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

Ecole Nationale Polytechnique

Département d'Electronique

Thèse de Doctorat d'Etat

Présentée par

Fatiha MERAZKA

Magister en Signal & Communications de l'ENP

Thème

Techniques de codage de la parole : applications aux LSPs et aux systèmes VoIP

Devant le Jury :

Président	**Mr. Rabia Aksas**	**Professeur à l'ENP**
Rapporteur	**Mr. Daoud Berkani**	**Professeur à l'ENP**
Examinateur	**Mr. Abderezak Guessoum**	**Professeur à l'université de Blida**
Examinateur	**Mr. Fares Boudjama**	**Professeur à l'ENP**
Examinatrice	**Mm.Mhania Guerti**	**Maître de Conférences à l'ENP**

27/09/2004

Remerciements

Je tiens à exprimer ma profonde gratitude à Monsieur le professeur Daoud Berkani de m'avoir diriger, pour ses conseils judicieux, son suivi attentif et sa confiance .Je tiens à le remercier également de m'avoir intégrer au sein de son équipe de recherche ''codage & compression '' du laboratoire signal & communications.

J'adresse mes vifs remerciements à Monsieur le professeur Rabia Aksas d'avoir accepter de présider le juty et pour son aide précieuse.

Je tiens également à remercier vivement les examinateurs de ce travail, monsieur Abderazak Guessoum Professeur à l'université de Blida, monsieur Fares Boudjama Prosesseur à l'ENP et madame Mahia GUerti Maître de conferences à l'ENP.

في خوارزميات التنبؤ ا لخطي الترميزي، إرسال عوامل التنبؤ الخطي للترميز ، عادة محولة إلى التمثيل بزوج من الخطوط الطيفية تستهلك نسبة عالية من(bit rate) للمرمز. في bit rat إنه ضروري تكميم هذه العوامل بدقة باستعمال أقل عدد ممكن من دون التضحية بالنوعية الشفافة وبعدد قليل من العمليات الحسابية. حتى وإن كانت أشعة التكميم أكثر فعالية من التكميم الجبري، استعمالاتها في التكميم الدقيق للمعلومة منحصرة في احتمالات ارتفاع عدد العمليات الحسابية. في هذا البحث حل مقدم لمشكل العمليات الحسابية بمقارنة ثلاثة بنيات لشعاع المكمم بالانقسام. الشعاع المكمم بالانقسام **4/3/3** وجد أنه الأحسن من بين الثلاثة بنيات المقدمة. حلين لمشكل تعاكس العوامل الذي يمكن أن ينجم عن عملية التكميم انقسامي مقترحة هنا. . في هذا البحث ثلاثة قياسات للمسافة فحصت باستعمال الشعاع المكمم بالانقسام **4/3/3** و المسافة المتوسطة العكسية وجدت أنها الأكثر ملأمة بأحسن مؤهلات. خوارم مقترح لعامل الضجيج على مؤهلات المكمم **4/3/3** و النتائج الناتجة مقدمة هنا. .

Abstract -- Line Specrum Pairs (LSPs) have been the prevailing parameter set to represent LP coefficients in speech coding. At low bit rate, it is essential to quantize these parameters accurately, using as few bits as possible without sacrificing the transparent quality and with low complexity. Though the vector quantizers are more efficient than the scalar quantizers, their use for accurate quantization of LPC information is restricted due to their probable high complexity. In this work, a remedial solution to the problem of complexity is proposed by comparing three structures of the split vector quantizer. Two solutions to the problem of LSP inversion, which may arise after split quantization, are proposed. An algorithm is proposed for the effect of noise on the performance of the split 3/3/4 quantizer. With the emergence of voice over packet networks (VoIP), erasure-robustness has become an important issue for coder performance. Coding standards designed for these applications are presented by the ITU dual rate voice coder G.723.1 and ITU toll quality coder G. 729. However, both coders inherited the inter-frame predictive split VQ coding of LSPs, which cause error propagation. To elevate this problem, we propose to investigate the trade-offs between LSPs coding rate and erasure performance based on G. 729 and G.723.1 by using intra-frame methods for coding the LSPs with an interpolative concealment method. Our results show that intra-frame coding with interpolative method is much more robust to frame erasure than the methods used by the two standards (ITU G.729, ITU G723.1).

Index terms – Speech coding, LSP parameters, Split Vector quantization, spectral distortion measure, VoIP, erasure robustness, ITU G.729, ITU G723.1 and error propagation.

Résumé: - Les LSPs représentent les coefficients de prédiction linéaire LP en codage de la parole. A bas débit, il s'agit de les quantifier avec un nombre de bits aussi faible que possible tout en gardant une qualité transparente et une faible complexité algorithmique. Dans notre travail, nous utilisons une quantification vectorielle (VQ) des LSPs. Nous proposons une solution au problème de complexité en comparant trois structures de la VQ par split (SVQ). Nous proposons également deux solutions au problème d'inversion des LSPs après SVQ. Un algorithme est proposé pour les effets du bruit sur la robustesse du SVQ 3-3-4. Avec l'émergence de la transmission de la voix à travers les réseaux IP (VoIP), la robustesse contre l'effacement des trames est devenue un critère important pour l'évaluation des performances d'un coder. Les standards G 723.1 et le G.729 sont conçus pour ces applications. Pour éviter la propagation des erreurs causées la VQ prédictive utilisée par ces deux codeurs, nous proposons l'étude du compromis entre le débit de codage des LSPs et les performances contre l'effacement, en employant une VQ intra-trame et une méthode de masquage par interpolation. Nos résultats montrent une amélioration de la robustesse.

Mots clés – codage de la parole, les paramètres LSPs, quantification vectorielle par split, mesure de distorsion spectrale, VoIP, robustesse contre effacement, ITU G.729, ITU G723.1 et propagation des erreurs.

Sommaire

Introduction

Nous proposons dans ce travail une solution au problème de réduction de la complexité de calcul par une comparaison de trois structures de codage intra-trame SVQs (Split Vector Quantization). Nous proposons pour la première fois deux solutions aux problèmes d'inversions des LSPs aux frontières des sous-vecteurs, après quantification par split. Un algorithme est également proposé pour les effets du bruit sur la robustesse du quantificateur par split utilisé. Ces aspects avec les améliorations apportées aux systèmes VoIP (Voice over Internet Protocol) constituent notre contribution dans ce domaine.

Les applications des communications numériques modernes, comme la téléphonie cellulaire, ont conduit à un besoin croissant d'une haute qualité des schémas de codage à très bas débit. Le rôle des codeurs de la parole est de transmettre le signal parole de la manière la plus efficace possible. Certaines contraintes comme le débit binaire, la complexité et la robustesse aux erreurs de transmission sont imposées à la conception d'un système de codage de la parole avec une reconstruction acceptable du signal parole comme but principal.

La plupart des codeurs actuels (contemporains) sont basés sur le codage par prédiction linéaire LPC/ LP où un signal d'excitation est envoyé à un filtre tous-pôles représentant l'information spectrale de la parole. Pour plusieurs applications, le spectre LP contient la majeure partie de l'information et par conséquent il est important de coder les paramètres LP en utilisant un nombre de bits aussi faible que possible tout en gardant une qualité satisfaisante de la parole.

Le spectre du signal parole est modélisé par un filtre à prédiction linéaire (LP) particulièrement d'ordre dix dans les codeurs LP. Ce filtre LP est généralement actualisé chaque 5 à 20ms d'intervalle. En compétition avec les coefficients de réflexion et les LARs (log area ratio), les LSPs (Line Spectrum Pairs) ont montré qu'ils sont les mieux adaptés au codage de la parole aujourd'hui du à leurs propriétés favorables en terme de stabilité,

distribution et sensibilité spectrale. Notre but est d'étudier le problème de la transmission efficace de l'information spectrale en utilisant un nombre de bits aussi faible que possible et une moindre complexité pour une qualité transparente de la parole.

Plusieurs études ont été faites dans le passé pour quantifier les paramètres LP dans le but de représenter l'enveloppe spectrale par un nombre de bits aussi faible que possible. En quantification scalaire, plusieurs représentations paramétriques LP ont été utilisées Récemment plusieurs études utilisant les LSPs ont montré que 32-40 bits sont nécessaires pour quantifier une trame du signal parole avec une précision raisonnable. La quantification vectorielle (VQ) est une méthode importante en codage de source et de compression des signaux et elle représente aujourd'hui un champ de recherche majeur. Le codage de la parole est un domaine où la VQ a été continuellement exploitée durant la dernière décade. Les codeurs LP sont basés sur la VQ de la séquence d'excitation. Le codage de la partie spectrale du codeur CELP a été initialement donné par des méthodes scalaires. Jusqu'environ 1980, la plupart des schémas de codage étaient fondés sur la quantification scalaire dans une certaine mesure. Cependant leur complexité a limité l'utilisation de la VQ. Le premier travail incorporant la VQ a été décrit par Buzo et al. En 1980, mais les performances obtenues avec un VQ à 10 bits/ trame étaient inacceptables. A la place des méthodes décrites précédemment, plusieurs méthodes hybrides de la quantification scalaire et vectorielle ont été étudiées. L'application directe de la VQ reste encore inconcevable en pratique mais plusieurs schémas dits avec contraintes ou sous-optimales qui réduisent la complexité de stockage au détriment d'une dégradation des performances ont montré qu'ils dépassent celles d'un quantificateur scalaire. Cependant la VQ reste complexe: pour un nombre fixe de bits par paramètre (débit) le nombre de vecteurs de reconstruction (taille) du VQ augmente de manière exponentielle avec la dimension des vecteurs. La mémoire de stockage ainsi que la complexité de recherche croient alors de manière exponentielle. Une base de données des vecteurs représentatifs couvrant la variabilité des paramètres à quantifier est nécessaire pour le design d'un VQ qui constitue en pratique un problème à prendre en considération. Intra-trame Split VQ (SVQ) et multi-étages VQ (MSVQ) sont des techniques qui remédient à ce problème mais en payant le prix d'une dégradation des performances. Pour des applications, pratiques la qualité transparente est très coûteuse. Les standards récents ont établi le compromis entre coût-performance à des exigences inférieures à la qualité transparente de la parole. Un exemple

typique est le standard ITU (International Telecommunications Union) G .729, où une qualité non transparente tout en restant bonne, peut être obtenue à 18 bits/trame en employant un MSVQ à faible complexité combiné à un SVQ dans un schéma prédictif.

Avec l'émergence de la transmission de la voix à travers les réseaux par paquets comme Internet, la robustesse contre l'effacement des trames est devenue un point important dans l'évaluation des performances d'un codeur. Les standards conçus pour ces applications sont présentés par l'ITU G.732.1 [24] à deux débits et l'ITU G.729 à qualité téléphonique (toll)[23]. Cependant, les deux codeurs utilisent une méthode SVQ prédictive pour le codage des paramètres LSPs à partir des développements précédents où le souci majeur étaient le débit et la complexité. Avec le codage inter-trame qui tient compte de la corrélation entre les trames, lorsque l'effacement d'une trame se produit, l'état du décodeur change et cause ainsi une propagation de l'erreur aux trames suivantes. Malgré que la plupart des codeurs possédaient la possibilité de lisser (smooth out) les trames effacées, au moins 2-3 des trames sont affectées. En tenant compte de la robustesse contre l'effacement des trames, nous avons étudié le compromis entre le débit de codage des LSPs et les performances contre l'effacement des trames pour les standards de l'ITU G. 723.1 et G. 729. Pour cela, nous avons comparé plusieurs méthodes de quantification avec celles utilisées par les deux standards dans le but d'améliorer les performances de ces derniers. Nous avons comparé une quantification intra-trame avec le prédictive split VQ pour le Standard ITU G. 729 et G.723.1.

Ce travail de recherche a été consacré à l'étude et au développement de techniques de compression de la parole et à sa transmission à travers les réseaux IP (VoIP). Nous avons partagé notre travail en deux parties. La première partie, comportant les trois premiers chapitres, est consacrée aux techniques de codage des paramètres LSPs à bas débit. La seconde comporte les deux derniers chapitres et elle est consacrée aux systèmes VoIP.

Dans la première partie, le premier chapitre comporte des généralités sur le modèle humain de production de la parole, le système auditif humain et un aperçu sur le codage de la parole. Le second chapitre est consacré à la prédiction linéaire en codage de la parole par la présentation de différentes méthodes d'estimation des coefficients de prédictions LP ainsi que d'autres représentations équivalentes de ces derniers obtenus par des transformations.

Le troisième chapitre présente un aperçu sur les quantifications scalaire et vectorielle et les conditions pour l'optimalité. Il traite également de la quantification intra-trame que nous avons proposé pour coder les paramètres LSPs, les résultats obtenus ainsi que leurs interprétations.

Dans la deuxième partie, le quatrième chapitre introduit les systèmes de transmission de la voix à travers les réseaux IP (*VoIP*), les problèmes de perte de paquets rencontrés et les méthodes de récupération (ou de masquage) de ces pertes.

Le cinquième chapitre expose les améliorations des performances, que nous avons apportées aux deux standards ITU G.729 et G.723.1.

On terminera par une conclusion générale sur les méthodes utilisées et les résultats obtenus ainsi que les perspectives futures.

Chapitre 1
Notions sur le signal Parole

1.1 Introduction

Notre système auditif n'est pas équitablement sensible aux distorsions à différentes fréquences et possède une gamme dynamique limitée. Alors, la compréhension de la physiologie de la production de la parole chez l'être humain, les propriétés de base du signal parole et sa perception sont cruciales dans la conception d'un codeur de la parole qui doit, idéalement, paramétré uniquement l'information perceptuellement importante et par conséquent représenter le signal de manière compacte.

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. Une classification des différents codeurs terminera ce chapitre.

1.2 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.

1.3 Le Mécanisme de la 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 1.1.

L'appareil respiratoire fournit l'énergie nécessaire à la production de sons, en poussant de l'air à travers la trachée-artère. Au sommet de celle-ci se trouve le ***larynx*** 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 entourent une cavité située à la partie supérieure de la trachée. Les ***cordes vocales*** sont en fait deux lèvres symétriques placées en travers du larynx. Ces lèvres peuvent fermer complètement le larynx et en s'écartant progressivement, 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.

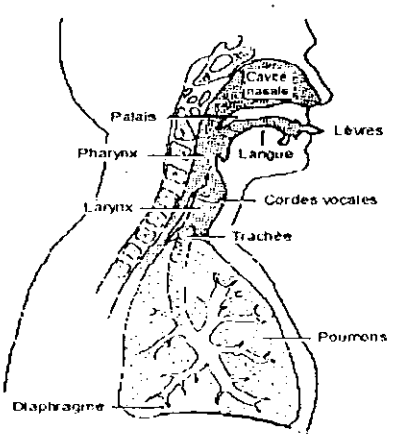

Figure 1.1 Appareil phonatoire.

Les sons voisés résultent, au contraire, d'une vibration périodique des cordes vocales. Le larynx est d'abord complètement fermé, ce qui accroît la pression en amont des cordes vocales et 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] :

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

Un *son voisé* est un signal quasi périodique dont le spectre est tracé à la Figure 1.2. On y observe les raies qui correspondent aux harmoniques du fondamentale F_0 (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 1.3. 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 à d produite avec la glotte largement ouverte. Dans ce cas, le spectre du signal résulte de l'excitation du conduit vocal par une source aléatoire. C'est un spectre continu qui présente une structure formantique semblable à celle d'une voyelle voisée mais ne possède pas de structure de pitch (raies dues aux harmoniques du fondamental).

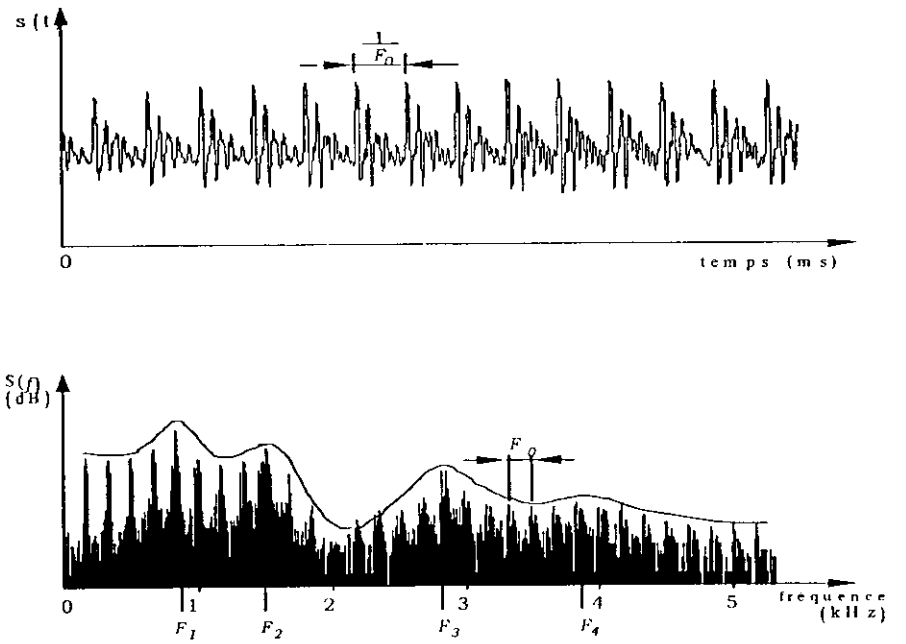

Figure 1.2 Un signal vocal voisé et son spectre.

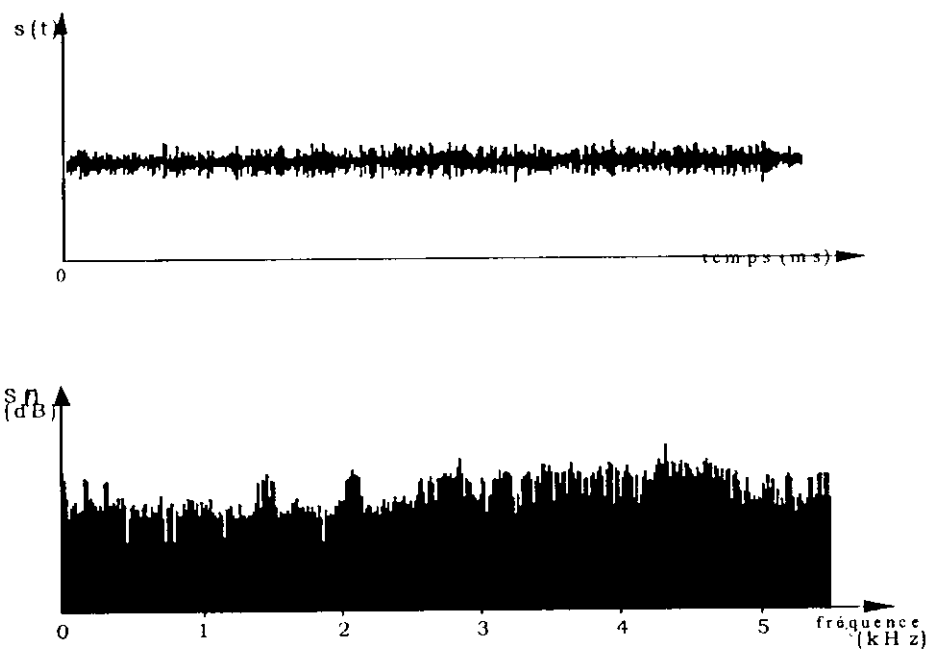

Figure 1.3 Un signal vocal non voisé et son spectre.

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 ***psychoacoustique***. Sans vouloir entrer dans trop de détails sur la contribution majeure des psychoacousticiens dans l'étude de la parole, il est intéressant d'en connaître les résultats les plus marquants. Ainsi, l'oreille ne répond pas également à toutes les fréquences. Le seuil d'audition de l'oreille est non linéaire par rapport aux fréquences. L'oreille atteint sa sensibilité maximale entre 3 et 4 kHz.

1.4 Les Redondances dans le Signal Parole

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. Les techniques de codage de la parole ont pour but la diminution du débit ou la minimisation de la redondance tout en gardant une qualité satisfaisante, sans nuire à son intelligibilité.

En réalité, le problème est plus complexe car le signal parole présente une très grande variabilité. Cette variabilité est causée par de nombreux paramètres dont la redondance, la prosodie et les caractéristiques propres à chaque locuteur.

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

Tableau 1.1 Exemple de messages à coder

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

L'entropie de la source est :

$$H(X) = 0.5\log 2 + 0.25\log 4 + 0.125\log 8 + 0.125\log 8 = 1.75 bits \qquad (1.1)$$

Dans le deuxième type de codage (n=2), l'efficacité est alors

$$\eta = \frac{H(X)}{2} = 0.875 \qquad (1.2)$$

La redondance est alors

$$\rho = 1 - \eta = 0.125 \qquad (1.3)$$

L'efficacité de 87.5% (ou la redondance de 12.5%) provient du fait qu'on a avantagé de la même manière des messages de fréquences différentes.
Le troisième codage, au contraire, affecte les mots les plus courts aux messages les plus fréquents.

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

dans ces conditions, on aura

$$\eta = 1 \text{ et } \rho = 0$$

Le code est efficace à 100%. Sa redondance est nulle.
Dans une conversation courante, environ dix phonèmes sont prononcés chaque seconde. L'information moyenne est donc inférieure à 50 bits/s [5]. D'un autre coté pour garder une haute qualité de la parole avec une représentation numérique du signal parole, l'utilisation d'un système de conversion A/D réclame plus de 100 kbits/s. Il y a donc apparemment une redondance énorme dans le signal vocal. La suppression partielle des redondances permet une représentation plus efficace des données.
La compression des données peut se faire sans pertes d'information ou avec pertes en exploitant dans ce cas la tolérance de l'organe récepteur (l'oreille). La compression du signal consistera à réduire les redondances du signal parole.

1.5 Le Modèle de Production de la Parole

L'analyse de la parole est une étape indispensable à toute application de synthèse, de codage ou de reconnaissance. Elle repose en général sur un modèle. Celui-ci possède un ensemble de paramètres numériques dont les plages de variation définissent l'ensemble des signaux couverts par le modèle.

Fant [3] a proposé en 1960 un modèle de production qui spécifie qu'un signal voisé peut être modélisé par le passage d'un train d'impulsions $u(n)$ à travers un filtre numérique récursif de type tous-pôles. On montre que cette modélisation reste valable dans le cas des sons non voisés, à condition que $u(n)$ soit cette fois un bruit blanc. Le modèle final est illustré à la Figure 1.4. Il est souvent appelé modèle auto régressif (AR), parce qu'il correspond dans le domaine temporel à une régression linéaire de la forme

$$X(n) = G \cdot u(n) + \sum_{i=1}^{p} - a_i X(n-1) \tag{1.4}$$

où *u(n)* et p sont respectivement le signal d'excitation et l'ordre du système. Chaque échantillon est obtenu en ajoutant un terme d'excitation à une prédiction obtenue par combinaison linéaire des p échantillons précédents.

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

Les paramètres du modèle AR sont : la période du train d'impulsions (sons voisés uniquement), la décision son voisé/non voisé, le gain G et les coefficients du filtre $1/A(z)$, appelé filtre de synthèse.

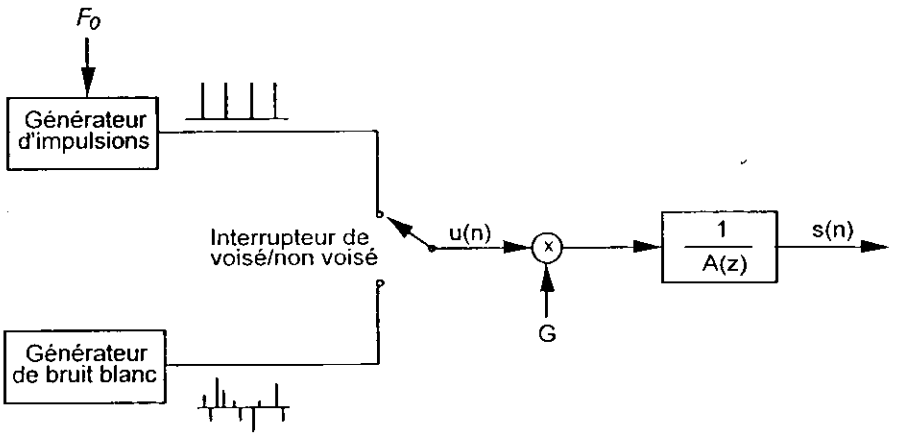

Figure 1.4 Modèle simplifié de production de la parole.

Les relations d'équivalences entre le modèle physique et le modèle mathématique (Figure 1.4) peuvent être données comme suit :

Conduit vocal	$\Leftrightarrow$	$1/A(z)$, le filtre LPC
Le flux d'air	$\Leftrightarrow$	Le signal d'excitation $u(n)$
Vibration des cordes vocales	$\Leftrightarrow$	voisé
Période de vibration des cordes vocales	$\Leftrightarrow$	$T = 1/F_0$, période du pitch
Fricatives et plosives	$\Leftrightarrow$	non voisé/voisé
Volume d'air	$\Leftrightarrow$	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.

1.6 Classification des Codeurs de la Parole

L'objectif principal d'un codec de la parole est la compression du signal parole en un signal informationnel aussi petit que possible tout en minimisant la distorsion résultante. Le codage de la parole invoque le compromis entre compression, distorsion résultante et complexité. Généralement, plus la compression est obtenue, plus la distorsion augmente, provoquant une dégradation de la qualité de la parole. Cependant, des algorithmes plus

compliqués peuvent atteindre une plus grande efficacité de compression pour une distorsion inférieure. Pour aider à la classification de la qualité de la parole, il existe un nombre de termes standard regroupés dans le Tableau 1.2, qui aident à l'évaluation des différents algorithmes de codage de la parole.

Tableau 1.2 – Les termes généraux de classification de la qualité de la parole.

Qualité	Description
Broadcast	Broadcast quality (0-7000 Hz)
Toll	Qualité téléphonique (300-3400 Hz)
Digital Cellular	Quelques distorsions de la qualité téléphonique relevées, mais les sons restent naturels (naturalness)
Communications	Quelques distorsions relevées avec pertes de la sonorité naturelle (naturalness)
Synthetic	Distorsion type synthétique mais reste intelligible.

Plusieurs algorithmes de codage de la parole exploitent les propriétés du signal parole à des degrés différents. Par conséquent, ils peuvent être divisés en trois catégories : les *codeurs de formes d'ondes* (waveform coders) qui généralement nécessitent un haut débit utilisant très peu ou presque pas la modélisation du signal et donnent une très bonne qualité de la parole. De l'autre côté, *les vocodeurs* ou *codeurs paramétriques* (vocoders or parametric coders) nécessitent un plus faible débit, ils modélisent le conduit vocal et le signal d'excitation et produisent une qualité synthétique de la parole. Récemment, une nouvelle classe de codeurs, appelés l*es codeurs hybrides* (hybrid coders), est introduite. Ils utilisent des techniques des deux classes de codeurs précédents et donnent une bonne qualité à des débits moyens.

1.6.1 L es Codeurs de Formes d'Ondes

Les codeurs de formes d'ondes essayent de reproduire les formes d'ondes du signal d'entrée. Ils sont généralement conçus indépendamment du signal pour être utilisé au codage d'une grande variété de signaux. Généralement, ils sont à faible complexité et produisent une haute qualité à des débits supérieurs à environ 16 kbits/s. Les codeurs de formes d'ondes peuvent opérer dans le domaine *temporel* ou *fréquentiel*.

1.6.1.1 Codeurs dans le Domaine Temporel

Les codeurs de formes d'ondes dans le domaine temporel réalisent le processus de codage sur des échantillons temporels du signal. Les méthodes de codage les plus connues dans le domaine temporel sont [4]: le codage PCM (Pulse Code Modulation), le codage APCM (Adaptive Pulse Code Modulation), le codage DPCM (Differential Pulse Code Modulation), le codage ADPCM (Adaptive Differential Pulse Code Modulation), le codage DM (Delta Modulation), le codage ADM (Adaptive Delta Modulation) et le codage APC (Adaptive Predictive Coding). Dans ce qui suit, on décrit brièvement quelques schémas importants de codage dans le domaine temporel.

o Les Codeurs PCM

C'est le plus simple type de codage de formes d'ondes. C'est essentiellement un processus de quantification échantillon par échantillon. N'importe quelle quantification scalaire peut être utilisée avec ce schéma, mais la forme de quantification la plus utilisée est la quantification logarithmique. La recommandation du codeur G.711 du CCITT (International telegraph and Telephone Consultative Committee's) définie 8 "A-Law et μ-Law PCM" comme une méthode standard pour le codage du signal téléphonique.
Dans cette catégorie de codeurs, on distingue les codeurs temporels et fréquentiels. Ces derniers n'utilisent aucune connaissance à priori sur la façon dont le signal est généré. Le codeur temporel fait correspondre à l'amplitude du signal analogique une suite d'échantillons discrets.

o Les Codeurs DPCM et ADPCM

La technique PCM ne fait aucune supposition sur la nature des formes d'ondes à coder, par conséquent, elle fonctionne bien pour des signaux différents de ceux de la parole. Cependant, en codant la parole, il existe une très forte corrélation entre les échantillons successifs obtenus. Cette corrélation peut être exploitée pour réduire le débit binaire. Une méthode simple de le faire est de transmettre uniquement la différence entre deux échantillons. Le signal différence possèdera alors une gamme dynamique plus réduite que le signal original, et peut être alors quantifié moyennant un nombre de niveaux

reconstitution plus réduit. Dans la méthode citée plus haut, l'échantillon précédent est utilisé pour prédire la valeur de l'échantillon présent. La prédiction sera améliorée si un bloc plus large de la parole est utilisé pour la prédiction. Cette technique est connue sous le nom de DPCM.

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

1.6.1.2 Codeurs dans le Domaine Fréquentiel

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

o Les Codeurs en Sous Bande

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

o Les Codeurs par Transformée

Cette technique transforme par bloc un segment du signal d'entrée dans le domaine fréquentiel ou un domaine similaire. Le codage adaptatif est réalisé en attribuant plus de

bits aux coefficients de transformation les plus importants. Au niveau du récepteur, le décodeur fait la transformation inverse pour obtenir le signal reconstruit. Plusieurs transformées comme la DFT (Discrete Fourier Transform) ou DCT (Discrete Cosine Transform) peuvent être utilisées.

1.6.2 L es Vocodeurs

Les performances des vocodeurs, connus aussi sous le nom de codeurs de source, sont fortement dépendantes de la précision des modèles de production de la parole. Ces codeurs sont conçus spécifiquement pour des applications à bas débit (comme les communications militaires ou satellitaires) et sont principalement destinés à maintenir une qualité intelligible de la parole. Les vocodeurs les plus efficaces sont basés sur la prédiction linéaire LPC (Linear Predictive Coding). Les détails de cette technique seront abordés dans le chapitre suivant. Une qualité des communications (tableau 1.2) peut être obtenue à des débits inférieurs à 2 kbits/s avec les vocodeurs LPC [5].

1.6.3 Les Codeurs Hybrides

La qualité des codeurs de formes d'ondes chute rapidement pour des débits inférieurs à 16 kbits/s, mais il existe une amélioration négligeable dans la qualité des vocodeurs à des débits supérieurs à 4 kbits/s. Les codeurs hybrides sont alors utilisés pour combler ce vide, donnant ainsi une 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 [6].
Les codeurs de formes d'ondes par interpolation WI (Waveform Interpolation) modélisent le signal résiduel par des formes d'ondes caractéristiques qui peuvent être interpolées aussi bien dans le domaine temporel que fréquentiel pour la reconstruction. Pour des débits inférieurs à 4 kbits/s, les codeurs WI donnent de meilleurs performances comparés à d'autres codeurs opérant à des débits similaires [7]. Cependant, les codeurs WI sont actuellement alourdis par leur complexité élevée et leur large retard (typiquement 40 ms).

1.7 Critères de Performance dans le Codage de la Parole

En général, il existe un ensemble de différents attributs qui décrivent les performances d'un codeur de la parole. Plusieurs d'entre eux nécessitent d'être considéré lors du design d'un système de transmission de la parole. Nous donnerons, dans ce qui suit, certains de ces attributs.

1.7.1 L a Qualité du signal Parole

Un des critères majeur est la qualité de la parole. Les codeurs de la parole tentent de produire moins de distorsions audibles pour un débit donné. La sonorité naturelle (naturalness) et l'intelligibilité des sons produits sont importants et constituent les critères désirés. La qualité de la parole peut être déterminée par des tests d'écoutes qui calculent l'opinion moyenne des écouteurs. La qualité de la parole peut être aussi déterminée par des mesures objectives comme la prédiction du gain, la distorsion spectrale logarithmique, etc. Le but essentiel des codeurs de la parole est que le signal décodé ou synthétisé soit le plus proche possible du signal original.

1.7.2 L e Débit Binaire

Un autre critère important est le débit binaire. Le débit binaire d'un codeur est le nombre de bits par seconde dont le codeur a besoin pour transmettre le signal. L'objectif d'un

algorithme de codage est la réduction du débit binaire tout en maintenant une qualité bonne de la parole.

1.7.3 La Complexité

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

1.7.4 L e Retard de Communication

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

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

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

1.7.5 La Bande Passante

La bande passante d'un signal à coder est aussi un critère important. Précisément, le signal téléphonique nécessite une bande passante de 200-3400 Hz. Les techniques de codage à large bande (utiles dans les transmissions audio, tele-conferencing et tele-enseignement) nécessitent une bande passante de 7-20 kHz.

1.7.6 La Robustesse

Les algorithmes de codage doivent être robustes contre les effets des erreurs du canal. Les erreurs du canal sont causées par le bruit du canal, interference inter-symbol, l'évanouissement du signal (fading du signal), etc.

1.7.7 La Transparence

Les signaux de la parole sont transmis en temps réel et sont distordus par plusieurs types de bruits de fond acoustiques comme le bruit de la rue, le bruit des voitures et le bruit du bureau. Par conséquent, les algorithmes de codage doivent être capables de maintenir une bonne qualité en présence de tels bruits.

1.8 Conclusion

A fin de mieux réduire le débit binaire, il est nécessaire de connaître le plus de propriétés du signal information. La décomposition du signal permet de connaître certaines des propriétés du signal. Bien souvent la prédiction linéaire (LP) est utilisée à cette fin.

Chapitre 2

La Prédiction Linéaire en Codage de la Parole

2.1 Introduction

Dans ce chapitre, on présentera la prédiction linéaire LP (Linear Prediction), qui est communément utilisé dans les algorithmes de codage de la parole à bas débit. La LP permet de modéliser le signal parole par une combinaison linéaire des échantillons précédents. On se focalisera, particulièrement sur la prédiction à court terme des paramètres spectraux. Plusieurs représentations paramétriques des coefficients de prédiction seront introduites pour améliorer l'efficacité du codage spectral. Nous présenterons également, les mesures de distorsions qui évaluent les performances du codage spectral.

2.2 L'Analyse par Prédiction Linéaire

Le codage des paramètres spectraux est une composante intégrale de codage de la parole. Le modèle du filtre source pour la production de la parole nous permet d'utiliser la prédiction linéaire pour analyser le comportement court terme des signaux parole au sein d'une trame de la parole. 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][11]. Son expression est alors

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

où le gain G, les coefficients $\{a_k\}$et (b_i) sont les paramètres du système, p et q sont les ordres des polynômes. L'équation (2.1) 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

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

$H(z)$ est le modèle pôle-zéro dans lequel les racines du dénominateur et numérateur sont, respectivement, les pôles et les zéros du système. Lorsque $a_k = 0$ pour $1 \le k \le p$, $H(z)$ devient un modèle tous-zéros ou modèle à *moyenne ajustée* (MA). Contrairement, si $\{b_i = 0\}$ pour $1 \le i \le q$, $H(z)$ devient un modèle tous-pôles ou modèle *auto régressive* (AR), exprimé par :

$$H(z) = \frac{G}{A(z)} \tag{2.3}$$

Dans l'analyse de la parole, les classes de phonèmes comme les fricatives et les nasales contiennent des vallées spectrales qui correspondent aux zéros dans $H(z)$. Par contre, les voyelles contiennent des résonances qui peuvent être modélisées par le modèle 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 :

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

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

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

L'ordre p du système est choisi de façon que l'estimation de l'enveloppe spectrale soit la plus adéquate possible. Une des façons de procéder est d'allouer une paire de pôles pour chaque formant présent dans le spectre. On ajoute 2 ou 3 pôles pour une meilleure approximation des zéros dûs aux sons non voisés.

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

2.2.1 Méthode d'Autocorrélation

La méthode d'autocorrélation 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 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 $s_w(n)$[5].

$$s_w(n) = \begin{cases} w(n) \cdot s(n) & \text{pour } 0 \le n \le N-1 \\ 0, & \text{ailleurs} \end{cases} \qquad (2.6)$$

La fonction de pondération la plus courante est la fenêtre de Hamming définie par

$$w(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) & \text{pour } 0 \le n \le N-1 \\ 0, & \text{ailleurs} \end{cases} \qquad (2.7)$$

Chaque échantillon peut être prédit approximativement à partir de p échantillons précédents. Ceci est valable pour toutes les valeurs du temps de moins l'infini à plus l'infini ($-\infty \prec n \prec +\infty$.)

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

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

$$R(i) = \sum_{n-i}^{N-1} s_w(n) s_w(n-i) \qquad 1 \le i \le p \tag{2.8}$$

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 $0 \le n \le N-1$:

$$E = \sum_{n=-\infty}^{\infty} e^2(n) = \sum_{n=-\infty}^{\infty} (s_w(n) - \sum_{k=1}^{p} a_k s_w(n-k))^2 \tag{2.9}$$

En procédant à l'annulation des dérivations partielles par rapport aux coefficients du filtre

$$\frac{\partial E}{\partial a_k} = 0 \qquad 1 \le k \le p \tag{2.10}$$

On obtient p équations linéaires avec p coefficients inconnus a_k

$$\sum_{k=1}^{p} a_k \sum_{n=-\infty}^{\infty} s_w(n-i) s_w(n-k) = \sum_{n=-\infty}^{\infty} s_w(n-i) s_w(n). \qquad 1 \le i \le p \tag{2.11}$$

alors, l'équation linéaire (2.11) peut être écrite sous la forme :

$$\sum_{k=1}^{p} R(|i-k|) a_k = R(i) \qquad 1 \le i \le p \tag{2.12}$$

La forme matricielle de l'ensemble des équations linéaires est représentée par $R \cdot a = v$ et peut être réécrite comme suit

$$\begin{bmatrix} R(0) & R(1)\cdots R(p-1) \\ R(1) & R(0)\cdots R(p-2) \\ \vdots & \vdots \\ R(p-1) & R(p-2)\cdots R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ \vdots \\ R(p) \end{bmatrix} \tag{2.13}$$

La matrice d'autocorrélation $p \times p$ obtenue est symétrique dont tous les éléments de la diagonale sont égaux, c'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 (WLD).

2.2.2 Méthode de Covariance

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

$$E = \sum_{n=-\infty}^{\infty} e_w{}^2(n) = \sum_{n=-\infty}^{\infty} e^2(n) w^2(n) \tag{2.14}$$

En annulant les dérivées partielles $\delta E / \delta a_k$ pour $1 \le i \le p$, on obtient p équations linéaires :

$$\sum_{k=1}^{p} \Phi(i,k) a_k = \Phi(i,0), \qquad 1 \le i \le p \tag{2.15}$$

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

$$\Phi(i,k) = \sum_{n=-\infty}^{+\infty} w(n)s(n-i)s(n-k) \tag{2.16}$$

Sous une forme matricielle, les p équations deviennent

$$\begin{bmatrix} \Phi(1,1)\ \Phi(1,2)....\Phi(1,p) \\ \Phi(2,1).\Phi(2,2)....\Phi(2,p) \\ . \quad ..... \\ . \quad ..... \\ \Phi(p,1)\Phi(p,2)....\Phi(p,p) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ . \\ . \\ a_p \end{bmatrix} = \begin{bmatrix} \Psi(1) \\ \Psi(2) \\ . \\ . \\ \Psi(p) \end{bmatrix} \tag{2.17}$$

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

La matrice Φ n'est pas une matrice de Toeplitz, 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

$$\Phi = LU \tag{2.18}$$

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

$$\Phi = CC^T \tag{2.19}$$

où $C = L$ et $C^T = U$. Le vecteur a est obtenu en résolvant d'abord l'équation (2.20)

$$Ly = \Psi \tag{2.20}$$

Puis:

$$Ua = y \tag{2.21}$$

2.2.3 Co nsidérations pratiques

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

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

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 et 16 kHz pour les applications de synthèse.

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

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

Les durées des trames d'analyse et leur décalage sont choisies inférieures à 30 ms. Les valeurs choisies sont liées au caractère quasi-stationnaire du signal parole.

Enfin, pour compenser les effets de bord, on multiplie général préalablement chaque trame d'analyse par une fenêtre de pondération *w(n)*, la fenêtre la plus utilisée reste celle de Hamming (équation 2.7).

2.4 Représentation des Paramètres Spectraux

Les coefficients de prédiction linéaire (LP) sont calculés à base de "bloc par bloc", généralement sur des trames de 5-40ms [9]. 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 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) [9] et les LSPs(Line Spectrum Pairs) [10]. Elles seront détaillées dans ce qui va suivre.

2.3.1 Les Coefficients de Réflexion et les LARs

Les coefficients de réflexion sont obtenus à partir de la procédure récursive de Wiener-Levinson-Durbin (WLD) [8][9] qui utilise une matrice de Toeplitz R. En résolvant l'ensemble des équations ordonnées récursivement de $m = 1,2,\cdots, p$, on aura

$$k_m = \frac{R(m) - \sum_{k=1}^{m-1} a_{m-1}(k) R(m-k)}{E_{m-1}} \tag{2.22}$$

$$a_m(m) = k_m \tag{2.23}$$

$$a_k(m) = a_k(m-1) - k_m a_{m-k}(m-1) \tag{2.24}$$

$$E_m = (1 - k_m^2) E_{m-1} \tag{2.25}$$

où initialement $E_0 = R(0)$ et $a_0 = 0$. A chaque cycle m, les coefficients $a_m(k)$ décrivent le prédicteur linéaire optimal d'ordre m. Puisque E_m, une erreur quadratique, n'est jamais négative, $|k_m| \prec 1$. Cette condition sur les coefficients de réflexion garantit aussi la stabilité du filtre de synthèse LP. Les valeurs négatives des coefficients de réflexion sont appelées les corrélations partielles (Partial Correlation) ou coefficients PARCOR.

On peut trouver les coefficients de réflexion à partir des coefficients LP $a_k = a_p(k)$, en calculant de manière récursive les deux équations suivantes pour $m = p, p-1, \cdots, 3, 2$:

$$a_{m-1}(i) = \frac{a_m(i) k_m a_m(m-i)}{1 - k_m^2}, \quad 1 \le i \le m-1 \tag{2.26}$$

$$k_{m-1} = a_{m-1}(m-1) \tag{2.27}$$

Un désavantage des coefficients de réflexion est la forme de leur sensibilité spectrale qui possède de grandes valeurs si l'amplitude des coefficients est proche de l'unité. Cependant,

ce désavantage peut être contourné par les transformations non linéaires qui élargissent la région au voisinage de la valeur $|k_m| = 1$. Deux de ces transformations sont les LARs [9] et la transformation sinus inverse [11]. Les LARs sont définis par

$$g_m = \log\left(\frac{1+k_m}{1-k_m}\right), \quad 1 \le m \le p. \tag{2.28}$$

La reconversion en coefficients de réflexion peut se faire par l'expression suivante

$$k_m = \frac{e^{g_m} - 1}{e^{g_m} + 1}, \; 1 \le m \le p. \tag{2.29}$$

A partir des coefficient de réflexion, on peut définir les coefficients arcsinus par

$$j_m = \arcsin(k_m), \; 1 \le m \le p. \tag{2.30}$$

2.3.2 Les Paires de Lignes Spectrales LSPs

Une autre représentation des coefficients LP, appelés les LSPs (Line Spectrum Pairs), a été introduite par Itakura [10]. Les LSPs sont les solutions des deux équations suivantes :

$$\left.\begin{matrix} P(z) \\ Q(z) \end{matrix}\right\} = A(z) \pm z^{-(p+1)} A\left(z^{-1}\right) \tag{2.31}$$

avec

$$A(z) = \frac{1}{2}[P(z) + Q(z)], \tag{2.32}$$

Soong et Juang [12] ont montré que si $H(z)$ est stable, où $A(z)$ est à phase minimale, alors les racines de $P(z)$et $Q(z)$ se trouvent sur le cercle unité et sont alternées entre les deux polynômes lorsque w augmente. Les LSPs correspondent à des positions angulaires. Les racines apparaissent sous forme de paires conjuguées et par conséquent, il existe p LSPs positionnés entre 0 et π. Il a été montré dans [12] que si les p LSPs, notés par w_i, sont dans un ordre descendant et uniques, alors le filtre inverse $A(z)$correspondant est à phase minimale, ce qui garantit la stabilité.

$$0 \prec w_1 \prec w_2 \prec \cdots \prec w_p \prec \pi \ [radians/\sec] \tag{2.33}$$

Kabal et Ramachandran [13] ont utilisé les polynômes de Chebyshev pour calculer les LSPs. Ils sont disposés sur un cercle unité du plan z et entrelacés entre valeurs paires et impaires [14].

$$T_m(x) = \cos(mw) \tag{2.34}$$

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

$$G'(x) = 2\sum_{i=0}^{l} g_i T_{l-i}(x), \tag{2.35}$$

$$H'(x) = 2\sum_{i=0}^{m} h_i T_{m-i}(x), \tag{2.36}$$

avec $l = m = p/2$ lorsque p est pair et $l = (p+1)/2$ et $m = (p-1)/2$ lorsque p est impair. Les racines de ces polynômes de Chebyshev vont donner les LSPs après une transformation inverse $w = \arccos(x)$. Les racines sont déterminées itérativement par la recherche des changements de signe dans les expressions de Chebyshev sur l'intervalle [-1, 1]

2.4 Mesure de la Qualité

La qualité du signal peut être évaluer par deux types de mesures: les mesures objectives et les mesures subjectives.

2.4.1 Mesures subjectives de la qualité de la parole

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 de la qualité généralement utilisées.

o Le test DRT (Diagnostic Rhyme Test) est une mesure d'intelligibilité dont la tâche du sujet est de reconnaître un de deux mots possibles dans un ensemble de paires minimales comme par exemple, meat – heat [15].

o Le test DAM (Diagnostic Acceptability Measure) évalue la qualité d'un système de communication sur la base d'acceptabilité de la parole comme perçue par un écouteur entraîné.

o Le test MOS (Mean Opinion Score) est une mesure très utilisée pour quantifier la qualité de la parole codée. Le MOS utilise 12 à 24 écouteurs [16] (les tests CCITT et TIA utilisent 23-64 écouteurs) qui sont entraînés à débiter phonétiquement des records selon une échelle de cinq niveaux comme représenté par le tableau 2.1.

Tableau 2.1- Description de l'échelle MOS [17]

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

2.4.2 Mesures Objectives de la Qualité de la Parole

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

2.4.2.1 Mesures Objectives de Distorsions dans le Domaine Temporel

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

o Le rapport signal à bruit SNR (Signal to Noise Ratio) reste la mesure objective de la qualité la plus couramment utilisée pour les codeurs qui essaient de préserver la forme du signal

Si $s(n)$ est le signal de parole original et $\hat{s}(n)$ le signal de parole synthétisé, alors le signal erreur est donné par l'expression suivante:

$$e(n) = s(n) - \hat{s}(n) \tag{2.37}$$

Pour un signale de N échantillons, on définit l'énergie du signal et de l'erreur par les expression suivantes

$$E_S = \sum_{n=-\infty}^{+\infty} s^2(n) \tag{2.38}$$

$$E_e = \sum_{n=-\infty}^{+\infty} e^2(n) \tag{2.39}$$

Le rapport signal à bruit (*SNR)* est alors donné par

$$SNR = 10\log_{10}\left(\frac{E_S}{E_e}\right) dB \tag{2.40}$$

o Le SNR Segmental (SEGSNR):

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

$$SEGSNR = \frac{1}{M}\sum_{j=0}^{M-1} 10\log_{10}\left(\frac{\sum_{n=1}^{N} s^2(n)}{\sum_{n=1}^{N} e^2(n)}\right) \text{dB} \tag{2.41}$$

Le *SEGSNR* est meilleur que le *SNR*. Mais ne constitue pas une bonne mesure lorsque toute la trame est constituée de silences. Ces types de trames causent un grand SNR négatif, qui

va biaiser l'ensemble des performances. Pour remédier à ce problème, des valeurs seuils peuvent être utilisées pour détecter les silences proches et les écarter.

D'autres mesures objectives dans le domaine temporel sont également utilisées comme la prédiction du gain, l'erreur énergie et les pourcentages des distorsions (outliers) statistiques.

2.4.2.2 Mesures Objectives de Distorsions dans le Domaine Fréquentiel

La mesure de distorsion $d(x,y)$ entre deux vecteurs de la parole x et y respectivement satisfait les conditions suivantes [18]:

$$d(x,y) \geq 0 \tag{2.42}$$

$$d(x,y) = 0 \text{ si } x = y \tag{2.43}$$

Une mesure plus rigoureuse est la mesure de distance ou métrique, qui doit satisfaire les conditions de symétrie et l'inégalité angulaire

$$d(x,y) = d(y,x) \tag{2.44}$$

$$d(x,z) \leq d(x,y) + d(y,z) \tag{2.45}$$

En général, l'ensemble de mesures de performance est la moyenne des distances:

$$D = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} d(x_i, y_i), \tag{2.46}$$

Une mesure de distorsion doit avoir quelques significations importantes selon les propriétés spectrales du signal parole. Cette mesure se fait généralement sur des trames de 5-30 ms de longueur.

La distorsion ou la différence entre les deux spectres affecte la perception du son. Dans ce qui suit, les différences entre les enveloppes spectrales originale et codée peuvent perceptuellement conduire à des sons phonétiquement différents:

- Si les formants de l'enveloppe spectrale du signal original et les formants de l'enveloppe spectrale du signal codé possèdent des fréquences sensiblement différentes.
- Si les bandes passantes des formants de ces enveloppes spectrales sont très différentes.

Une description très brève des différentes mesures de distorsions dans le domaine fréquentiel est présentée dans ce qui suit

o La mesure de distorsion spectrale Logarithmique (Log spectral distortion measure) la plus fréquemment utilisée, appelée souvent distorsion spectrale, basée sur la norme L_p est exprimée par

$$d_{SD}^{p} = \frac{2}{2\pi}\int_{\pi}^{\pi}\left|10\log S(w) - 10\log \hat{S}(w)\right|^{p} dw \qquad (2.47)$$

où le spectre de l'amplitude fréquentielle $S(w)$ est donné par

$$S(w) = \frac{G}{\left|A(e^{jw})\right|^{2}} = \frac{G}{\left|1-\sum_{n=1}^{P} a_n e^{jnw}\right|^{2}} \qquad (2.48)$$

G et $a_n\ \{n = 1,\cdots,p\}$ sont respectivement le facteur de gain et les coefficients du filtre LP. Lorsque $p = 2$

$$d_{SD} = \sqrt{\frac{1}{w_u - w_l}\int_{w_l}^{w_u}\left|10\log_{10}\frac{S(w)}{\hat{S}(w)}\right|^{2} dw} \quad \text{dB} \qquad (2.49)$$

où w_l et w_u définissent respectivement les fréquences inférieure et supérieure. Idéalement, w_l est égale à zéro et w_u correspond à la moitié de la fréquence d'échantillonnage.
En pratique, la distance spectrale logarithmique est calculée sur des valeurs discrètes d'une bande passante limitée. Pour un signal de parole échantillonné à 8 kHz et d'une bande passante de 3 kHz, la distance spectrale logarithmique est une sommation sur un ensemble de 96 points uniformément espacés de 0 Hz à 3 kHz. Ceci peut s'exprimer par

$$d_{SD} = \sqrt{\frac{1}{n_2 - n_1}\sum_{n=n_1}^{n_2-1}\left|10\log_{10}\frac{S\left(e^{j2\pi n/N}\right)}{\hat{S}\left(e^{j2\pi n/N}\right)}\right|^2} \quad \text{dB} \tag{2.50}$$

avec $N = 256$, n_1 et n_2 correspondent respectivement à 1 et 96.

o La distorsion d'Itakura-Saito connue sous le nom de mesure de distance du rapport de vraisemblance (likelihood ratio distance measure) mesure le rapport d'énergie entre le signal résiduel obtenu en utilisant le filtre LP avec les coefficients quantifiés et le signal résiduel obtenu en utilisant le filtre LP avec les coefficients non quantifiés.

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

avec:

$$V(\omega) = \log(s(\omega) - \log(\hat{s}(\omega))) \tag{2.52}$$

L'intégrale 2.51 peut s'ecriresous la forme du polynôme

$$d_{IS} = \left(\frac{G}{\hat{G}}\right)^2 \frac{\hat{a}^T R \hat{a}}{a^T R a} - 2\log\left(\frac{G}{\hat{G}}\right) - 1 \tag{2.53}$$

Avec $a = [1, a_1, a_2, \cdots, a_p]^T$, $\hat{a} = [1, \hat{a}_1, \hat{a}_2, \cdots, \hat{a}_p]^T$ et R la matrice des autocorrélations. Lorsque les gains sont supposés être égaux, la mesure d'Itakura-Saito est simplifiée comme suit:

$$d_{IS} = \frac{\hat{a}^T R \hat{a}}{a^T R a} - 1 \tag{2.54}$$

Cependant, la mesure d'Itakura-Saito n'est pas symétrique. Pour la symétrie une mesure d'Itakura modifiée peut être utilisée, elle s'écrit

$$d_{IS} = \frac{1}{2}\left[\frac{\hat{a}^T R \hat{a}}{a^T R a} - \frac{\hat{a}^T \hat{R} a}{a^T \hat{R} \hat{a}} - 2\right] \tag{2.55}$$

Un terme de poids peut être introduit à la mesure d'Itakura-Saito pour tirer avantage des propriétés de discrimination perceptuelle de l'oreille de l'être humain. Il est formulé comme suit

$$d_{WIS} = \frac{1}{2\pi}\int_{\pi} W\left(e^{j\omega}\right)\left[e^{V(\omega)} - v(\omega) - 1\right]d\omega \tag{2.56}$$

Quelques schémas de poids de $W\left(e^{j\omega}\right)$ sont proposés dans [19].

o La distance euclidienne pondérée est utilisée dans le domaine des LSPs, car les LSPs possèdent une bonne correspondance entre la forme de l'enveloppe, les formants et les

vallées. Pour accentuer une portion particulière du spectre, les LSPs de cette partie se verront attribués plus de poids comparés aux autres. Si f et $\hat{f}$ sont respectivement deux vecteurs des LSPs original et codé, alors leur distance euclidienne $d(f,\hat{f})$ est définie par

$$d(f,\hat{f}) = \|f-\hat{f}\|^2 \tag{2.57}$$

si une analyse LP d'ordre p est employée, alors l'Equation 2.57 devient

$$d(f,\hat{f}) = \sum_{i=1}^{p}(f_i-\hat{f}_i)^2 \tag{2.58}$$

La distance euclidienne pondérée par des des poids w_i s'écrit

$$d(f,\hat{f}) = \sum_{i=1}^{p} w_i(f_i-\hat{f}_i)^2 \tag{2.59}$$

Paliwal et Atal [20] ont défini le poids w_i comme suit

$$w_i = [S(f_i)]^r \tag{2.60}$$

où $S(f)$ est le spectre de puissance LP et r est une constante empirique qui contrôle les poids relatifs des LSPs. Expérimentalement, 0.15 est une valeur satisfaisante de r. Par conséquent, dans ce schéma le poids dépend de la valeur du spectre de puissance LPC à la position de cet LSP. Les formants de grandes amplitudes sont allouées plus de poids que les formants de faibles amplitudes. Les vallées sont attribuées moins de poids.

En plus, on sait que l'oreille est plus sensible aux basses fréquences que les hautes fréquences. Pour exploiter cette propriété, les basses fréquences de LSPs doivent avoir plus de poids. Paliwal et Atal [20] ont introduit un nouveau terme c_i dans l'équation (2.59) pour réduire la distance euclidienne pondérée

$$d(f,\hat{f}) = \sum_{i=1}^{p} c_i w_i \left(f_i - \hat{f}_i\right)^2 \tag{2.61}$$

avec,

$$c_i = \begin{cases} 1.0, & 1 \le i \le 8, \\ 0.8, & i = 9, \\ 0.9, & i = 10. \end{cases} \tag{2.62}$$

o Une autre distance proposée par Laroia et al.[21] appelée distance moyenne harmonique inverse (IHM) est exprimée comme suit

$$\frac{1}{(lsp(i+1) - lsp(i))} + \frac{1}{(lsp(i) - lsp(i-1))} \tag{2.63}$$

Les mesures objectives ne peuvent remplacer les tests subjectifs, mais ils peuvent aider dans le développement de nouveaux algorithmes.

2.5 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 largement utilisée dans le codage du signal de parole. Nous avons utilisé le concept des "geostatistics" et son schéma linéaire prédictif pour l'analyse par prédiction linéaire de la parole[22].

Chapitre 3
Codage Intra-trame des Paramètres LSPs

3.1 Introduction

Dans ce chapitre, nous introduisons la notion de quantification vectorielle qui est largement utilisée dans les codeurs de la parole à bas débit. En particulier, le codage des paramètres spectraux à base de trame par trame avec la quantification vectorielle, est une étape efficace dans la compression des signaux. Nous introduisons trois schémas du Split VQ comme solution au problème de la complexité qui se pose dans la quantification vectorielle. Une solution au problème de la robustesse face au bruit est présentée également dans ce chapitre.

3.2 La Quantification

La quantification est l'opération de discrétisation d'une ou plusieurs variables, c'est aussi l'approximation de la valeur instantanée exacte d'un signal par la plus voisine valeur tirée d'un ensemble de N valeurs discrètes [23].

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

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

3.2.1 La Quantification Scalaire

La quantification scalaire (QS) consiste à quantifier séparément chaque échantillon du signal d'entrée. Comme l'illustre la figure 3.1, un échantillon x du signal d'entrée est spécifié par l'indice i s'il tombe dans l'intervalle suivant:

$$I_i : \{x_i < x \le x_{i+1}\} \qquad i = 1,2, \ldots, N \tag{3. 1}$$

où x_i et x_{i+1} sont *les niveaux de décision* ou *seuils*.

Tous les échantillons situés dans l'intervalle I_i seront remplacés par une valeur y_i appelée *niveau de reconstruction* ou *représentant*. On peut par conséquent utiliser N niveaux de reconstruction pour représenter les échantillons du signal d'entrée compris entre les valeurs x_1 et x_{N+1}.

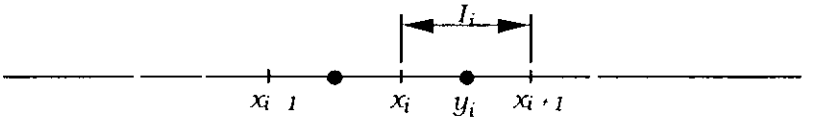

Figure 3.1 Quantification scalaire

Les niveaux de reconstruction peuvent s'écrire:

$$y = y_i \quad \text{si } x \in I_i \tag{3. 2}$$

ou s'exprimer par la fonction suivante

$$y = Q(x) \tag{3. 3}$$

L'expression (3.3) montrant la relation de quantification entre x et y est appelée *caractéristique de quantification.* La figure 3.2 montre que la fonction $y = Q(x)$ de la quantification scalaire est en marche d'escalier.

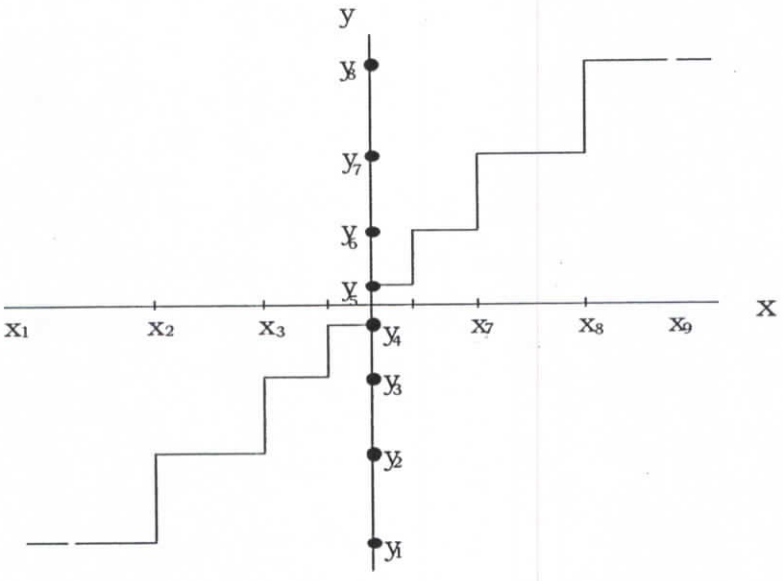

Figure 3.2 Caractéristique typique d'un quantificateur scalaire.

La SQ possède l'avantage de nécessiter une faible mémoire de stockage et une complexité minimale.

3.2.2 La Quantification Vectorielle

La quantification vectorielle (VQ) est l'extension de la quantification scalaire à un espace multidimensionnel. Shannon [24][25] a montré que pour un débit donné, le codage de blocs longs de l'information atteindra toujours de meilleures performances en terme de distorsions. L'amélioration de performance de la VQ par rapport à la SQ est le résultat de son habilité à exploiter toute corrélation (linéaire ou non linéaire) entre les composantes du vecteur et d'imiter la forme de la densité du vecteur source [26]-[28].

La collection des représentations possibles d'un vecteur est dite *dictionnaire* de R^k . On utilise en général plus d'un dictionnaire pour représenter le vecteur. Plusieurs procédures ont été proposées pour créer, organiser et tester les dictionnaires [23].

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'entée sont quantifiés sur N niveaux et codés avec R bits. Contrairement à la quantification scalaire, un quantificateur vectoriel peut fonctionner avec un débit fractionnaire ($R<1$).

La quantification vectorielle contient deux parties: codage et décodage. La figure 3.3 illustre le principe de la quantification vectorielle.

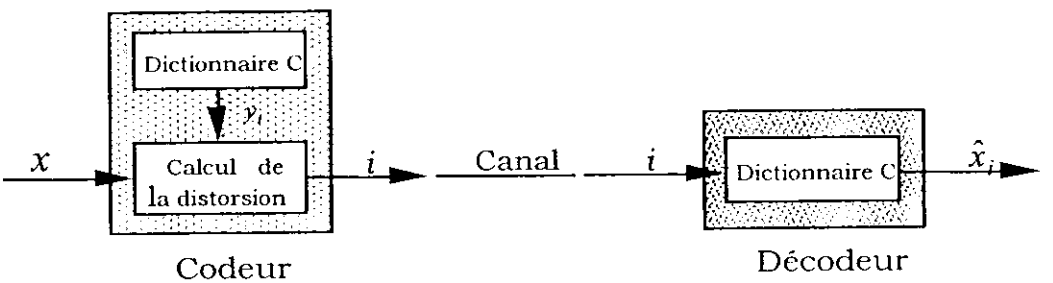

Figure 3.3 Principe de la quantification vectorielle.

Dans le codeur, on associe au vecteur d'entrée $\mathbf{x}$ un mot de code $\mathbf{y}_i$ du dictionnaire C selon le critère du plus proche voisin. Seul l'indice i est transmis au décodeur. On a N indices correspondant aux N mots de code du dictionnaire. Dans le décodeur, on utilise le même dictionnaire, c'est à dire $\hat{x}_i = y_i$. Le mot de code y_i est retrouvé à partir de l'indice i reçu, et il deviendra alors représentant du vecteur x.

Dans la quantification vectorielle, on utilise le fait que les composantes du vecteur d'entrée ont généralement une corrélation et qu'elles sont reparties principalement dans une certaine

région de l'espace R^k. On peut atteindre une plus petite valeur de distorsion avec le même nombre de vecteurs de reproduction qu'en quantification scalaire en construisant convenablement le dictionnaire. Autrement dit, on obtiendra la même distorsion avec ces deux types de quantification en utilisant un plus petit nombre de vecteurs de reproduction. Le signal d'entrée pourra donc être représenté et transmis à plus bas débit. Même si le signal d'entrée est non corrélé, c'est à dire que les composantes du vecteur sont indépendantes, il y a avantage à utiliser la quantification vectorielle.

Il n'y a rien de mystérieux à considérer des espaces de grandes dimensions; il suffit de savoir que tout s'organise autour des coordonnées des vecteurs et qu'il n'y a pas lieu de s'imposer une représentation géométrique. A titre d'illustration, nous précisons qu'un vecteur de l'espace R^m est simplement une matrice colonne constituée de m nombres réels $x_i : x = (x_1, x_2, ..., x_m)^T$. Par exemple, une sphère entièrement caractérisée par son centre $u = (u_1, u_2, ..., u_m)^T$ et son rayon ρ, est constituée de points dont les coordonnées vérifient la relation (3.4):

$$\sum_{i=1}^{m} (x_i - u_i)^2 = \rho^2 \tag{3.4}$$

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

$$D_Q = E[d(x, Q(x))] = \int_{-\infty}^{+\infty} d(x, Q(x).p(x).dx \tag{3.5}$$

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

$$D_Q \cong \frac{1}{N}\sum_{j=1}^{N} d(x_j, Q(x_j)) \tag{3.6}$$

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

$$S_i = Q^{-1}(y_i) = \{x : Q(x) = y_i\} \tag{3.7}$$

Les régions S_i sont appelées des régions de Voronoï, de Dirichlet ou du plus proche voisin. Nous appellerons centroïde de la classe S_i le vecteur c_i tel que sa distance moyenne à tous les éléments de la classe soit minimale (analogue au centre de gravité en géométrie euclidienne)

$$E[d(x,c_i); x \in S_i] = \text{Inf}\ \{E[d(x,x_i); x \in S_i]\} \tag{3.8}$$

Étant données une distance et une taille du dictionnaire, on cherche un quantificateur optimal qui minimise la distorsion moyenne ou qui se rapproche de la valeur optimale.

3.2.2.1 Les Conditions pour l'Optimalité

Les performances d'un VQ dépendent de la partition de l'espace du codeur et les vecteurs de reproduction ou codevecteurs du décodeur. Un quantificateur vectoriel est optimal lorsque la distorsion moyenne $E[d(x,Q(x))]$ est minimisée pour une séquence d'entrée X. Deux conditions sont nécessaires pour l'optimalité du dictionnaire durant le design : une pour le codeur et l'autre pour le décodeur. Ces conditions sont la condition du plus proche voisin et la condition du centroid qui ont été introduites pour la première fois par Lloyd pour la conception d'un QS [23].

Le quantificateur est dit localement optimal s'il vérifie les exigences suivantes:

o Un codage optimal (pour un dictionnaire fixé) respectant " la règle du plus proche voisin " que nous allons décrire ;

Le décodage optimal (pour une partition S_i donnée): le vecteur représentant y_i doit minimiser la distorsion associée au voronoï S_i, y_i est donc le centroïde de cette cellule : $y_i = cent(S_i)$.

Condition du plus proche voisin

Etant donné un décodeur et son ensemble fini de mots codes de sortie C, les classes de partition S_i du codeur sont optimales si, elles vérifient la condition suivante :

$$S_i \subset \left\{x \in R^n \middle| d(x, y_i) \leq d(x, y_j); \forall i \neq j\right\} \quad i = 1,2, \ldots, N \tag{3.9}$$

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

$$Q(x) = y_i \text{ seulement si } d(x, y_i) \leq d(x, y_j) \ \forall i \neq j \quad i = 1,2, \ldots, N \tag{3.10}$$

Condition du centroïde

Etant donné une partition d'encodeur $P = \{S_i | i = 1, \cdots, N\}$, les mots codés optimaux y_i dans C sont les centroïdes dans chaque partition S_i :

$$\begin{aligned} y_i &= Cent(S_i) \\ y_i &= \min E(d(x, y) | x \in S_i) \end{aligned} \tag{3.11}$$

3.2.3 Construction d'un Quantificateur Statistique

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

Dans le cas ou les poins d'entrée sont distribués d'une façon non uniforme, on adoptera une approche statistique visant à tirer partie de cette non-uniformité. Le dictionnaire sera construit par apprentissage : à partir d'une large base de vecteurs d'entrée où sera sélectionné un nombre réduit de points susceptibles d'en refléter les propriétés statistiques.
En revanche, si la distribution des vecteurs d'entrée est plutôt uniforme, on aura intérêt à conférer à l'espace de représentation une structure mathématique forte, indépendamment de la réalité des données à traiter. Cette approche algébrique utilise généralement les propriétés des réseaux réguliers de points.

3.3 Quantification Vectorielle par Split des Paramètres LSPs

D'un point de vue théorique, la quantification idéale, en termes de performances, consiste à quantifier le vecteur des dix LSPs en une seule entité. Par conséquent, la distorsion produite par un VQ des 10 LSPs sera inférieure à celle obtenue par une SQ à n'importe quel débit. Juang et al. [29] ont étudié la VQ des paramètres LPC avec la mesure de distorsion de vraisemblance et ont monté que les performances obtenues avec une VQ à 10 bits/ trame sont comparables à celles d'une SQ à 24 bits/frame. Ce VQ possède une distorsion spectrale moyenne de 3.35 dB, ce qui n'est pas acceptable pour une haute qualité de la parole dans les codeurs. Pour une qualité transparente de la parole 24 bits sont nécessaires pour quantifier une trame de la parole. Or, en VQ à recherche complète (full-search), un dictionnaire contenant $N = 2^b$ mots codes est utilisé pour quantifier un vecteur x de dimension k par un débit de r bits par composante vecteur avec $b = rk$ bits par vecteur. Cependant, en augmentant le nombre de bits par vecteur, la taille du dictionnaire ainsi que la complexité de ce type de recherche dans le dictionnaire croient de manière exponentielle. Plusieurs schémas de quantificateurs sous-optimales qui réduisent la complexité et la taille du dictionnaire au détriment d'une dégradation de la qualité transparente de la parole ont été longuement étudiés. On peut citer la quantification vectorielle par Split (SVQ) qui est une forme de la technique de code produit [20][23][30], et le multi-étages VQ [31][32]. Nous avons étudié ce problème et les résultats trouvés sont publiés dans [35]-[47]. Plusieurs autres méthodes de quantifications des paramètres LSPs ont été proposées pour avoir une qualité transparente de la parole à bas débit [48]-[58].

Nous proposons dans ce travail, l'étude de trois structures de SVQs, qui sont représentées par les Figures 3.4, 3.5 & 3.6. Dans ces structures les vecteurs des dix LSPs sont divisés en trois sous vecteurs Q1, Q2 et Q3. Le premier sous vecteur contient les trois premiers LSPs, le second contient les trois suivants et le dernier sous vecteur contient les quatre derniers LSPs. La quantification selon le schéma de la figure 3.4 utilise un quantificateur vectoriel sur 8-9 bits pour quantifier les trois premiers LSPs ensuite un quantificateur vectoriel sur 8-9 bits pour quantifier les trois différences (LSP4-LSP3Q; LSP5-LSP3Q; LSP6-LSP3Q) et finalement un quantificateur vectoriel sur 8-9 bits pour quantifier les quatre dernières différences LSP7-LSP6Q; LSP8-LSP6Q; LSP9-LSP6Q; LSP10-LSP6Q). Le second schéma (figure 3.5) utilise un quantificateur vectoriel sur 8-9 bits pour les quatre derniers LSPs ensuite un quantificateur vectoriel sur 8-9 bits pour quantifier les trois différences (LSP7Q-LSP4; LSP7Q-LSP5; LSP7Q-LSP6) et finalement un quantificateur vectoriel sur 8-9 bits pour quantifier les trois différences (LSP4Q-LSP1; LSP4Q-LSP2; LSP4Q-LSP3). Le premier et le second schéma semblent intéressants puisq'ils exploitent la corrélation intra-trame en quantifiant les différences. Cependant ces deux solutions restent compliquées car l'utilisation des différences peut perturber la distribution des LSPs dans tout l'espace. On propose alors comme schéma de comparaison le SVQ (3-3-4) (figure 3.6). Ces trois solutions présentées quantifient les LSPs d'une manière simple et relativement efficace sans nécessiter une grande capacité de stockage puisque la taille des dictionnaires utilisés ne dépasse pas 512 mots de code. Malgré le fait que ces trois quantificateurs qui s'étalent sur 24 à 28 bits ne semblent pas les plus optimales, ils possèdent les avantages de rapidité, de robustesse et d'une faible capacité de stockage [59].

Devant ce type de quantification se posent divers problèmes. Par exemple, une fidèle reconstitution du signal parole nécessite un bon ordonnancement des LSPs pour assurer la stabilité du filtre LP. Or, l'utilisation des trois quantificateurs proposés peut dans certaines cas modifier l'ordre des LSPs aux frontières des sous vecteurs. Ceci cause une instabilité du filtre LP puisque ce dernier possèdera des pôles en dehors du cercle unité. Nous proposons deux solutions face à ce type de problème (figure 3.7). La première consiste à effectuer la quantification sans se préoccuper de l'inversion. Après l'opération de quantification, l'ordre est vérifié. Dans le cas d'une inversion, les deux LSPs concernés sont permutés et suffisamment espacés pour assurer la stabilité du filtre LP. La seconde solution consiste à

restreindre le choix, au sein des dictionnaires Q1, Q2 et Q3 aux vecteurs des LSPs qui n'engagent pas la stabilité du filtre LP. De cette manière, le LSP choisi sera le vecteur le plus proche vérifiant la condition de non-inversion des LSPs. Ceci revient à diminuer la taille des dictionnaires utilisés. Ce type de problème se produit rarement, néanmoins il risque de réduire la qualité de l'écoute.

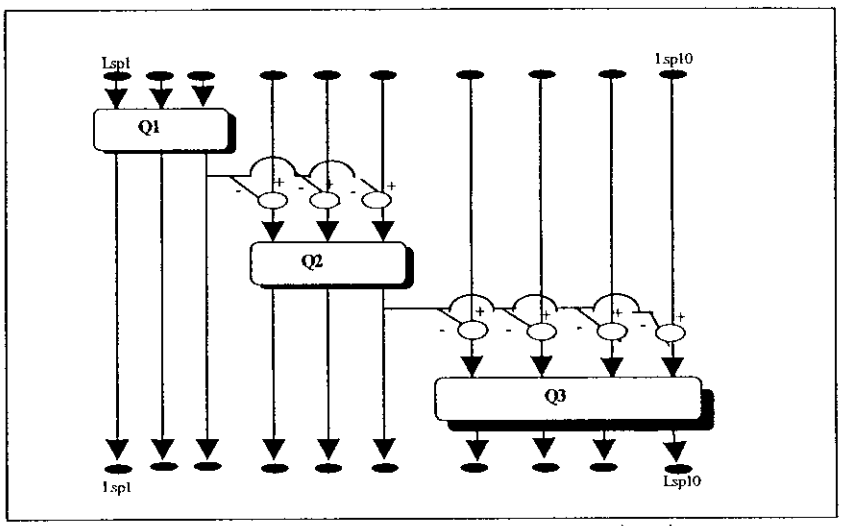

Figure 3.4 Quantificateur vectoriel par Split (3- Δ 3- Δ 4)

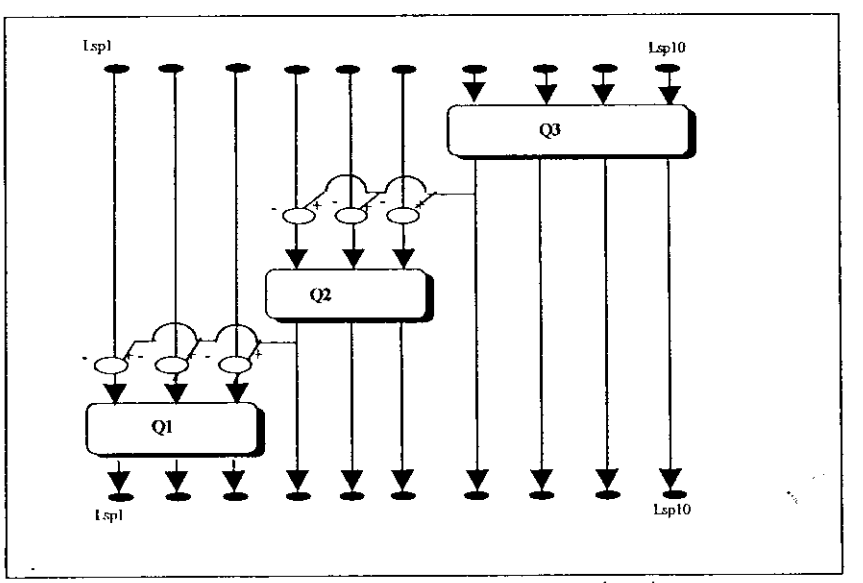

Figure 3.5 Quantificateur vectoriel par Split (Δ 3- Δ 3-4).

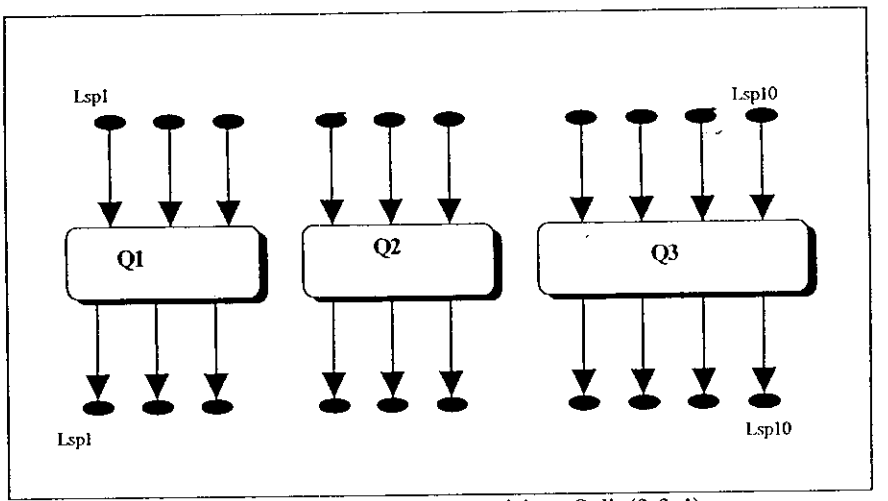

Figure 3.6 Quantificateur vectoriel par Split (3-3-4)

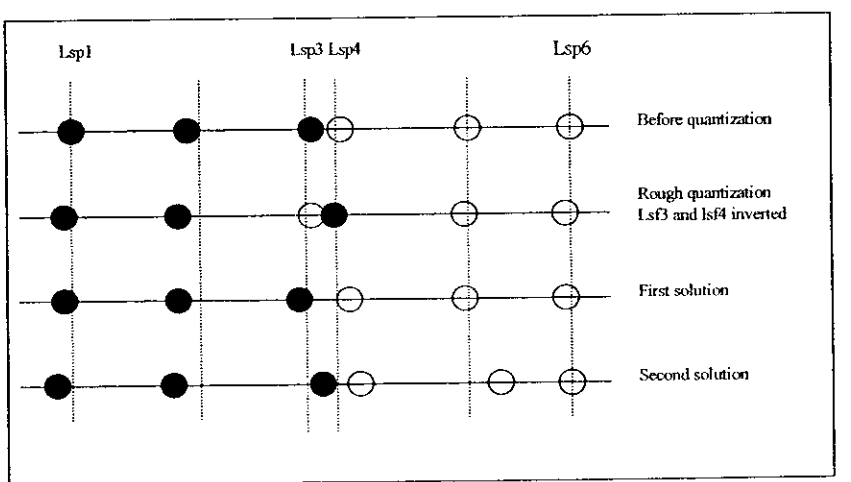

Figure 3.7 Solutions au problème d'inversion des LSPs

Plusieurs études de la quantification des paramètres LSPs peuvent être trouvées dans la littérature [20][21][54]-[58]. Cependant, les résultats ne peuvent être comparables en général puisqu'il peut exister de larges déviations dues à l'expérimentation. Nous avons remarqué que différentes bases de données peuvent conduire à des performances objectives différentes pour la même méthode de quantification. En plus, il existe plusieurs méthodes pour effectuer l'analyse LPC. Par exemple, la méthode d'autocorrelation et de covariance stabilisée sont très communes en analyse LPC, la procédure d'expansion de la bande passante (bandwidth expansion) affecte aussi les résultats. La longueur des trames qui varie de 5-40 ms dans

différents articles, et la la fenêtre d'analyse de recouvrement (analysis window overlap) diffère d'un travail à un autre. Dans nos simulations nous avons utilisé des fenêtres de 256 échantillons avec un overlap de 50%. En ce qui concerne l'expansion de la bande passante (bandwidth expansion), nous avons utilisé la méthode présentée dans [61] qui consiste à ajoute une fenêtre d'autocorrelation appelée "Lag-window". Cette fenêtre permet d'atténuer les fortes résonances sans altérer le reste du signal contrairement aux compensations fréquentielles classiques qui agissent sur la totalité du signal.

Dans le but de comparer les performances des trois structures présentées, nous avons utilisé la distorsion spectrale moyenne exprimée en dB^2 par l'expression suivante:

$$SD = \frac{1}{N_f} \sum_{n=1}^{N_f} \left(\frac{1}{\pi} \int_0^{\pi} \left[\log S_n(w) - \log \hat{S}_n(w) \right]^2 dw \right), dB^2 \tag{3.14}$$

où N_f est le nombre total des trames. Cette mesure de distorsion spectrale est connue pour avoir une bonne correspondance avec les mesures subjectives [61]. La mesure de distorsion utilisée dans la littérature est celle donnée par l'équation (2.50) , mais cette dernière a le défaut de masquer légèrement les trames dont la distorsion spectrale est considérablement différente de la valeur moyenne, c'est à dire diminuer la variance de la distribution des distorsions. Donc la première distorsion permet une évaluation plus objective et c'est celle ci que nous utiliserons lors de nos simulations.

Les dictionnaires utilisés sont obtenus par l'algorithme de la *k*-moyenne [62]. Les résultats de simulations sont représentés dans tableau 3.1. Ces résultats montrent que la meilleure structure est la division 3-3-4.

Tableau 3.1 Performances en terme de distorsion spectral (DS) des trois split QV

Schéma de Quantization	Débit (Bits/trame)	Av. SD (dB^2)
3-Δ3-Δ4	24	1.568
Δ3-Δ3-4	24	1.543
(3-3-4)	24	1.342

En effet, les dictionnaires construits à partir des différences sont moins optimaux. On en déduit que l'obtention des ΔLSP disperse quelque peu les vecteurs dans tout l'espace.

A fin de trouver la meilleure distance, en termes de performances, à utiliser, nous avons comparé trois mesures de distances, la première est la distance euclidienne équation 2.58, la seconde est la distance euclidienne pondérée équation 2.61 et la troisième est la distance moyenne harmonique inverse (IHM) équation 2.63 [21].

Les résultats de comparaison sont regroupés dans le tableau 3.2. On remarque bien que les résultats obtenus avec la distance euclidienne pondérée et la distance IHM sont en bon accord. On gardera pour la suite des simulations la distance euclidienne pondérée, puisqu'elle est simple comparée à la distance IHM. La distance IHM nécessite le calcule de la DFT (Discrete Fourier transform) après l'algorithme de Wiener Levinson Durbin (WLD).

Tableau 3.2. Quantization par split (3-3-4)
Pour Différentes mesures de distances

Distance	Av. SD (dB^2)
Distance Euclidienne	1.342
(IHM) distance	1.222
Distance Euclidienne pondérée	1.220

Les résultats obtenus avec le split 3-3-4 sont présentés dans le tableau 3.3.

Tableau 3.3. Performances pour différentes allocation de bits du QV par split (3-3-4)

Débit (bits/trame)	Bit Allocation (bits) Q1	Q2	Q3	Avg. SD (dB^2)
24	8	8	8	1.222
25	9	8	8	1.134
25	8	9	8	1.056
25	8	8	9	0.987
26	9	9	8	0.9635
26	9	8	9	0.8457
26	8	9	9	0.7564
27	9	9	9	0.6936

Table 3.4 Bit allocation for the scalar quantizer
at 34 bits/frames with IIM distance

Indice	1	2	3	4	5	6	7	8	9	10	total
Bits	4	4	4	4	3	3	3	3	3	3	34

Nos résultats sont exprimés en dB^2. Dans le but de faire une comparaison avec les résultats présentés dans les références et dB [20][21][54]-[58]qui sont exprimés en dB , nous avons calculé une nouvelle fois la distorsion relative au SVQ 3-3-4 à 24 bits/trame. Nous avons trouvé une distorsion spectrale moyenne de 1.23 dB avec 4.14 % les pourcentages des distorsions ([2-4]dB et >4dB). Une distorsion spectrale moyenne de 1.26 dB est obtenue pour le quantificateur scalaire à 34 bits/trame (tableau 3.4) avec 3.44 % et les pourcentages des distorsions ([2-4]dB et >4dB).

Pour le SVQ 4-6 présenté dans la référence [20], une qualité transparente de la parole de 1.03 dB avec 1.03 % et les pourcentages des distorsions ([2-4]dB et >4dB) est obtenue à 24 bits/trame avec la distance euclidienne pondérée. Ce qui reste complexe pour nos applications puisque les dictionnaires seront de 12 bits. Pour le quantificateur hybride scalaire-vectoriel décrit dans [63] et utilisé de [20], une qualité transparente comparable à celle du quantificateur à 24 bits/trame est obtenue à 31-32 bits/trame. Nos résultats sont meilleurs que ceux trouvé par [20] puisque notre objectif primordial est la réduction de la complexité. Cet objectif est atteint puisque les dictionnaires ne dépassent pas 8 bits et la distance utilisée est moins complexe que celle utilisée par [20]. Aussi, nous avons sauvé 8 à 9 bits par trame par rapport à la quantification scalaire à 34 bits/trame.

3. 4 La Robustesse

La robustesse contre les erreurs est un critère important dans la mesure des performances d'un quantificateur vectoriel. La robustesse est un critère sensible de la quantification vectorielle car une erreur sur un indice du dictionnaire peut conduire au choix du vecteur représentant complètement différent du vecteur original. La quantification scalaire dans tous les cas donne un représentant le plus proche du vecteur original. Nous présentons ici un algorithme qui permet de remédier à ce problème. Lorsque tous les bits de l'indice ont le même poids, il est insignifiant de trier le dictionnaire. Mais lorsqu'on considère que les 4 ou 5 bits de poids fort

sont convenablement protégés, alors on peut trier les dictionnaires en fonction des bits de poids faible. Il suffit de trier les vecteurs en fonction de leurs proximités relatives.

En supposant par exemple que sur les huit bits, seuls les quatre bits de poids faible peuvent être erronés (pas protégés). L'algorithme consiste à trier le dictionnaire en regroupant les vecteurs proches en terme de distance Euclidienne. On cherche à chaque étape le vecteur le plus proche des n vecteurs déjà triés. Les résultats obtenus pour les trois dictionnaires sont représentés dans tableau 3.5, 3.6 & 3.7 [59].

Il est clair que le tri apporte des améliorations remarquables sur les erreurs commises sur les quatre bits de poids faible.

Tableau 3.5. Résultas du premier dictionnaire

Distorsion en (dB^2) Selon le type De tri	sans tri	n =1	n =3	n =5	n =7
Modification du bit 0	16.07	2.07	3.07	2.04	2.56
Modification du bit 1	25.87	3.57	3.27	2.77	2.97
Modification Du bit 2	29.51	5.00	4.37	3.57	3.87
Modification Du bit 3	30.47	7.68	7.00	5.37	5.37

Tableau 3.6. Résultas du second dictionnaire

Distorsion en (dB^2) Selon le type De tri	sans tri	n =1	n =3	n =5	n =7
Modification De bit 0	26.47	3.47	4.37	7.10	5.60
Modification De bit 1	39.77	6.67	5.20	7.80	7.10
Modification De bit 2	47.47	13.47	10.17	9.27	8.87
Modification De bit 3	52.16	20.67	19.77	16.67	14.17

Tableau 3.7. Résultas troisième dictionnaire

Distorsion en(dB^2) Le type type De tri	Sans tri	n =1	n =3	n =5	n =7
Modification du bit 0	21.27	3.8	4.9	5.4	5.7
Modification du bit 1	26.5	7.5	5.7	6.5	6.4
Modification du bit 2	32.1	11.9	9.3	7.9	7.3
Modification du bit 3	33.5	18.0	13.5	12.4	10.9

3.5 Conclusion

Dans ce chapitre, nous avons examiné des schémas de la quantification vectorielle intra-trame par split dans le but de réduire la complexité. En particulier, trois schémas, Split $\Lambda 3/\Delta 3/4$ VQ, Split $3/\Delta 3/\Delta 4$ VQ et 3/3/4 VQ sont proposés et leurs performances comparées et testées individuellement. On a montré que la structure 3/3/4 VQ est la meilleure en terme de distorsion spectrale parmi les trois et que les performances en termes de complexité sont meilleures que celles existant dans la littérature. Nous avons présenté deux solutions aux problèmes d'inversion des LSPs qui peut surgir aux frontières des sous vecteurs après une quantification vectorielle par split. Nous avons montré que la distance IHM donne de meilleures performances e terme de distorsion spectrale que la distance euclidienne et la distance euclidienne pondérée. La IHM est simple à implémenter. Pour la robustesse, nous avons proposé une méthode contre les effets du bruit basée sur le tri des bits de poids faible des dictionnaires. Nous avons donc, réalisé un quantificateur à faible complexité et débit binaire et robuste contre les effets du bruit.

Chapitre 4

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

4.1 Introduction

Dans le monde des télécommunications modernes, la tendance récente est de remplacer les réseaux de circuits à commutation comme le PSTN (Public Service Telephone Network), conçu pour la transmission de la voix, par des réseaux à commutation par paquets comme l'Internet.

Cette recherche se focalise sur l'amélioration de la qualité de la parole dans la transmission de la voix à travers les réseaux IP (VoIP), qui est endommagée par le retard des paquets, les pertes de paquets et la gigue (Jitter) dans les réseaux IP. Dans ce chapitre, on présentera les systèmes VoIP et les différentes méthodes de masquage des erreurs.

4.2 La Voix sur les Réseaux IP

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

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 [64]-[66].

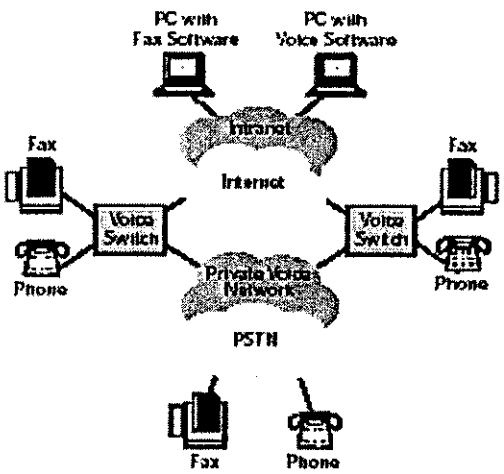

Figure 4.1 Infrastructure du système VoIP [66].

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 qui déterminent la qualité de service délivrée par le réseau [69] : on peut citer les codecs, la bande passante, le retard, la gigue et les pertes de paquets dans le réseau.

4.3 Les Facteurs Affectant la Qualité de Service

4.3.1 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 [48][67][68] ont été conçu pour travailler avec ces contraintes. Conçus pour différentes applications, ils représentent de bons candidats pour les transmissions VoIP. Le tableau 4.1 montre les performances pour différents codecs.

Tableau 4.1 Les principaux codecs en VoIP

Standards	Algorithme	Débit (kbits/s)	Taille trame(ms)/ lookahead	Complexité (MIPS)	Qualit é (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

4.3.2 Le Retard

Les retards rencontrés dans la transmission par paquets sont classés en plusieurs types : retard d'accumulation, retard paquetage, retard du réseau et retard de propagation.

4.3.3 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. 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. Cette opération rajouté un retard aux retards cités précédemment.

4.3.4 Les Pertes de Paquets

La transmission de la voix sur Internet (réseau IP) se fait par paquets. Au récepteur, certains paquets peuvent manquer, à cause des délais, à l'encombrement ou aux erreurs de transfert. Cette perte de paquets dégrade la qualité de la voix reçue dans un système de transmission VoIP. Etant donné que la transmission de la voix se fait en temps réel, le récepteur ne fait pas pas appel à la retransmission des paquets perdus à cause des délais de transferts trop importants. Dans le paragraphe suivant, on développera les techniques employées pour le masquage des pertes.

4.4 Les Techniques de Masquage des Paquets perdus

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. Dans ce paragraphe, nous exposons les différentes techniques utilisées pour récupérer les paquets perdus.

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 4.2 [70]- [75].

4.4.1 Masquage Basé sur l'Emetteur

Dans cette catégorie de masquage, nous présentons quelques techniques qui exigent la participation de l'émetteur pour masquer les trames perdues [76].

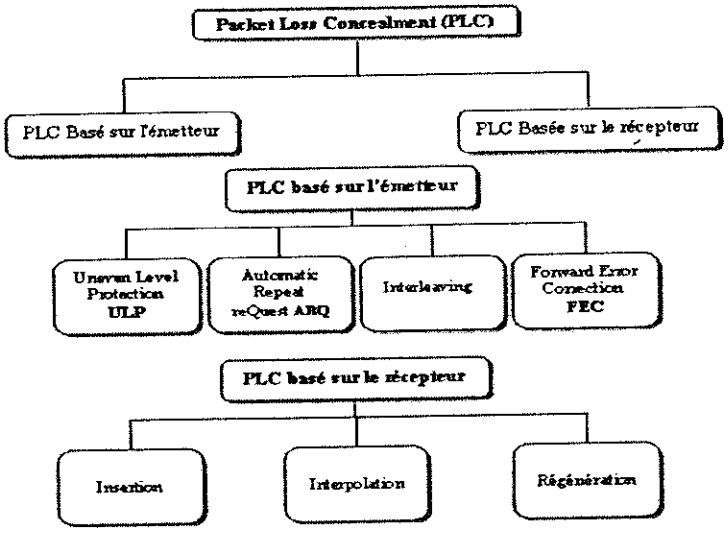

Figure 4.2 Les techniques de masquage des paquets perdus [76].

4.4.1.1 Correction d'Erreurs Progressive

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

Les données redondantes proviennent des données originales en utilisant l'opération logique (*XOR*) : un paquet de parité est généré pour les k paquets originaux de données.

La *FEC* transmet k paquets de données originaux (D) et h paquets supplémentaires redondants de parité (P). La figure 4.3 présente un exemple pour $k = 3$ et $h = 2$.

Le codeur *FEC* produit deux paquets redondants (P_1, P_2) des trois paquets de données. Si un paquet de données (exemple D_3) et un paquet de parité (exemple P_1) sont mal reçus, le récepteur peut récupérer le paquet perdu (D_3) en utilisant les bons paquets reçus D_1, D_2 et P_2 [76][77].

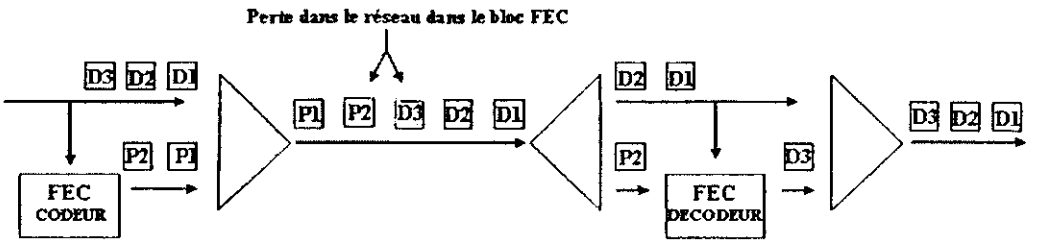

Figure 4.3 Exemple du *FEC* [76].

La *FEC* reste efficace pour un rapport (h/k) faible. Pour le décodeur *FEC*, les pertes de paquets consécutives peuvent être corrigées pour une grande valeur de k. Si k augmente, le délai de la reconstruction au niveau du récepteur augmente aussi. On peut citer plusieurs avantages et inconvénients du *FEC* :

- L'opération de la *FEC* ne dépend pas du contenu des données originales et la réparation est le remplacement exact du paquet perdu.
- Le paquet original de la donnée peut être utilisé par des récepteurs qui ne sont pas compatibles avec la *FEC* puisque les données redondantes sont envoyées habituellement comme un flux séparé.
- Les inconvénients de cette technique sont: le retard supplémentaire imposé, une bande passante croissante et la difficulté d'implémentation du décodeur.

4.4.1.2 L'Entrelacement

L'entrelacement (Interleaving) est une technique utile, pour réduire les effets de pertes, lorsque la taille des trames est plus petite que celle des paquets et que le retard de bout-à-bout (end-to-end) n'est pas important [77]. Dans cette technique, les trames originales de données ne sont pas combinées dans le même ordre séquentiel telles qu'elles sont produites par le codeur mais elles sont entrelacées par l'émetteur comme il est illustré sur la figure 4.4.

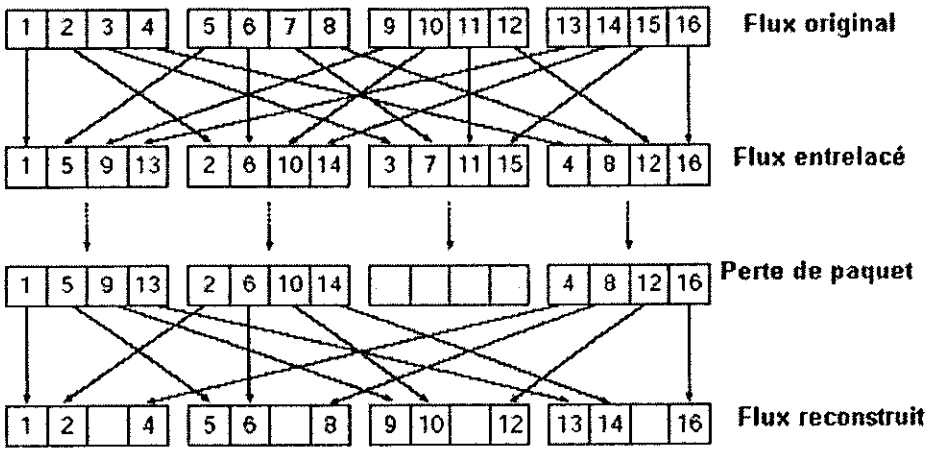

Figure 4.4 Exemple d'entrelacement [76].

Au récepteur, les trames de données sont rassemblées à leur ordre original. La figure 4.4 montre que l'effet d'une perte de paquet est réparti sur de petits intervalles correspondant aux trames de données distribuées au lieu d'être adjacentes. La réduction de l'effet des pertes est dûe aux raisons suivantes:

Les petits intervalles vides résultants correspondent aux intervalles de la parole qui sont considérablement plus courts qu'une longueur de phonème. Par conséquent, les êtres humains peuvent interpoler mentalement les intervalles vides et l'intelligibilité de la parole ne sera pas diminuée. Contrairement à la situation du non-entrelacement (no-interleaving), où une simple perte d'un paquet peut avoir comme conséquence la perte d'un phonème complet, ce qui diminue l'intelligibilité de la parole.

Si le récepteur emploie une certaine technique de masquage des erreurs (par exemple les lacunes dues à la perte de paquets sont remplies en utilisant l'interpolation des trames de données adjacentes reçues), un rendement plus élevé sera obtenu si l'interpolation est effectuée sur de petits intervalles au lieu de longs intervalles.

L'augmentation de latence constitue un inconvénient à l'utilisation de l'entrelacement dans des applications interactives. Alors que l'avantage majeur de cette technique est qu'elle ne nécessite pas une augmentation de la bande passante.

4.4.1.3 La Requête de Répétition Automatique

La requête de répétition automatique ARQ (*Automatic Repeat Request*) est une technique de retransmission, dont les stratégies de base sont :

- La détection du paquet perdu qui se fait par le récepteur ou par l'émetteur.
- La stratégie de reconnaissance : Le récepteur envoie des informations indiquant les données reçues et les données manquantes.
- La stratégie de rediffusion: elle determine les données retransmises par l'émetteur.

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

4.4.1.4 La Protection à Niveau Inégal

Lorsque les sous-divisions constituant la donnée n'ont pas la même importance (Les données de la parole, en particulier), une technique dite protection à niveau inégal *ULP* (*Uneven Level Protection*) peut être appliquée. Cette technique attribue plus de protection aux données les plus importantes. Elle est très souvent employée avec la technique *FEC*. Les unités de données sont arrangées dans un paquet de type *RTP* (*Real-time Transfert Protocol*) par ordre d'importance décroissant. Plus de protection est appliquée aux débuts des unités, c'est à dire aux données les plus importantes.

4.4.2 Masquage Basé sur le Récepteur

Dans ce paragraphe, nous résumons plusieurs techniques de masquage de pertes qui peuvent être effectuées au niveau du récepteur et qui n'exigent pas la contribution de l'émetteur [79]. Ces techniques sont en général moins efficaces que les techniques précédentes (basées sur l'émetteur). Elles consistent à produire des remplacements semblables aux paquets originaux perdus. Ces techniques fonctionnent bien pour des taux de pertes relativement faibles (< 15%) et pour des lacunes petites (< 40ms). Quand la longueur des pertes consécutives approche de

la longueur d'un phonème, ces techniques ne fonctionnent pas correctement. Il existe trois catégories de méthodes de dissimulation : insertion, interpolation et régénération.

4.4.2.1 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 parcequ'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 [77].

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

4.4.2.2 L'Interpolation

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

4.4.2.3 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.

4.5 Conclusion

Nous avons présenté un aperçu sur la transmission de la voix via les réseaux IP. Nous avons abordé un des problèmes affectant la 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. On verra une application de ces techniques au chapitre suivant.

Chapitre 5

Amélioration des performances des codeurs basés sur LPC dans les réseaux IP

5.1 Introduction

Dans ce chapitre nous présentons les résultats obtenus pour l'amélioration des performances des standards de l'ITU G.729 et G.723.1 [6][67] en termes de robustesse contre les pertes de paquets lors de la transmission de ces derniers à travers les réseaux IP. Dans un premier temps, nous donnerons une description de la méthode de masquage interpolative utilisée. Ensuite, nous présenterons pour le standard G.729, les résultats obtenus par l'implémentation de la méthode de masquage par interpolation et le codage intra-trame pour la quantification des paramètres LSPs. Nous exposerons une étude comparative des distorsions spectrales causées par le masquage prédictif et interpolatif des trames effacées. Nous donnerons également les résultats obtenus avec le standard G.723.1 par l'application des techniques intra-trame pour la quantification des paramètres LSPs et la méthode interpolative pour le masquage des trames perdues.

5.2 Masquage des Pertes dans le Standard ITU G.729

Le codeur de l'ITU G.729 possède une procédure de traitement des trames effacées basée sur une méthode de masquage prédictif. Ce type de méthodes n'introduit aucun délai supplémentaire car les paramètres des trames perdues seront récupérés à partir des bonnes

trames précédentes. Cependant, ce codeur quantifie les paramètres LSPs par une méthode prédictive, donc la procédure de masquage utilisée peut causer une propagation de l'erreur aux autres trames comme illustré par la figure 5.1, où les distorsions spectrales d'une séquence de parole codée avec et sans effacement de trame sont représentées. Sur le graphe, on peut très bien voir la propagation de l'erreur de la distorsion après chaque perte (voire l'indicateur des trames effacées et la divergence des deux graphes).

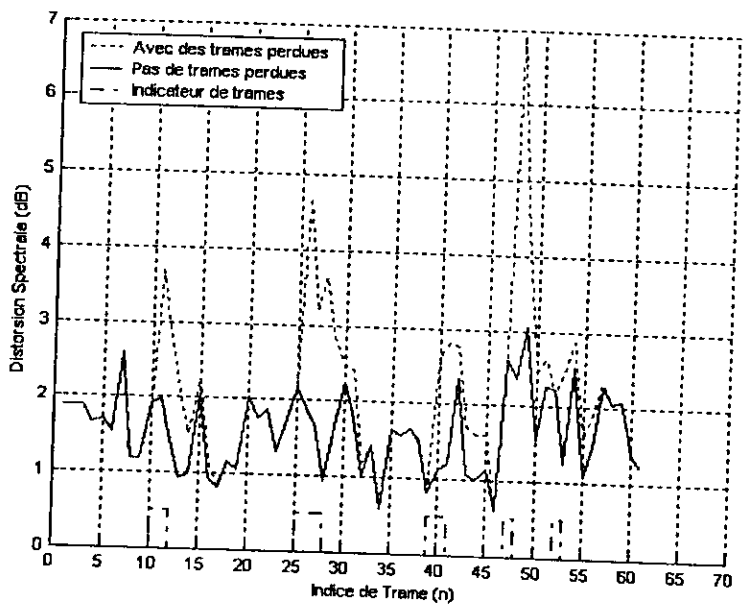

Figure 5.1 Propagation de l'erreur de la distorsion spectrale dans le G.729

5.3 Masquage par Interpolation

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

L'approche par interpolation, pour les codeurs CELP, a été à peine exploitée en raison du retard supplémentaire imposé par cette technique, qui n'est pas acceptable dans certaines applications, comme le cas de l'émission sans fil où le retard est fortement contrôlé.

L'apparition d'une nouvelle et importante application, la voix sur des réseaux IP (VoIP), a rendu la méthode par interpolation très attirante. Dans les systèmes VoIP, en fait, une ou plusieurs trames futures sont, au moins la plupart du temps, disponibles au décodeur, chargées dans un tampon appelé le "tampon du playout ". Un tel tampon est introduit pour minimiser les effets d'instabilité du retard et c'est un composant essentiel pour tous les récepteurs VoIP. Par conséquent, on peut exploiter le délai introduit par ce tampon pour masquer les trames effacées avec l'approche interpolative et donc améliorer les performances du codec sans aucun coût supplémentaire en terme de délai. La figure 5.2 illustre l'application du masquage par interpolation dans un récepteur VoIP.

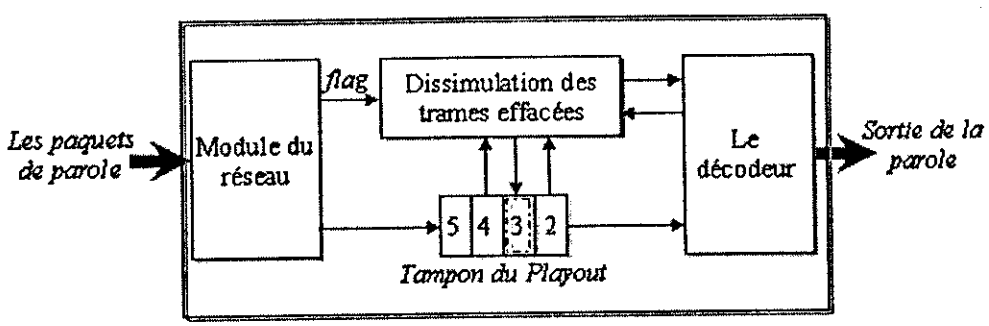

Figure 5.2 Récepteur VoIP typique : Masquage par Interpolation

Les paquets arrivant du réseau sont traités d'abord par le module du réseau. Les statistiques sont collectées, les paquets sont rangés et transférés au tampon du playout. Si après le temps du playback, le paquet n'est pas arrivé, il est déclaré perdu et le module de masquage des trames effacées le reconstruit en utilisant les bonnes trames futures et précédentes. Sur la figure 5.2, le paquet 3 manque, alors on le reconstruit en interpolant le précédent (2) et le suivant (4).

5.3.1 Application du Masquage par Interpolation au G.729

Les paramètres LSPs sont bien connus par leur propriété d'être ordonné strictement en ordre ascendant avec leurs indices pour chaque trame. Ils sont connus aussi par leurs corrélations

inter-trames et *intra-trames*. A cet effet, et pour appliquer le masquage des trames effacées par interpolation au standard G.729, nous allons chercher une quantification *intra-trame* qui donne des performances comparables ou meilleures que la quantification prédictive (*inter-trame*), utilisée par le standard G729.

5.3.1.1 Quantification Intra-trame des Paramètres LSPs

Nous avons choisit la technique SVQ, décrite dans le chapitre 3, pour la quantification des vecteurs de paramètres LSPs. Un avantage de la SVQ est la possibilité de modifier facilement les allocations des bits entre les différents sous vecteurs. Cette propriété est très utile si quelques paramètres des vecteurs d'entrée exigent une quantification plus exacte que les autres. Les questions principales pour concevoir une SVQ concernent :

-En combien de parties le vecteur devrait être divisé et combien de composants devrait en contenir chaque partie.

– L'allocation des bits attribuée à chaque partie. C'est une tâche compliquée du fait que les LSPs en fréquences intermédiaires varient plus que les LSPs en hautes et basses fréquences comme le montre la figure 5.3

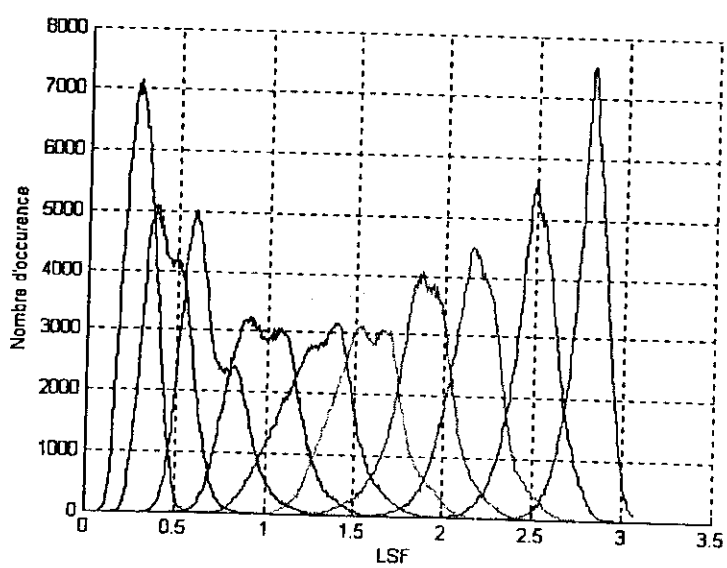

Figure 5.3 Distribution des paramètres LSPs

5.3.1.2 Bases de Données et les Mesures de Distorsions Utilisées

Le signal parole utilisée dans nos expériences consiste en deux bases de données distinctes qui incluent 229829 vecteurs LSPs pour l'apprentissage (*Training*) et 72839 vecteurs ont été réservés pour l'évaluation ou les tests.

Les vecteurs LSPs ont été produits à partir de la base de données TIMIT [78] qui contient un total de 6300 phrases, 10 phrases prononcées par chacun des 630 orateurs des 8 régions du dialecte des États-Unis. La fréquence d'échantillonnage du signal parole des fichiers est de 8 kHz. Les phrases sont prononcées par des hommes et des femmes.

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

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

Dans notre évaluation nous avons employé deux mesures objectives de la qualité: la distorsion spectrale (SD) (équation 2.50) et la distorsion EMBSD (*Enhanced Modified Bark Spectral Distortion*) [79].

La mesure objective "EMBSD" est connue d'avoir une corrélation très élevée avec les essais subjectifs (Tableau 5.1) et convient à l'évaluation de la parole dégradée par les erreurs de transmission dans des environnements réels de réseaux [79], telles que les erreurs de bits et d'effacements des trames.

Tableau 5.1 Table provisoire de conversion des valeurs du MOS au EMBSD

Catégorie	Qualité de la parole	EMBSD Perceptuelle Distorsion
1	Médiocre	8
2	Faible	6
3	Passable	4
4	Bonne	2
5	Excellente	0

5.4 Quantification des Paramètres LSPs

Pour trouver la partition optimale des vecteurs LSPs, une corrélation intra-trame a été calculée pour les 229829 vecteurs, c'est-à-dire la corrélation entre LSPi et LSPj de la même trame,(i, j = 1,2,...,10). Les coefficients de corrélation intra-trame sont présentés sur le tableau 5.2. Ces résultats montrent que la corrélation entre les LSPs consécutifs est considérable. La méthode de division la plus intéressante est (3,3,4). Cependant, le quatrième LSP est fortement corrèlé avec LSP3 qu'avec LSP5 et la corrélation entre LSP4 et LSP1 est approximativement la même que celle entre LSP4 et LSP6. Par conséquent, le quatrième LSP devrait être déplacé du second sous-vecteur au premier. De plus, à cause de la faible corrélation entre LSP8 et LSP9, le choix de la division (4,4,2) est théoriquement meilleur.

Tableau 5.2 La corrélation entre LSP_i et LSP_j de la même trame

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

Nous avons fait l'apprentissage des quantificateurs par l'algorithme GLA (generalized Lloyd Algorithm), pour différents débits binaires. Comme le standard G.729 utilise 18 bits/trame pour la quantification des LSPs, nous avons commencé avec 18 bits/trame en essayant de trouver une distorsion spectrale moyenne égale ou meilleure que celle du standard.

La distorsion spectrale moyenne pour le standard G.729, calculée pour les vecteurs de la base de donnée de test est : $SD_{G.729} = 1.543$ dB. Pour 18 bits/trame et 19 bits/trame, nous avons trouvé des distorsions spectrales >1.8 dB. Les résultas qui ont des SD proches de $SD_{G.729}$ sont mentionnés dans le tableau 5.3.

Tableau 5.3 Distorsions spectrales pour les différentes divisions et les différentes allocations de bits.

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

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

5.5 Interpolation des Paramètres LSPs

5.5.1 E spérance de l'Erreur Quadratique du Masquage par Interpolation

On calcule d'abord l'erreur quadratique pour une récupération par interpolation [80] des LSPs à partir des LSPs codés par une quantification intra-trame. Commençant à l'instant $n+1$, L trames consécutives sont perdues. La méthode d'interpolation récupère les vecteurs LSPs perdus par une interpolation linéaire entre les bonnes trames "antérieures" et "suivantes". Soit $F_n = (f_1, f_2, ..., f_p)$ le vecteur LSP de la nième trame de dimension p et $\hat{F}_n$ sa version quantifiée ou interpolée. Le vecteur LSP interpolé peut être écrit

$$\hat{F}_{n+x} = \frac{L+x-1}{L+1}\hat{F}_n + \frac{x}{L+1}\hat{F}_{n+L+1} \tag{5.1}$$

Les paramètres LSPs peuvent être considérés comme stationnaires au sens large. Alors, on peut représenter par approximation les vecteurs LSPs quantifiés par leur version non quantifiée et prendre l'espérance de la distorsion quadratique moyenne

$$D_L = \frac{1}{L}\sum_{x=1}^{L}\sum_{p=1}^{P}(f_{n+x,p} - \hat{f}_{n+x,p})^2 \tag{5.2}$$

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

$$ED_{\text{int}} = \frac{\Phi(0)}{L}\sum_{x=1}^{L}[1 + \frac{(L+1-x)^2 + x^2}{(L+1)^2} - \frac{2(L+1-x)}{L+1}\phi(x) - \frac{2x}{L+1}\phi(L+1-x) + \frac{2x(L+1-x)}{(L+1)^2}\phi(L+1)] \tag{5.3}$$

où $\Phi(.)$ et $\phi(.)$ sont respectivement, la somme des autocorrélations et la somme normalisée des autocorrélations des vecteurs LSPs et sont définis par

$$\Phi(\tau) = \sum_{p=1}^{P} E[f_{n,p} f_{n+\tau,p}]$$
$$\phi(\tau) = \frac{\sum_{p=1}^{P} E[f_{n,p} f_{n+\tau,p}]}{\sum_{p=1}^{P} E[f_{n,p}^2]} \tag{5.4}$$

5.5.2 L 'Espérance de l'erreur Quadratique du Masquage Prédictif

Pour le masquage prédictif, les LSPs perdus sont récupérés par un estimateur scalaire fixe [80], à partir des vecteurs, codés par une quantification inter-trame prédictive, reçus des "bonnes" trames antérieures par un prédicteur scalaire fixe β et le vecteur LSP masqué

$$\hat{F}_{n+x} = B^x \hat{F}_n \tag{5.5}$$

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

$$\hat{F}_{n+L+1} = \beta^{L+1}\hat{F}_n + e_{n+L+1} \tag{5.6}$$

L'erreur quadratique totale de ces $L+1$ trames sera la somme de la partie prédite et de celle propagée. Donc l'espérance de la distorsion de la partie prédit est

$$L \times ED_{L,pred} = \Phi(0)\sum_{x=1}^{L}[1+\beta^{2x}-2\beta^{x}\phi(x)] \tag{5.7}$$

Pour la partie propagée, cette espérance s'écrit

$$D_{prop} = \sum_{p=1}^{P}(f_{n+L+1,p} - \beta^{L+1}f_{n,p} - e_{n+L+1,p})^2 \tag{5.8}$$

On prend l'espérance des deux côtés de l'équation (5.8). Tous les termes avec $e_n + L + 1$ seront égaux à zéro puisque $e_n + L + 1$ est indépendant de f_n et que l'espérance de $e_n + L + 1$ est égale à zéro. En négligeant les termes $e_n^2 + L + 1$ petits, on obtient :

$$ED_{prop} = \Phi(0)[1+\beta^{2(L+1)} - 2\beta^{L+1}\phi(L+1)] \tag{5.9}$$

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

$$\begin{aligned} ED_{\Pr ed} &= \frac{1}{L+1}(L \times ED_{L,pred} + ED_{prop}) \\ &= \frac{\Phi(0)}{L+1}\sum_{x=1}^{L+1}[1+\beta^{2x}-2\beta^{x}\phi(x)] \end{aligned} \tag{5.10}$$

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

$$ED_{rep} = \frac{\Phi(0)}{L+1}\sum_{x=1}^{L+1}[2 - 2\phi(x)] \tag{5.11}$$

5.5.3 Comparaison des Deux Méthodes

Nous avons calculé la somme des autocorrélations des 229829 vecteurs LSPs (voir Tableau 5.4)

Tableau 5.4 Les sommes des autocorrélations normalisées des paramètres LSPs

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

Les rapports $\frac{ED_{\text{int}}}{ED_{pre}}$ pour L = 1, 2, 3 sont calculés et présentés dans le tableau 5.5.

Tableau 5.5 Rapport des distorsions moyennes de la récupération prédictive et par interpolation des LSPs

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

On remarque d'après les résultats présentés dans le tableau 5.5, que l'espérance de la distorsion moyenne de la méthode prédictive est plus grande que celle de la méthode par interpolation d'un facteur de 2. On peut représenter ou approcher la relation entre la distorsion spectrale (SD) et l'erreur quadratique (sqe) par une fonction linéaire (dans l'échelle log-log. C'est une relation statistique calculée pour un nombre de vecteurs LSPs très grand) c'est-à-dire : $\log(SD) = r\log(sqe) + b$ où r est positif et b une constante.

Alors on obtient :

$$\frac{DS_{pre}}{DS_{\text{int}}} = (\frac{ED_{pre}}{ED_{\text{int}}})^r > 1 \tag{5.12}$$

D'après ces résultats théoriques, on peut réellement estimer l'efficacité de la méthode par interpolation appliquée au G.729.

5.6 Simulations et Résultats

Nous avons trouvé (tableau 5.3) qu'avec une quantification SVQ *intra-trame* à 20 bits/trame, la distorsion est minimale pour le codeur G.729. Malgré que, cette quantification nécessite 2 bits/trame plus que le débit utilisé par codage prédictif du G.729 original, elle présente des propriétés adéquates à l'application de la méthode de récupération par interpolation des LSPs. Nous allons simuler la voix en temps-réel sur des réseaux par paquets où chaque paquet contient une seule trame.

5.6.1 L e 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 ([75], [81]). 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

$$\begin{aligned} P_1 &= P_0 \cdot p + P_1 \cdot q \\ P_1 + P_0 &= 1 \end{aligned} \tag{5.13}$$

$$\Rightarrow P_0 = \frac{1-q}{p+1-q} \qquad P_1 = \frac{p}{p+1-q} \tag{5.14}$$

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 5.4 présente le modèle de Markov avec ses probabilités de transition et le tableau 5.6 cite les taux de perte utilisés dans notre simulation.

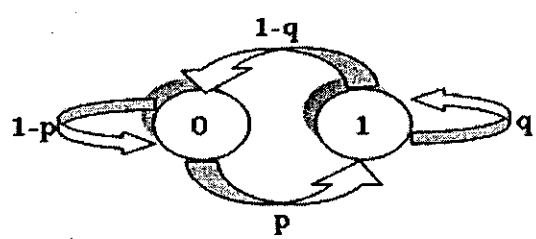

Figure 5.4 Pertes de paquets modélisées par un processus aléatoire de Markov

Tableau 5.6 Les taux de pertes simulés

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

5.6.2 Pro cédure de Masquage Implémentée

Le processus complet de masquage peut être résumé comme suit.

Si une trame est déclarée perdue :

1. Interpolation linéaire des paramètres LSPs de la bonne trame "précédente" et la bonne trame "suivante";
2. Interpolation du délai tonal ;
3. En se basant sur la bonne trame précédente, prendre une décision sur le type de la trame (voisée ou non voisée V/NV);
4. Si la trame précédente est voisée : -Mettre la contribution du dictionnaire fixe à zéro;
5. Si la trame précédente est non voisée: -Mettre l'information du dictionnaire adaptatif à zéro, -Utiliser l'information précédente du gain, -Remplacer les signaux d'excitation par une séquence de nombres aléatoires normalisée par le gain atténué.

5.6.3 Rés ultats de l'Implémentation de la Méthode par Interpolation au Standard G.729

Le tableau 5.7 regroupe les résultas de comparaison des performances de la méthode interpolative appliquée aux paramètres LSPs quantifiés avec SVQ *intra-trame* et la méthode prédictive adoptée par le G.729. La figure 5.5 montre les performances, en terme de distorsion spectrale moyenne, pour différents débits binaires [82].

Tableau 5.7 Distorsion spectrale moyenne et
les pourcentages [2-4]dB et >4dB avec des trames effacées

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

On voit bien qu'avec un extra débit de 0.2 kbits/s, c'est-à-dire 2,5% du débit total, notre méthode de quantification et de récupération des LSPs ainsi appliquée, réalise une diminution des distorsions spectrales de 0.3 à 1.1 dB, comparée à la méthode adoptée par le standard G.729. Les pourcentages des distorsions ([2-4]dB et >4dB) sont aussi beaucoup plus petits, ce qui permet d'avoir une qualité perceptuelle considérable lorsque l'effacement des trames se produit.

La distribution, des distorsions spectrales, représentée par la figure 5.6, montre que la plupart des trames perdues sont interpolées avec de faibles distorsions par la quantification intra-trame SVQ.

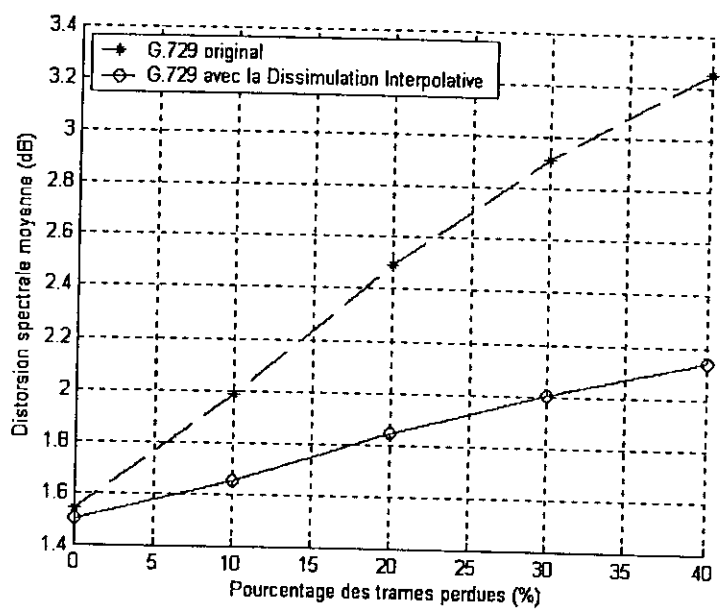

Figure 5.5 Distorsions spectrales moyennes avec des trames effacées

Nous avons appliqué la distorsion spectrale modifiée " EMBSD ", qui se rapproche des tests subjectifs [79] . La figure 5.7 montre les résultats de comparaison des performances obtenues par l'application d'un masquage par interpolation aux LSPs quantifiés avec une SVQ intra-trame, à la méthode prédictive adoptée par le G.729.

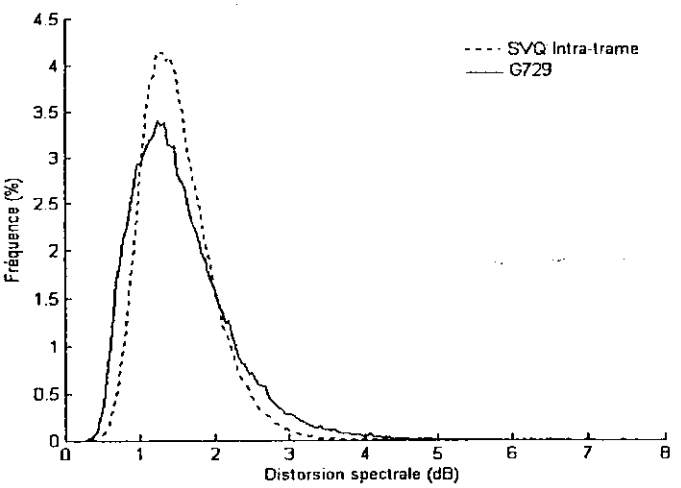

Figure 5.6 Histogrammes des Distorsions Spectrales (SD)

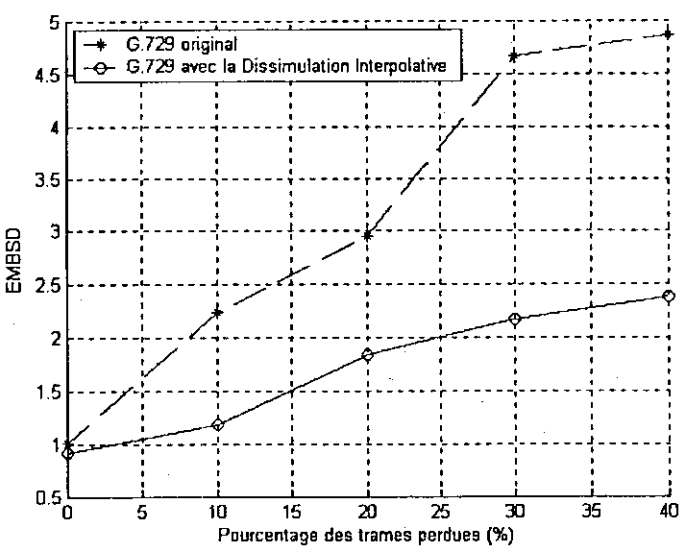

Figure 5.7 EMBSD avec des trames effacées pour G.729

On voit bien, qu'avec un débit de 0.2 kbits/s de plus, notre méthode de quantification et de récupération des LSPs ainsi appliquées, réalisent jusqu'à 2.5 de moins de la distorsion perceptuelle EMBSD, comparée à la méthode adoptée par le standard G.729. Et si on compare les résultats de la figure 5.7 avec ceux du tableau 5.1 on peut conclure que notre méthode

donne une bonne qualité jusqu'au 30% de taux de perte ce qui correspond à une qualité passable pour le G.729 original [81].

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

5.7 Application de la Quantification Intra-Trame au G723.1

Nous avons également appliqué la quantification intra-trame des paramètres LSPs, particulièrement le split 3-3-4 à 25 bits/trame avec le standard G.723.1, avec la méthode utilisée par ce dernier qui est le SVQ prédictif à 24 bits/trame. Les résultats obtenus [83], en termes de distorsions spectrales moyennes, sont regroupés sur le tableau 5.8.

Nous avons appliqué une interpolation entre la ''bonne'' trame précédente et suivante pour la récupération des trames perdues. La figure 5.8 montre les performances du SVQ comparé au PSVQ pour le G.723.1 lorsqu'il y a effacement pour différents débits de pertes. On peut noter à partir du tableau 5.8, que les distorsions obtenues par SVQ sont inférieures à celles obtenues par SVQ prédictif. Donc, en ajoutant uniquement un bit au second vecteur du SVQ on peut améliorer la distorsion spectrale moyenne lorsqu'il y a effacement de trames. On peut conclure que puisque les pertes sont inévitables, l'ajout d'un bit semble une solution intéressante.

Tableau 5.8 Distorsions spectrales moyennes et les
les pourcentages [2-4]dB et >4dB avec des trames effacées

Trames perdues (%)	PSVQ			(3-3-4) SVQ		
	SD_{moy} (dB)	2-4 dB (%)	>4 dB (%)	SD_{moy} (dB)	2-4 dB (%)	>4 dB (%)
0	1.56	19	0	1.55	14	0
10	2.03	24	14	1.91	19	7
20	2.54	32	16	2.30	24	12
30	2.99	36	26	2.69	27	18

Nous avons également comparé le SVQ (3-3-4) avec le quantificateur Two-stage VQ Lattice

VQ [84]. Nous avons trouvé que les performances obtenues avec le SVQ et le Two-stage VQ Lattice VQ sont comparables aux performances du prédictif split VQ employé par le standard sans extra bits. La figure 5.9 illustre les résultats de comparaison.

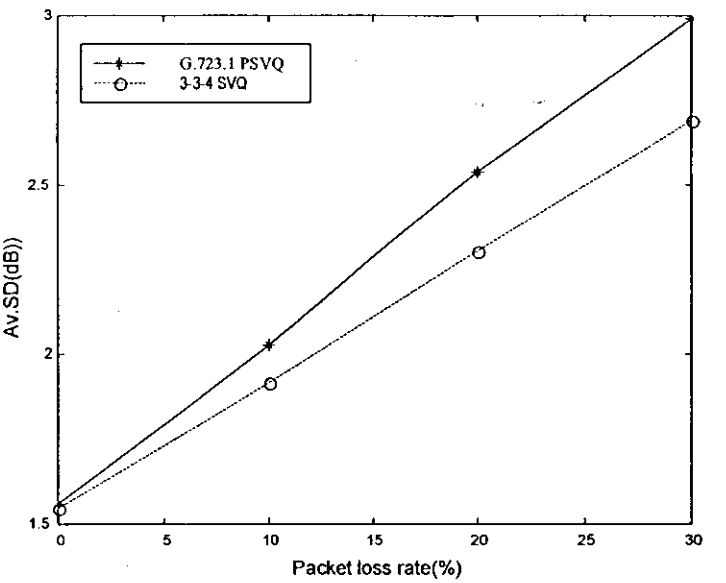

Figure 5.8 Les Distorsions spectrales moyennes avec des trames effacées de la SVQ et PSVQ pour G.723.1

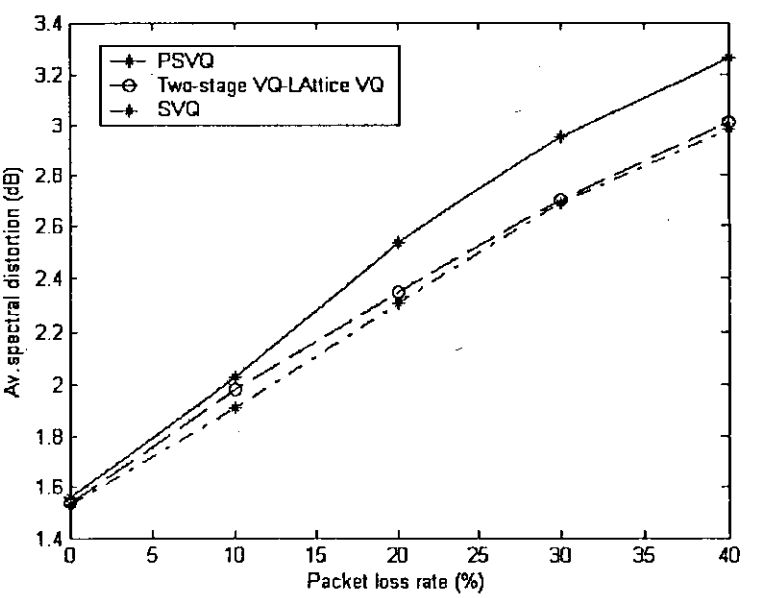

Figure 5.9 Les Distorsions spectrales moyennes avec des trames effacées de la SVQ, PSVQ et Two-stage VQ Lattice VQ pour G.723.1

5.8 Conclusion

Dans ce chapitre nous avons présenté les résultats obtenus par l'application pour la récupération des trames perdues d'une méthode de masquage par interpolation et une quantification intra-trame SVQ pour coder les LSPs pour le standard G.729. En ajoutant uniquement 2 bits/trame, les performances du G.729 sont améliorées.

L'inconvénient de cette méthode est le délai supplémentaire requis pour l'interpolation, mais comme nous l'avons vu, nous avons exploité le délai introduit par le " *tampon du play out* " (qui est un composant essentiel dans les applications VoIP), pour implémenter cette méthode.

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

En ce qui concerne le standard G.723.1, nous avons trouvé qu'en ajoutant un bit/trame, nous pouvons améliorer ses performances par une quantification intra-trame SVQ et une interpolation des t.0

rames perdues.

Conclusion générale

La tendance vers le codage des signaux à bas débit tout en gardant une haute qualité perceptuelle de la parole est une course qui ne s'arrêtera jamais. Les codeurs LP sont les plus utilisés parce qu'ils peuvent extraire des caractéristiques significatives à partir du signal parole et les transmettre à très bas débits. Chaque trame du signal parole est modélisée par un filtre tous-pôles par l'analyse par prédiction linéaire et les coefficients du filtre sont alors codés et transmis. Les coefficients LP sont transformés en paramètres LSPs, car ces derniers ont montré qu'ils sont mieux adaptés à la quantification vectorielle et assurent la stabilité du filtre LP.

L'ensemble des travaux de cette thèse concerne les techniques de codage de la parole et leurs applications dans un premier temps aux LSPs et, dans un deuxième, aux systèmes de communications VoIP.

Nous avons réussi à apporter des améliorations au niveau du codeur ainsi qu'au niveau des systèmes VoIP en proposant des algorithmes de codage capables de réduire la complexité, le débit et les pertes dues à la propagation des erreurs des systèmes actuels. Ce qui était le but recherché.

La première partie de nos travaux consistait à coder les paramètres LSPs par un nombre de bits aussi faible que possible. Donc, nous avons cherché à réduire la complexité des quantificateurs vectoriels tout en gardant une bonne qualité perceptuelle de la parole. Pour cela, nous avons présenté trois quantificateurs vectoriels basés sur le principe de la division en sous-vecteurs. Nous avons montré que la division 3-3-4 donnait de meilleures performances que les deux autres méthodes présentées. Nous avons montré également que moyennant 24 bits/trame, la structure 3-3-4, pouvait avoir des performances comparables à celle obtenues par un quantificateur scalaire à 34 bits/trame attribués aux LSPs par le codeur de la voix (standard U.S. federal) 4800 bits\s. Nous avons montré que grâce à ce découpage, nous pouvons avoir un quantificateur relativement performant et surtout simple à réaliser et que les dictionnaires des sous vecteurs ne dépassent pas 8 bits/vecteur.

Une des mesures de l'efficacité d'un quantificateur est la robustesse face au bruit. Nous avons pour cela présenté une méthode basée sur le tri des indices des dictionnaires.

Les recherches sur le codage des LSPs ont duré plus de vingt années et continuent toujours. Le seul souci était de voir le compromis entre débit et complexité pour une bonne qualité de la parole. Avec l'émergence de la transmission de la voix par paquets, la robustesse contre l'effacement des trames ou pertes devient un critère important dans l'évaluation des performances d'une technique de codage des LSPs.

Par nos travaux dans cette direction, nous avons amélioré les distorsions pour le standard G723.1 par l'utilisation d'une méthode intra-trame de quantification par split (3-3-4). Nous avons pu sauver 1 bit/trame par rapport à la méthode inter-trame utilisée par le codeur.

En ce qui concerne le standard G729 nous avons également implémenté une méthode de récupération des trames perdues. Cette méthode interpolative associée avec le SVQ pour le codage des LSPs a permis de sauver 2 bits/trame par rapport à la méthode inter-trame utilisée par le standard.

Les recherches futures dans le domaine du codage de la parole à bas débit se concentrent sur des méthodes de quantification des paramètres LSPs à faible complexité comme les méthodes algébriques ou hybrides.

Le codage conjoint de source et de canal (joint source channel coding) est une orientation des recherches pour la compression des signaux en présence de bruit.

En VoIP les efforts futurs se focaliseront sur l'implémentation sur des méthodes de récupération des trames, de l'excitation et la quantification des paramètres spectraux plus simples en termes de complexité.

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.1987.

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

[4]A. Gersho," Advances in speech and audio compression," *Proceedings of IEEE*, vol. 82, pp. 900-918, June 1994.

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

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

[7]W.B. Kleijn, K. K. Paliwal, eds., *Speech Coding and Synthesis.* Amsterdam: Elsevier, 1995.

[8]J. Makhoul, "Linear prediction: "A tutorial review speech," *Proc. IEEE.* Vol. 63, pp. 124-143, Apr. 1975.

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

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

[11] 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.

[12] 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.

[13]P. Kabal and R. P. Ramachandran, "The computation of Line Spectrum frequenciesusing Chebyshev polynomials,"in *Proc. IEEE Int Conf. Acoust. Speech, Signal Processing*, ASSP-34, pp. 1419-1426, Dec. 1986.

[14]F. Merazka, D. Berkani, "Extraction des paramètres LSF par prédiction Linéaire," COMAIE, Tlemcen, Algeria. pp. 127-130, Dec. 1996. ISSN 1111-357X. Editor OPU.

[15]A. H. et al, "Atriculation testing methods," *J. Acoust. Soc. Am.*, vol. 37, pp. 158-166, 1966.

[16] R. Kubichek, "Standards and Technology Issus in Objective Voice Quality Assessment," *Digital Signal Processing: A review Journal*, vol. DSP, pp. 38-44, Apr. 1991.

[17] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE J. Selected Area in Comm.*, vol. 10, pp. 819-829, Jun. 1992.

[18] A. H Gray, Jr, J. D. Markel, "Distance measures for speech processing," *IEEE Trans. Acoust. , Speech and Signal Processin.*, vol. ASSP-24, pp. 380-391, Oct. 1976.

[19]F. Soong, M. M. Sondhi, "A frequency-weighted Itakura spectral distorsion measure and its application to speech recognition in noise," in *Proc. IEEE Int Conf. Acoust. Speech, Signal Processing*, (Dallas, TX, April 1987), pp. II-625-II-628.

[20]K. K. Paliwal and B.Atal, "Efficient Vector Quantization of LPC Parameters at 24 bits/frame," *IEEE Trans. On Speech and Audio Processing*, vol. 1, No 1, pp.3-14, Jan. 1993.

[21]R. Laroia, N. Phamdo, and N. Farvardin. "Robust and efficient quantization of speech LSP parameters using structured vector quantizers, " in *Proc. IEEE Int Conf. Acoust. Speech, Signal Processing*, pp. 641- 644. Toronto, Ont. Canada, 1991.

[22]F. Merazka, D. Berkani, "All-Pole Modeling of Speech Based on a Geostatistical Model". Under review, Journal of speech Technology, Kluwer Academic Publisher.

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

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

[25]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.

[26]J. Makhoul, S. Roucous, and H. Gish "Vector Quantization in speech coding," *Proc. IEEE.* , vol. 73, pp. 1551-1558, Nov. 1985.

[27]T. Lookabaugh and R. M. Gray, "High-resolution quantization theory and the vector quantizer advantage," *IEEE Trans. Inform. Theory*, vol. IT-35, pp. 1020-1033, Sept. 1989.

[28]R. M. Gray, *Source Coding Theory*, Boston: Kluwer Academic Press, 1990.

[29]B. H. Juang, D. R Gray, A. H. Gray, Jr, "Distortion performance of vector quantization for LPC voice coding," *IEEE Trans. Acoust. , Speech and Signal Processin.*, vol. ASSP-30, pp. 294-303, Oct. 1982.

[30]R. M. Gray, D. L. Neuhoff "Quantization", *IEEE Trans. on information theory*, vol. 44, No 6, Oct. 1998.

[31]B.-H. Juang and A. H. Gray, Jr., "Multiple stage vector quantization for speech coding," in *Proc. IEEE Int Conf. Acoust. Speech, Signal Processing*, (Paris, France, April 1982), vol. 1, pp. 597-600.

[32]F. Merazka, D. Berkani, "Multistage Vector Quantization of LSP Parameters at Low Bit Rates," COMAIE'98, Bejaia, Algeria. Pp 417-420, Dec. 1998. ISSN 1111-357X. Editor OPU.

[33]F. Merazka and D. Berkani, *Very Low Bit-Rate Vector Quantization of LSP Parameters*, Computational Intelligence for Modeling, Control & Automation, Neural Networks & Advanced, Control Strategies,.*IOS Press in Holland. (Ohmsha)*, Vol. 54, pp. 374-379. Feb. 1999

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

[35]F. Merazka, D. Berkani, " Efficient Vector Quantization of LSP Parameters," CESA'98, IMACS / IEEE, Proceedings of the SMC Multi-conference, pp. 882-885. Tunisia. April 1998. ISBN 2-95123O9-0-7.

[36]F. Merazka, D. Berkani, " Spectral Coding by Fast Vector Quantization accepted to ICCTA'2000. The 10th International Conference on Computer Systems & Applications, Alexandria. Egypt, Sept.2000.

[37]F. Merazka, D. Berkani, " Classified Coding of Spectral Parameters," accepted to ICCTA'2000. The 10th International Conference on Computer Systems & Applications, Alexandria. Egypt, Sept.2000.

[38] F. Merazka, D. Berkani, Tree Structured Vector Quantization of LSP parameters at Low Bit Rate ," International Conference on Contribution of Cognition to Modeling. Accepted MS2000.

[39]F. Merazka, D. Berkani, " Robust Spectral Parameter Coding in Speech Processing, "Radio Africa, Botswana, Gaborone, 35-29 October 1999.

[40]F. Merazka, D. Berkani, " Intraframe and Interframe Coding of Spectral Speech Parameters," ICCTA'99. The 9th International Conference on Computer Systems & Applications, Alexandria. Egypt, August 1999.

[41]F. Merazka, " Vector Quantization of LSP Parameters," (Case Study). ICTP-URSI/BDT Workshop on the use of Radio for Digital Communications in Developing Countries. Trieste, Italy. Feb 1-19, 1999.

[42]F. Merazka, D. Berkani, "LSP Vector Quantization Algorithms Comparison" IASTED/ IEEE International Conference On Signal and Image Processing (SIP'98), Las Vegas, Nevada, USA. Oct.1998. Editor IASTED.

[43]F. Merazka, D. Berkani, "Low Complexity LSP Vector Quantization Using Splitting," ICCTA'98. The 8th International Conference on Computer Systems & Applications, session VII pp.6-9. Alexandria. Egypt, Sept.1998.

[44]F. Merazka, D. Berkani, "Low Bit Rates Vector Quantization of LSP Parameters by Splitting", (ICSPAT'98), Proceedings of the 9th International Conference on Signal Processing Applications and Technology, pp.1247-1251. Toronto, Canada. Sept. 1998. Editor & Distr. Miller Freeman, Inc.

[45]F. Merazka, D. Berkani, " Vector Quantization of LSP Parameters at Low Bit Rates," International Conference on Contribution of Cognition to Modeling. CCM'98, AMSE, France, July 1998.

[46] F. Merazka, D. Berkani, " LSP Vector Quantization," IASTED / IAC / IEEE Proceedings of the International Conference on Signal Processing and Communications, ICSPC'98, Canary Islands, Spain, February 1998. Editor IASTED.

[47]F. Merazka, D. Berkani, " A Robust Single Stage Vector Quantization for Spectral Coding," Proceedings of the Second International Conference on Electronic, Signals Systems & Control, SSC2'99, Blida, Algeria, May 10-12 1999.

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

[49]J. Skoglund and J. Lindén. "Predictive VQ for noisy channel spectrum coding: AR or MA ?, " *IEEE In proc. of ICASSP,* vol. 2 pp. 1351-1354, Munich, Germany. 1997.

[50]H. L. Vu and L. Lois "Optimal transform of LSP parameters using neural networks," in *Proc. IEEE Int Conf. Acoust. Speech, Signal Processing*, vol. 2 pp. 1339-1342, Munich, Germany, 1997.

[51]S. Sridharan and Leis. Two novel lossless algorithms to exploit index redundancy in VQ speech compression ," in *Proc. IEEE Int Conf. Acoust. Speech, Signal Processing*, Seattle, USA, May 1998.

[52]M. Skoglund and J. Skoglund," On non linear utilization of exploit index redundancy in VQ speech compression," in *Proc . IEEE Int Conf. Acoust. Speech, Signal Processing*, Seattle, USA, May 1998.

[53]C. S. Xydeas and T. M. Chapman." Multi codebook vector quantization of LPC parameters," in *Proc. IEEE Int Conf. Acoust. Speech, Signal Processing*, Seattle, USA, May 1998.

[54]M. Xie and J. P. Adoul "Fast and low complexity LSP quantization using Algebraic vector quantizer," in *Proc. IEEE Int Conf. Acoust. Speech, Signal Processing*, vol. 1, pp. 716-719, May 1995.

[55]J. Pan "Two-stage vector quantization pyramidal lattice vector quantization and application to speech LSP coding, " in *Proc. IEEE Int Conf. Acoust. Speech, Signal Processing*, vol. 2, pp. 737-740, 1996.

[56]U. Sinervo, J. Nurminen, A. Heikkinen & J. Saarinen, "Evaluation of split and multistage techniques in LSF quantization", in *Proc. Norsig* 2001, Trondheim, Norway, , pp. 18-22, Oct. 2001.

[57]J. Linden "Interframe quantization of spectrum parameters in speech coding" Lincent. Thesis Tech. Rep. 235L Chalmers Univ. Tech., 1996.

[58]K. K. Paliwal and W. B. Kleijn, *Quantization of LPC parameters*, in Speech coding and Synthesis, W. B. Kleijn and K. K. Paliwal, EDS. New York: Elsevier, 1995, pp. 433-466.

[59]F. Merazka, D. Berkani, "Robust Split Vector Quantization of LSP Parameters at Low Bit Rates," AJSE journal, vol. 29 No 1B, pp.31-48, April. 2004.

[60]Y. Tohkura, F. Itakura, and S. Hashimoto, "Predictive coding of speech signals and subjective error criteria," *in Proc.*IEEE *Int Conf. Acoust. Speech, Signal Processing*,vol. ASSP-26, pp. 587-596, Dec. 1978.

[61]Y. Kitawaki, K. Itoh. and K. Kakehi, "Speech quality measurement methods for synthesized speech," *Review of ECL.,*. Vol. 29 no. 9-1. NTT Japan Sept-Dec. 1981.

[62]Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun., Technol.*, vol. COM-28, No. 1, pp. 84-95, Jan. 1980.

[63]T. Moriya and M. Honda, "Speech coder using phase equalization and vector quantization," *in Proc.*IEEE *Int Conf. Acoust. Speech, Signal Processing*, Tokyo, Japan, pp. 1701-1704, Apr. 1986.

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

[65] G. Held, voice over data Networks, New York: McGraw-Hill, 1998.

[66]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.

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

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

[69]G. Held, voice over data Networks, New York: McGraw-Hill, 1998.

[70]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.* Vol. 29 NO. 2, Feb. 1981.

[71]J-C. Bolot, "Analysis and control of audio packet loss in the Internet," *NOSSDAV 95.*

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

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

[74]D. Goodman and G. Lockhart etc, "Waveform substitution techniques for recovering missing speech segments in packet voice communications," IEEE Trans. On Acoustics, Speech, and Signal Processing, vol. ASSP-34, No. 6, pp. 1440, Dec. 1986.

[75]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.

[76]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.

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

[78]NIST,"Timit Speech Corpus", NIST 1990.

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

[80]J.Wang and J.D.Gibson, "Parameter interpolation to Enhance the Frame Erasure Robustness of CELP Coders in Packet Networks," in Proc. IEEE *Int Conf. Acoust. Speech, Signal Processing*, pp.7-11, Salt Lake city, Ut., May 1991.

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

[82]F. Merazka & D. Berkani " under review", Wireless and personal communication Journal , Kluwer academic publisher.

[83]F. Merazka, D. Berkani, " Split VQ and Predictive split VQ of LSP Parameters in Packet Networks", 10th IEEE Digital Signal Processing Workshop, Georgia, USA 13-16 oct. 2002.

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