix féminine et de 0.07 à 0.09 pour une voix masculine.

La dernière étape qui consiste à élargir la bande passante LPC améliore la distorsion spectrale de 0.05 dB à 0.14 pour une voix féminine et de 0.03 dB à 0.04 dB pour une voix masculine, cela

L'implémentation de la méthode de dissimulation basée sur la répétition, nous a permet d'obtenir une plage d'amélioration de la distorsion meilleur que d'implémenter chaque caractéristique indépendamment, on constate aussi que cette algorithme donne une plage d'amélioration de distorsion de 0.16 dB à 0.27 dB pour les voix féminine, alors que pour les voix masculine la plage varie de 0.12 dB à 0.16 dB. On peut conclure que cet algorithme donne des bons résultats pour les voix féminines que pour les voix masculines.

La méthode d'interpolation de l'excitation permet de reproduire le vecteur du dictionnaire fixe avec deux impulsions sur huit (25% par trame) qui sont fausses, où chaque sous-trame est faussée par une seule impulsion, les deux impulsions fausses n'affectent pas trop la production du signal de sortie.

Conclusion

Nous avons vu le long des chapitres précédents que la réception des paquets envoyés n'est pas garantie, cela est dû à la nature "*best effort*" des réseaux *IP*, lorsque 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. Pour cela, l'application des méthodes de dissimulation des trames perdues est nécessaire. Le codec G.729 est équipé d'une procédure de masquage des erreurs.

Dans le but d'améliorer les performances du codec G.729, on a proposé d'implémenter une autre technique de dissimulation basée sur la répétition, cette méthode est composée de trois caractéristiques.

En premier lieu nous avons implémenté l'algorithme d'assourdissement, afin d'éviter une détérioration excessive du signal causée par l'utilisation des versions atténuées du gain (même après que l'effacement des trames est fini) pour la mise à jour du dictionnaire adaptatif.

Nous avons évalué les performances obtenues par cette méthode à l'aide de la distorsion spectrale, les résultats ont montrés que la plage d'amélioration varie entre 0.13 dB à 0.27 dB pour les voix féminines et de 0.10 dB à 0.13 dB pour les voix masculines est obtenue par rapport au standard.

Dans un deuxième lieu nous avons appliqué l'algorithme de l'ajout de 3% d'une gigue aléatoire, dans le but d'éviter l'accumulation des erreurs dans les mauvaises trames successives, nous avons également noté une amélioration de la distorsion spectrale de 0.14 dB à 0.18 dB pour une voix féminine et de 0.07 dB à 0.09 dB par rapport au standard lui-même.

En troisième lieu de notre algorithme, nous avons appliqué une fonction d'expansion à la bande passante LPC pour éviter les formants aigus qui sont caractérisés par une fréquence inférieure à 100 Hz, nous avons obtenu les améliorations suivantes de 0.05 dB à 0.14 dB pour une voix féminine et de 0.03 dB à 0.04 dB pour une voix masculine, cette méthode apporte plus d'amélioration pour les voix féminines car elles sont plus riches de formants aigus.

Dans une dernière étape nous avons rassemblé les trois algorithmes précédents dans un seul algorithme, pour voir l'amélioration qu'apporte cette méthode de récupération basée sur la répétition, cette méthode a renvoyée de bon résultat, et elle nous a permit d'améliorer la plage de distorsion de 0.16 dB à 0.27 dB pour une voix féminine et de 0.12 dB à 0.16 dB pour une voix masculine.

Pour mieux valider nos résultas nous avons utilisé une mesure perceptuelle, les résultats de cette mesure nous a montré une nette amélioration pour les voix féminine.

Afin d'améliorer encore les performances du G.729 nous avons appliqué une méthode d'interpolation de l'excitation, les résultats obtenus ont donnés une erreur de deux impulsions sur huit, les deux impulsions fausses n'affectent pas trop la production du signal de sortie.

Perspectives futures:

- Implémenter l'algorithme de l'interpolation de l'excitation avec celui de la dissimulation basé sur la répétition.
- Confirmer les résultats obtenus par un test d'écoute.
- Implémentation de l'algorithme final sur un chip (DSP).

Annexe A

Algorithme de Levinson-Durbin :

Les coefficients d'autocorrelation R(k), k=0,1,.....,P sont utilisées pour obtenir les coefficients du filtre LP après résolution du système linéaire (1.13)

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

L'algorithme qui va être décrit profite de la structure particulière (Toeplitz symétrique) de la matrice d'autocorrélation pour résoudre (1.13) par une 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 α_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: $m = 1,2,...p.$

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

pour k=1,2,...m-1.

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

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

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

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

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

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

[3] M. Xie et D.Berkani. "*Amélioration des performances des codeurs de parole*" Août 97

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

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

[6] F.Itakura and S. Saito, "*Analysis synthesis telephoney based upon the maximum likelihood method*" in Rep 6 th Int. Congr. on acoustics, Kohasi, Ed. Tokyo, Japan Aug. 21-28, 1968, C-5-5.

[7] J. D Markel and A. H. Gray, Jr "*A linear prediction vocoder simulation based upon the autocorrelation method*", IEEE Trans Acoust. Speech. Signal Processing vol ASSP622, PP.124-134, Apr. 1974.

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

[9] A. H. Gray, and J. D. Markel "*Quantization and bit allocation in speech processing*", IEEE *Trans, on Acoustic, Speech Signal Processing*, vol. ASSP-24, pp. 459-473, Oct. 1976.

[10] P. E. Papamichlis, "*Practical Approaches to Speech Coding*", Prentice-Hall, Englewood Cliffs, N. J. 1987.

[11] D.O'Shaugnessy,"*speech communication, Human and machine.* Reading", MA:Addison-Wesley, 1987.

[12] J.D. Markel and A. H. Gray, Jr., "Linear prediction of speech", New York: Springer-Verlag, 1976.

[13] F.Itakura, "Line *spectrum representation of linear predictive coefficients of speech signals*" *J. Acoust. Soc. Amer.*, vol. 57, suppl. 1 p. S35(A), 1975.

[14] F. K. Soong and B. H. Juang, "*Line spectrum pair (LSP) and speech data compression*", in *Proc. IEEE Int Conf. Acoust. Speech, Signal Processing*, San Diego, CA, pp.1.10.1-1.10.4, Mar.1984.

[15] B.S Atal, R,V Cox and P.Kroon,"*Spectral quantization and interpolation for CELP coders*", in Proc. IEEE int. Conf. On Acoustics, speech and signals.

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

[17] D. Lin, "*Real time voice transmission over the Internet*", Master's thesis, university of Illinois at Urbana-Champaign, Urbana, Illinois, 1999.

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

[19] G. Held, "*Voice Over Data Networks*", New York: McGraw-Hill, 1998.

[20] ITU," *ITU-T G.729: CS-ACELP Speech Coding 8 kbit/s*", ITU1998.

[21] ITU, ITU-T G.723.1: Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s, ITU 1996.

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

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

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

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

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

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

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

[29] C. E. Shannon, "*A mathematical theory of communication*", *Bell Systemes Technical Journal*, pp. 27:379-423, 623-656, 1948.

[30] C. E. Shannon, "coding *theorems for a discrete source with a fidelity criterion*", in *Proc.IRE National Convention Rec.*, Part 4, pp.142-163, 1959.

[31] A. Gersho and R.M. Gray "*Vector Quantization and Signal compression*", Kluwer Academic Publishers, Boston, 1992.

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

[33] Juan Carlos De Martin, Takahiro Unno and Vishu Viswanathan " *IMPROVED FRAME ERASURE CONCEALMENT FOR CELP-BASED CODERS*", DSPS R&D, Texas Instruments Dallas, Texas.

[34] R. Salami et al., "*Design and Description of CSACELP: A Toll Quality 8 kb/s Speech Coder*", IEEE Transactions on Speech and Audio Processing, vol. 6, pp. 116–130, March 1998.

[35] P. Kroon and Y. Shoham, "*Performance of the Proposed ITU-T 8 kb/s Speech Coding Standard for a Rayleigh Fading Channel*", in Proceedings IEEE Workshop on Speech Coding for Telecommunications, (Annapolis, Maryland), pp. 11–12, September 1995.

ملخص

لوحظ في الرّامزة النموذجية ج.729 المقدمة و المعدّلة من طرف الاتحاد الدولي للاتصالات عن بعد، أن بعد ضياع قطع من الكلام، الرّامزة النموذجية ج.729 تصحح هذه القطع الضائعة لكن قيمة المكاسب تصغر(مكسب القاموس التكيفي ومكسب القاموس الثابت). لكن، حتى بعد تصحيح القطعة الضائعة، القطع التالية تتلف بسبب استعمال المكاسب المصغرة.

لقد درسنا و جربنا في هذا البحث الطرق التي تصحح هذا التلف، لقد استعملنا طرقا أساسها تكرار بارا متر قطع الكلام الصحيحة التي وصلت (تصحيح أساسه التكرار وتصحيح أساسه استكمال التنبيه).

الميزة الأساسية لهذه الطرق هي عدم الحاجة لأي وقت زائد.

مفاتيح الكلمات

ترميز الكلام، الكلام عبر شبكة انترنت، إخفاء فقدان القطع، استكمال، تنبيه، تكرار .

RESUME

Dans le codec G.729 de l'ITU (International Telecom Union), nous avons observé qu'après un effacement de trame, la dissimulation standard du codec dissimule les trames perdues avec une atténuation des gains (le gain du dictionnaire adaptatif $g_p^{(n)}$ et le gain du dictionnaire fixe $g_c^{(n)}$). Mais, même après la correction de la trame effacée est fini, les trames suivantes seront détériorés à cause de l'utilisation d'une version atténuée des gains.

Nous avons étudié et tester des méthodes qui corrige cette détérioration, nous avons utilisé des méthodes qui sont basées sur la répétition des paramètres des bonnes trames reçues (Dissimulation basée sur la répétition et Interpolation de l'excitation).

L'avantage majeur de ces méthodes est qu'on n'utilise aucun délai supplémentaire.

Mots clefs :

Codage de la parole, Voix sur IP, gain du dictionnaire adaptatif, gain du dictionnaire fixe, masquage des pertes, interpolation, excitation, répétition.

ABSTRACT

In the codec G.729 of ITU (International Telecom Union), we have observed that after a frame erasure the standard concealment of the codec conceals the lost frames with attenuation of codebook gain (Adaptive Codebook Gain and Fixed Codebook Gain). However, even further the frame erasure is over, the speech signal is further decayed in the subsequent frames. This is because the adaptive codebook is updated with the attenuating excitation signal so the attenuation propagates to the subsequent frames.

We have studied and tested methods that are based on repetition of the parameters of the good frames received on the bad frames (Repetition Based Concealment and Excitation interpolation).

The advantage of these methods of not introducing any extra delay.

Key words:

Speech coding, Voice over IP, adaptive codebook gain, fixed codebook gain, packet loss concealment, interpolation, excitation, repetition.