Codage de la parole à bas et très bas débits

Geneviève BAUDOIN*,
Jan CERNOCKY**,
Philippe GOURNAY***,
Gérard CHOLLET***

Résumé

Cet article présente les principales techniques de codage de parole à bas et très bas débit, de 50 bit/s à 4 000 bit/s. Puis il présente en détail la méthode HSX pour le codage à 1 200 bit/s et une nouvelle approche segmentale utilisant des unités acoustiques obtenues de manière non supervisée pour des débits inférieurs à 400 bit/s.

Mots clés :

SPEECH CODING AT LOW AND VERY LOW BIT RATES

Abstract

This paper reviews the main algorithms for speech coding at low and very low bit rates, from 50 bps to 4 000 bps. Then the HSX technique for coding at 1 200 bps and a new segmental method with automatically derived units for very low bit rate coding are presented in details.

Keywords :

Sommaire

I. INTRODUCTION

Dans les systèmes de téléphonie filaire classiques, la parole est numérisée à 64 kbit/s. De nombreux algorithmes [59,26] ont été proposés pour diminuer ce débit tout en essayant de conserver une qualité subjective donnée fonction des exigences de l'application à laquelle le codeur est destiné. On distingue en général 3 plages de débits :

- Les hauts débits, supérieurs à 16 kbit/s, correspondant à des algorithmes de codage de la forme d'onde non spécifiques à la parole,
- Les débits moyens, de 4 kbit/s à 16 kbit/s, correspondant à des techniques de codage hybrides utilisant des méthodes de codage de la forme d'onde et prenant en compte certaines propriétés de la parole ou de la perception auditive[1]. Le principal représentant de cette classe est le codage CELP [55].
- Les bas et très bas débits, de quelques dizaines de bits par seconde à 4 kbit/s, correspondant aux vocodeurs (VOice CODER) spécifiques au codage de la parole.

Un système de codage de la parole comprend 2 parties : le codeur et le décodeur. Le codeur analyse le signal pour en extraire un nombre réduit de paramètres pertinents qui sont représentés par un nombre restreint de bits pour archivage ou transmission. Le décodeur utilise ces paramètres pour reconstruire un signal de parole synthétique.

La plupart des algorithmes de codage mettent à profit un modèle linéaire simple de production de la parole. Ce modèle sépare la source d'excitation, qui peut être quasi périodique pour les sons voisés ou de type bruit pour les sons fricatifs ou plosifs, du canal vocal qui est considéré comme un résonateur acoustique. La forme du conduit vocal détermine ses fréquences de résonance et l'enveloppe spectrale (formants) du signal de parole.

Le signal de parole est souvent modélisé (modèle « source-filtre ») comme la sortie d'un filtre tout pôle

* Département Signaux et Télécommunications, ESIEE, BP 99 93162 Noisy Le Grand cedex. Email : baudoing@esiee.fr
** Université Technique de BRNO, Institut de Radioélectronique, BRNO, République Tchèque. Email : cernocky@urel.fee.vutbr.cz
***Thomson-CSF Communications, 66 rue du fossé blanc, 92231 Gennevilliers cedex. Email : philippe.gournay@tcc.thomson-csf.com
****CNRS-URA-820, ENST-TSI, 46 rue Barrault, 75634 PARIS cedex. Email : 13 chollet@tsi.enst.fr
1. Certains codeurs à haut débit utilisent aussi les propriétés de la perception auditive.

(appelé filtre de synthèse) dont la fonction de transfert représente l'enveloppe spectrale, excitée par une entrée dont les caractéristiques (en particulier la fréquence fondamentale[2]) déterminent la structure fine du spectre.

Le signal de parole n'étant pas stationnaire, les codeurs le découpent généralement en trames quasi-stationnaires de durée comprise entre 5 et 30 ms. Sur chaque trame, le codeur extrait des paramètres représentant l'enveloppe spectrale et caractérise ou modélise l'excitation de manière plus ou moins fine soit par quantification vectorielle, soit à l'aide de paramètres tels que l'énergie, le voisement et la fréquence fondamentale F_0. D'autres paramètres peuvent être calculés pour représenter plus finement l'excitation. Les paramètres les plus souvent utilisés pour l'enveloppe spectrale sont les paires de raies spectrales ou LSF (« *Line Spectral Frequencies* ») qui sont déduites des coefficients de prédiction linéaire et qui possèdent de bonnes propriétés pour la quantification et l'interpolation.

De nombreux algorithmes de codage à moyen débit ont été normalisés au cours des 10 dernières années pour les systèmes de communications avec les mobiles, GSM plein débit (ou *Full Rate* GSM) et demi-débit (ou *Half Rate* GSM), GSM plein débit amélioré (ou *Enhanced Full Rate* GSM), IS95 par exemple. La numérisation de la parole permet une meilleure protection contre les distorsions et les bruits introduits par les canaux radiomobiles. Une diminution du débit en dessous de 4 kbit/s, à condition de conserver une qualité de type téléphonique permettra d'augmenter la capacité des réseaux de communications avec les mobiles.

Les autres applications des codeurs à bas ou très bas débits incluent l'amélioration des systèmes de téléphonie sécurisés par cryptage, la radiomessagerie vocale, la téléphonie sur Internet, les répondeurs vocaux, les communications sur le canal HF, les communications personnelles par satellites à faible coût, et les bas débits des communications à débit adaptatif où le codeur de source et le codeur de canal s'adaptent à la qualité du canal et à la nature du signal.

L'évaluation des codeurs à bas et très bas débits ne peut pas se faire par des critères objectifs de rapport signal à bruit. Le signal décodé doit être perçu comme proche de l'original, mais les formes d'onde peuvent être très différentes. On évalue ces codeurs par des tests subjectifs, tels que le test ACR (Absolute Category Rating) délivrant un score MOS (*Mean Opinion Score*) ou le test d'acceptabilité DAM (*Diagnostic Acceptability Measure*) pour la qualité, et le test de rimes DRT (Diagnostic Rhyme Test []) pour l'intelligibilité. Ces tests sont menés sous certaines conditions de bruit ambiant ou de taux d'erreurs canal. Pour qualifier la qualité d'un codeur, on utilise les termes anglais : « *broadcast* », « *toll* », « *telecommunication* », « *synthetic* ». Une qualité de type « *broadcast* » correspond à un codage large bande (audioconférence par exemple), la qualité de type « *toll* » est celle du téléphone analogique filaire. Pour une qualité de type « *telecommunication* », l'intelligibilité et le naturel sont conservés mais quelques distorsions sont audibles. Un codeur de qualité « *synthetic* » est intelligible mais le signal manque de naturel.

La limite théorique minimum de débit pour un codage conservant l'information sémantique contenue dans la parole est d'environ 60 bit/s, si l'on compte environ 60 phones dans une langue et une vitesse d'élocution moyenne d'une dizaine de phones par seconde. Pour un débit aussi faible, les informations concernant le locuteur et ses émotions sont perdues.

Cet article s'intéresse à la catégorie des codeurs à bas et très bas débits. Il comprend une introduction (section 1) puis 3 parties principales (sections 2, 3, 4). La section n° 2 effectue un état de l'art des principes de codage à bas et très bas débit. Les sections 3 et 4 exposent comment nos travaux y trouvent leur place. La section n° 3 présente le codeur à 1 200/600 bit/s de type HSX (Harmonic Stochastic Coder) développé par Thomson-CSF Télécommunications. La section n° 4 décrit une nouvelle approche segmentale pour le codage à très bas débit par indexation d'unités acoustiques de taille variable obtenues automatiquement sur les données, approche développée dans le cadre de la thèse de Jan Cernocky.

II. LES CODEURS DE PAROLE À BAS ET TRÈS BAS DÉBIT

Pour les bas débits, typiquement de 800 bit/s à 4 000 bit/s, les techniques de codage de la forme d'onde ne donnent pas de bons résultats. Les codeurs doivent éliminer les informations sans pertinence pour la perception. Les vocodeurs utilisent certaines caractéristiques de la perception et de la production de la parole, aussi sont-ils généralement très peu efficaces pour les signaux autres que la parole comme les signaux DTMF[3] de numérotation téléphonique ou le bruit ambiant.

II.1. Présentation du codage CELP et de ses limitations pour le codage à bas débit

Le codage CELP (Code Excited Linear Prediction) a été introduit par Schroeder et Atal [55]. Il est très efficace pour les débits moyens de 4,8 kbit/s à 16 kbit/s, comme en témoignent les nombreuses normes qui l'utilisent. La figure 1 représente le principe du codage CELP.

Dans chaque trame, une analyse spectrale par prédiction linéaire court terme permet d'estimer l'enveloppe spectrale et détermine le filtre de synthèse *1/A(z)*.

On découpe chaque trame en sous-trames plus courtes (durée typique 5 ms). On modélise la périodicité

2. On utilise (par abus de langage) les expressions fréquence fondamentale et pitch indifféremment dans cet article.
3. DTMF = Dual Tone Multi-Frequency.

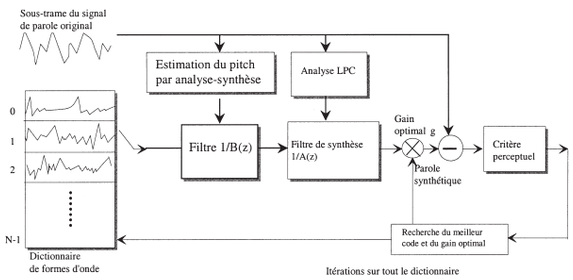

FIG. 1. — Principe du codage CELP.

Légende anglaise

de l'erreur de prédiction court terme (résiduel) à l'aide d'un prédicteur linéaire long terme représenté par un filtre $B(z) = 1 - bz^{-Q}$, où Q est une estimation de la période fondamentale. Sur chaque sous-trame on effectue une quantification vectorielle du signal par une technique d'analyse par synthèse. La quantification vectorielle utilise un dictionnaire de $M = 2^k$ séquences de bruit blanc normalisées en énergie. La longueur de ces séquences est égale à une sous-trame. Chaque séquence du dictionnaire est filtrée par le filtre de synthèse $1/(A(z)B(z))$ et multipliée par un gain. La sortie obtenue est le signal de parole synthétique qui est comparé au signal original. Le codeur teste toutes les séquences du dictionnaire, calcule le gain optimum pour chacune et retient celle qui minimise un critère « perceptuel[4] » de comparaison entre le signal synthétique et le signal original. Le codeur transmet l'indice de la séquence qui minimise le critère (sur k bits) ainsi que le gain associé, les paramètres spectraux et le pitch. Le critère « perceptuel » est un critère de moindres carrés calculé sur la différence entre le signal original et le signal synthétique après filtrage de cette différence par un filtre de pondération de type $A(z)/A(z/\gamma)$ où γ est compris entre 0 et 1 (typiquement $\gamma = 0.85$). Ce filtre pondère l'erreur dans le domaine fréquentiel, il atténue l'erreur dans les zones où l'amplitude de $1/|A(f)|$ est importante (zones de formants) et amplifie l'erreur dans les zones de faible amplitude de $1/|A(f)|$. Il met ainsi à profit les propriétés de masquage des bruits par les zones de fortes amplitudes du spectre, d'où le nom de critère perceptuel.

En pratique, pour diminuer la complexité du codeur, on remplace le filtre $1/B(z)$ par un dictionnaire qui contient les séquences de résiduel précédentes. Ce dictionnaire est appelé adaptatif, sa sortie est ajoutée à la sortie du dictionnaire de bruit blanc qui est appelé dictionnaire stochastique. Certains codeurs utilisent plusieurs dictionnaires stochastiques et forment le signal synthétiques en ajoutant les sorties des différents dictionnaires.

Quelques tentatives ont été faites pour diminuer les débits obtenus avec les codeurs CELP [28]. Mais en dessous de 3 kbit/s la méthode est inférieure aux approches de type vocodeurs.

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

Pour les sons voisés, le signal synthétique présente parfois des harmoniques de F_0 jusqu'à $f_e/2$ même si le signal original n'a plus d'harmoniques au-delà d'une fréquence f_{max}. On parle dans ce cas d'artéfact tonal. La figure 2 illustre ce phénomène pour un signal codé par un codeur CELP GSM demi-débit à 5 600 bit/s.

D'une manière générale la partie hautes fréquences du spectre est mal représentée car malgré le filtre de pondération, son amplitude est très faible par rapport à la partie basses fréquences qui est de ce fait favorisée par le critère des moindres carrés.

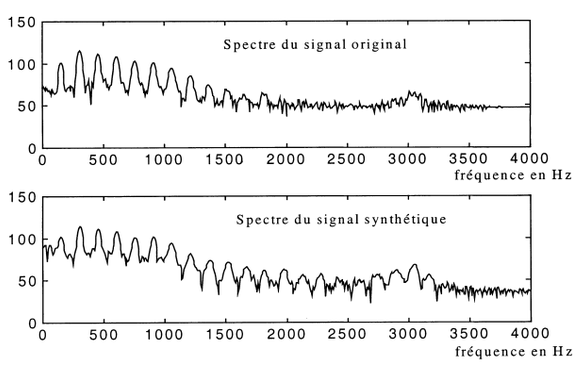

FIG. 2. — Artefacts tonals introduits par un codage *CELP* à 5 600 bit/s.

Légende anglaise

II.2. Les vocodeurs classiques à 2 états d'excitation

Dans les vocodeurs classiques, vocodeurs à canaux, vocodeurs à formants, ou vocodeurs LPC, les différentes trames de signal sont classées en trames voisées (V) et trames non-voisées (NV).

Ces vocodeurs classiques utilisent le modèle « source-filtre ». La synthèse du signal décodé utilise un signal d'excitation reconstruit formé d'un bruit blanc pour les trames non-voisées et d'un train périodique d'impulsions à la fréquence F_0 pour les trames voisées. La figure 3 représente le synthétiseur d'un vocodeur à 2 états d'excitation.

Ces vocodeurs diffèrent essentiellement dans leur façon d'estimer et d'appliquer l'enveloppe spectrale.

Dans les vocodeurs à canaux introduits par Dudley en 1939 [17], le codeur évalue l'énergie, le voisement, F_0, et les puissances relatives du signal dans un ensemble de bandes de fréquences adjacentes (de l'ordre de 10 bandes). Le décodeur génère la parole synthétique en

4. On utilise le néologisme « perceptuel pour indiquer un critère ou un filtre essayant de tenir compte de la perception auditive.

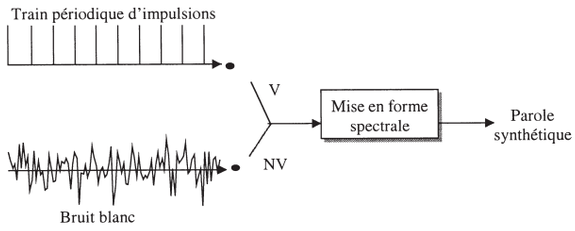

FIG. 3. — Synthèse dans un vocodeur à 2 états d'excitation.

Légende anglaise

passant le signal d'excitation dans un banc de filtres passe-bande dont les sorties sont pondérées par les puissances relatives du signal original dans ces différentes bandes. Les sorties des filtres sont ensuite ajoutées et cette somme est mise à l'échelle en fonction de l'énergie de la trame originale. Ces codeurs ont été utilisés jusqu'à des débits de 400 bit/s.

Dans les vocodeurs à formants [17], le codeur détermine la position, l'amplitude et la largeur de bande des 3 premiers formants, ainsi que l'énergie de la trame, le voisement et F_0. Au décodeur, l'excitation synthétique est filtrée par 3 filtres accordés sur les formants. Le signal résultant est mis à l'échelle en fonction de l'énergie de la trame. On obtient avec cette technique un signal intelligible pour des débits de 1 200 bit/s, mais la détermination des formants est une tâche difficile et peu robuste.

Dans les vocodeurs à prédiction linéaire LPC (Linear Predictive Coding) [1,64], l'enveloppe spectrale du signal de parole est modélisée par l'amplitude de la fonction de transfert d'un filtre tout pôle *1/A(z)*. Les coefficients a_i du filtre sont obtenus par prédiction linéaire. Le signal de parole x_n est prédit par $\hat{x}_n$ qui est une combinaison linéaire des échantillons précédents $\hat{x}_n = -\sum_{i=1}^{p} a_i x_{n-1}$. L'enveloppe spectrale est très sensible la quantification des coefficients ai. De plus l'interpolation de ces coefficients peut conduire à des filtres de synthèse instables. Aussi les transforme-t-on souvent en un autre jeu de coefficients pour la quantification et la transmission. Les coefficients classiques sont les logarithmes de rapports d'aires (Log Area Ratio ou LAR), les coefficients de réflexion (ou k_i), et les paires de raies spectrales (Line Spectrum Frequencies ou LSF). Le nombre de coefficients a_i est compris entre 8 et 16 pour une fréquence de 8 kHz, de façon à ce que la fonction de transfert du filtre présente un nombre suffisant de résonances pour modéliser correctement les 3 à 5 premiers formants. En plus des coefficients déduits des coefficients LPC, le codeur transmet l'énergie, le voisement et la fréquence fondamentale de la trame. Le décodeur génère le signal synthétique en filtrant l'excitation reconstruite par le filtre de synthèse *1/A(z)* et en mettant à l'échelle la sortie en fonction de l'énergie de la trame.

Les codeurs LPC à 2 états ont été développés pour des débits d'environ 2 400 bit/s. Des débits de 600 à 800 bit/s ont été atteints en appliquant une quantification vectorielle aux coefficients spectraux [48, 30, 16, 29].

Le codage LPC à 2 400 bit/s a été normalisé par l'OTAN (Voice coding standard STANAG 4198 [45]), le département de la défense américain DOD (Federal Standard 1015 [64]). Plus récemment l'OTAN a normalisé un codeur LPC à 800 bit/s pour les communications HF [43].

Dans ces 3 codeurs, l'excitation est représentée de manière trop succincte. Pour un codeur à 2 400 bit/s, environ 1 850 bit/s sont dédiés à l'enveloppe spectrale et seulement 550 bit/s à l'excitation. La classification de l'excitation en 2 classes (V ou NV) n'est pas adaptée aux sons mixtes comme les fricatives voisées. Elle ne peut pas représenter les sons qui présentent un spectre harmonique jusqu'à une fréquence f_{max} puis une structure de bruit au-delà de f_{max}. Les sons plosifs ne sont pas correctement modélisés à l'aide d'un bruit blanc à l'énergie répartie sur la trame. Pour ces différentes raisons, le signal synthétique manque de clarté, est perçu comme bruité et présente des artefacts tonals. De plus si la classification V/NV est erronée ou si F_0 est mal estimée, la qualité du signal synthétique est fortement dégradée. Les défauts les plus audibles se produisent sur les zones voisées ou aux transitions. Ils sont essentiellement dus à une mauvaise représentation de l'évolution des paramètres de voisement.

II.3. Les nouveaux algorithmes de codage à bas débit

Dans les 10 dernières années, plusieurs algorithmes ont été proposés qui permettent un codage à bas débit avec une qualité de type *communication* (MOS autour de 3.5). Ces nouveaux algorithmes ont en commun une meilleure représentation des parties voisées du signal et de l'évolution des paramètres de voisement aux transitions entre sons. La plupart du temps, les paramètres spectraux sont codés par quantification vectorielle [18,35] sans distorsion audible pour un débit de 1 500 bit/s. Une pondération perceptuelle peut-être appliquée autour des formants, les paramètres LSF se prêtant bien à ce type de pondération.

Parmi les nouvelles méthodes de codage à bas débit, on peut distinguer les algorithmes de type codeurs harmoniques (MBE[5], STC[6]), les algorithmes à interpolation de forme d'onde (WI[7]) et les algorithmes à excitation mixte (MELP[8], HSX[9]).

5. MBE = Multiband Excited Coder.
6. STC = Sinusoidal Transform Coder.
7. WI = Waveform Interpolation.
8. MELP = Mixed Excitation Linear Prediction.
9. HSX = Harmonic Stochastic Coder.

La complexité de ces nouvelles approches est nettement supérieure à celle des codeurs LPC classiques, mais il est possible de les implanter sur un seul DSP en virgule fixe.

Les codeurs à modèles sinusoïdaux ou STC (Sinusoïdal Transform Coders)

Les codeurs STC (McAulay et Quatieri [38, 39, 40, 41]) modélisent la parole par une somme de sinusoïdes dont les amplitudes, les fréquences et les phases évoluent au cours du temps. Pour les parties voisées, les fréquences sont reliées aux harmoniques de F_0 et évoluent lentement au cours du temps. Les pics de la transformée de Fourier à court terme peuvent être utilisés pour déterminer les paramètres des sinusoïdes. Le nombre de sinusoïdes dans le modèle est variable car il dépend de F_0. Il a donc fallu développer des techniques de quantification vectorielle de vecteurs de longueur variable. Pour les sons non-voisés, l'excitation est un bruit blanc obtenu par une somme de sinusoïdes dont les fréquences sont uniformément réparties entre 0 et $f_e/2$. Différents modèles d'évolution de la phase ont été proposés [41].

Le codage STC donne de très bons résultats pour les débits moyens et pour la plage supérieure des bas débits. Un codeur sinusoïdal multi-débit a été développé aux MIT Lincoln Labs [40] avec des débits de 1.8 à 8 kbit/s. Pour les débits les plus faibles, les informations de phase ne sont pas transmises.

Les codeurs à excitation multibande ou Multi-Band Excited Coders (MBE)

Dans les codeurs MBE [22] et leurs variantes IMBE (*Improved MBE*) ou AMBE (*Advanced MBE*), le signal est analysé dans plusieurs bandes de fréquence adjacentes et est déclaré voisé ou non-voisé dans chacune des bandes. Le nombre de bandes d'analyse est de l'ordre du nombre d'harmoniques de F_0 entre 0 et $f_e/2$. La figure 4 représente l'amplitude de la transformée de Fourier discrète d'une trame de signal de parole avec ses zones harmoniques (voisées) ou non harmoniques (non-voisées) représentées par les signes V et UV.

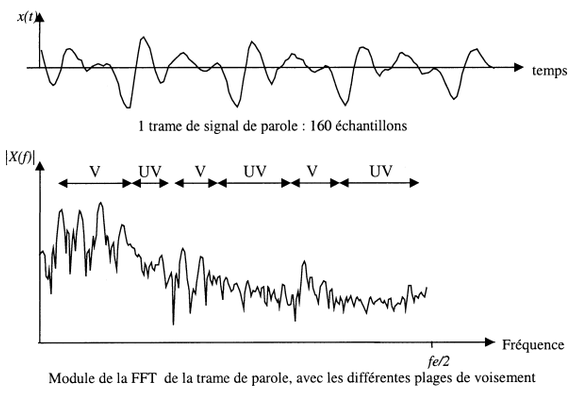

FIG. 4. — Zones voisées (V) ou non voisées (UV) du spectre d'une trame de parole.

Légende anglaise

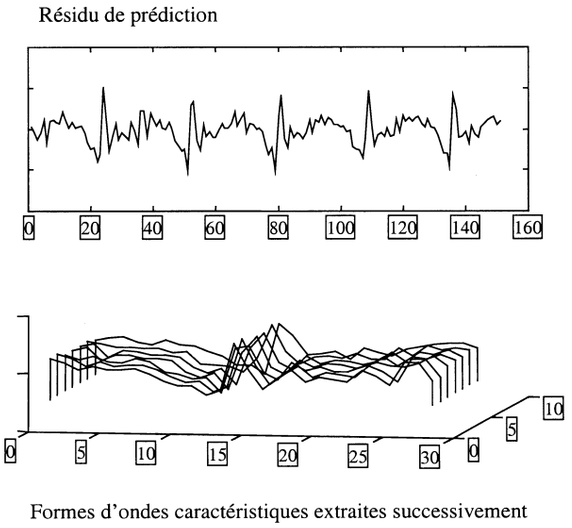

FIG. 5. — Trame de résiduel avec les CW correspondantes extraites toutes les 2.5 ms.

Légende anglaise

L'enveloppe spectrale $H(f)$ et la structure fine $E(f)$ de la transformée de Fourier discrète à court terme $X(f)$ de la trame de signal sont approchées séparément par $\hat{E}(f)$ et $\hat{H}(f)$. Le signal synthétique $\hat{x}_n$ est obtenu dans le domaine fréquentiel par $\hat{X}(f)=\hat{E}(f)\,\hat{H}(f)$. Les paramètres transmis par le codeur sont : la fréquence fondamentale, l'information de voisement pour chaque bande, et les paramètres décrivant l'enveloppe spectrale. Pour les bas débits, le voisement est estimé par groupe de quelques harmoniques.

L'algorithme IMBE (ou *Improved MBE coding*) a été normalisé à 4 150 bit/s pour le système Inmarsat-M avec 2 250 bit/s pour la correction d'erreur, d'où un débit total de 6 400 bit/s.

WI prototype Waveform Interpolation coders

Dans les codeurs WI à interpolation de formes d'onde (WI = Waveform Interpolation coders) [31,32,33,58], les paramètres spectraux correspondent aux coefficients de prédiction linéaire. La fréquence fondamentale est estimée et le résiduel de prédiction est calculé par filtrage du signal de parole par *A(z)*. Puis une forme d'onde caractéristique (CW = Characteristic Waveform) est extraite du signal résiduel à intervalles réguliers (typiquement à un rythme de 480 Hz). Cette extraction se fait en plaçant des marqueurs de pitch par détection de pics sur le signal résiduel suréchantillonné. Pour les sons voisés, la longueur des CW correspond à une période de pitch $p(t_m)$ à l'instant de calcul t_m. La figure 5 représente le signal résiduel d'une trame voisée de 20 ms et les CW correspondantes calculées toutes les 2,5 ms.

Pour les sons non voisés, la longueur des CW est arbitraire. La longueur de l'onde caractéristique $z(t_m,\tau)$ calcu-

lée à l'instant t_m est normalisée à 2π par la relation $u(t_m, \tau) = z(t_m,[p(t_m)/(2\pi)]\tau)$, puis alignée en temps avec l'onde précédente $u(t_{m-1},\tau)$. À chaque instant t est associé un signal périodique $u(t, \tau)$ de période 2π, représenté par les coefficients de sa série de Fourier. Ce signal est obtenu par interpolation linéaire (sur les coefficients de Fourier) entre 2 CW successives aux instants t_m et t_{m+1}. L'équation 1 donne la formule d'interpolation.

$$(1) \qquad u(t, \tau) = (1-\alpha(t))\, u(t_m,\tau)+\alpha(t)\, u(t_m+1,\tau)$$

Dans l'équation 1, $\alpha(t)$ est une fonction monotone croissante avec $\alpha(t_m) = 0$ et $(t_{m+1}) = 1$. La longueur dénormalisée d'une période de ce signal est obtenue par interpolation linéaire du pitch par l'équation 2.

$$(2) \qquad p(t) = (1-\alpha(t))\, p(t_m)+\alpha(t)\, p(t_{m+1})$$

Pour les segments voisés, la forme d'onde caractéristique évolue lentement tandis que pour les segments non-voisés elle évolue rapidement. Ces 2 composantes sont séparées par filtrages passe-bas et passe-haut de fréquence de coupure de 20 Hz appliqués à $u(t, \Phi)$ le long de l'axe t. Les 2 composantes à 2 dimensions résultant de ces filtrages sont appelées SEW (*Slowly Evolving Waveform*) et REW (*Rapidly Evolving Waveform*), Elles sont numérisées séparément de façon à exploiter au mieux la différence de perception de ces 2 signaux. Il est en effet inutile du point de vue perception de coder précisément la composante rapide REW. Une représentation grossière de la forme de son amplitude spectrale est suffisante. Mais ce signal évoluant rapidement, il faut transmettre ces informations à un rythme suffisamment élevé (par exemple 240 Hz). Il faut au contraire coder la composante SEW avec beaucoup de précision car l'oreille perçoit les distorsions même faibles sur ces sons périodiques. Mais on peut transmettre les paramètres de la composante SEW à un rythme lent (typiquement à 40 Hz, c'est-à-dire toutes les 25 ms). La quantification de la composante SEW est faite par quantification vectorielle des coefficients de sa série de Fourier.

Les paramètres d'analyse sont donc transmis à des rythmes différents, par exemple le pitch à 80 Hz, les paramètres LPC à 40 Hz, la puissance du signal à 80 Hz, les amplitudes des coefficients de la série de Fourier de la composante REW à 240 Hz, et les paramètres de la SEW à 40 Hz.

Le synthétiseur reconstruit les 2 composantes SEW et REW à partir de leurs coefficients de Fourier. La composante REW est obtenue en combinant les amplitudes reçues du codeur avec une phase aléatoire. À chaque instant t la forme d'onde $u(t, \tau)$ de période 2π, peut être calculée par interpolation linéaire des CW transmises (voir l'équation 1). La longueur dénormalisée de la forme d'onde est obtenue par interpolation linéaire sur le pitch par l'équation 2. L'excitation synthétique correspondante $e(t)$ est obtenue par l'équation 3.

$$(3) \qquad e(t)=u(t, \Phi(t))=u\left(t, \Phi(t_m) + \int_{-\infty}^{t} \frac{2\pi}{p(u)}\, du\right)$$

L'excitation totale reconstruite $e(t)$ est obtenue en ajoutant les coefficients de Fourier des composantes REW et SEW. Elle est ensuite filtrée par le filtre de synthèse LPC. Les paramètres LPC sont interpolés linéairement à chaque instant. Un filtre de renforcement des formants est appliqué pour améliorer la qualité subjective du signal.

Un codeur WI travaillant à 2 400 bit/s [33] donne de meilleurs résultats subjectifs que la norme FS1016 à 4800 bit/s utilisant un codage CELP. Le modèle WI n'est pas limitatif, on peut obtenir une meilleure qualité en augmentant le débit.

Les codeurs lpc à excitation mixte ou meelp Mixed Excitation Linear Prediction Coders

Le nouveau standard DOD à 2 400 bit/s [42,61,62] est un codeur LPC à excitation mixte (MELP = *Mixed Excitation Linear Prediction*).

Il utilise une excitation mixte c'est-à-dire formée de la somme d'une composante impulsionnelle et d'une composante de bruit. La composante impulsionnelle est formée d'un train d'impulsions périodique ou non. Cette excitation est une excitation multibande avec une intensité de voisement définie pour chaque bande de fréquence.

Le codeur fait une première estimation de la fréquence fondamentale, puis il calcule l'intensité de voisement dans 5 bandes de fréquence adjacentes. L'intensité de voisement est déterminée dans chaque bande par la valeur de l'autocorrélation normalisée par la valeur de la période de pitch. Dans la norme, cette intensité est codée sur 1 bit, chaque bande est donc classée voisée ou non-voisée. Après analyse le codeur peut positionner un indicateur appelé indicateur d'apériodicité (« *aperiodic flag* ») pour indiquer au décodeur que la composante impulsionnelle doit être apériodique. Le codeur effectue par ailleurs une analyse spectrale par prédiction linéaire et calcule les amplitudes des 10 premières harmoniques du pitch sur la transformée de Fourier du signal résiduel. Ces amplitudes sont quantifiées de manière vectorielle. Les paramètres transmis par le codeur sont finalement : la période fondamentale, le drapeau d'apériodicité, les 5 intensités de voisement, 2 gains (correspondant aux énergies de 2 demi-trames), les paramètres spectraux et les 10 amplitudes d'harmoniques du pitch codés par quantification vectorielle.

Le synthétiseur interpole linéairement les différents paramètres de manière synchrone au pitch. La composante impulsionnelle est obtenue sur une période de pitch par transformée de Fourier inverse sur les 10 amplitudes de Fourier. Pour les sons non-voisés ou lorsque l'indicateur d'apériodicité est positionné, une perturbation aléatoire (*jitter*) est appliquée à la valeur de la période fondamentale. Cette possibilité d'excitation impulsionnelle non périodique est particulièrement intéressante pour les zones de transitions entre sons. La composante impulsionnelle et la composante de bruit sont filtrées puis ajoutées. Le filtrage appliqué à la composante impulsionnelle a pour réponse impulsionnelle la somme de toutes les réponses impulsionnelles des filtres passe-bande pour les bandes voisées. Le filtrage de la compo-

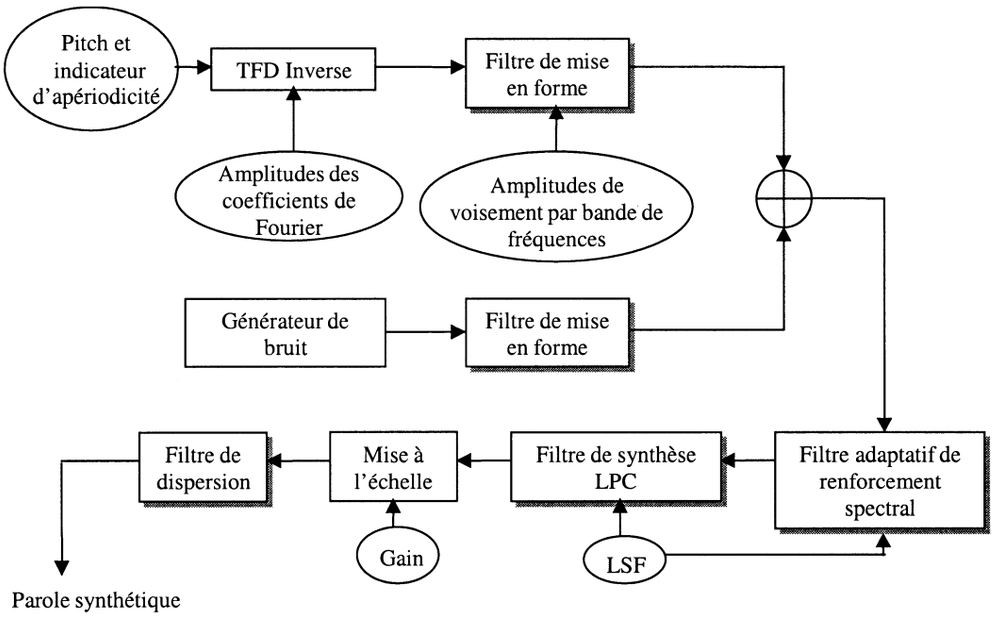

FIG. 6. — Synthèse MELP.

Légende anglaise

sante de bruit est déterminé de la même façon à partir des bandes non-voisées. L'excitation globale est ensuite filtrée par un filtre adaptatif de renforcement des formants et par le filtre de synthèse LPC. Le signal synthétique résultant est mis à l'échelle en fonction de l'énergie de la trame originale et passé dans un filtre dont le but est d'étaler l'énergie des impulsions sur une période de pitch (pulse dispersive filter).

La figure 6 représente la synthèse MELP.

La qualité obtenue avec cette norme correspond à la qualité dite de *communication* (MOS autour de 3,5) qui est légèrement inférieure à la qualité téléphonique classique. Cette qualité est nettement supérieure à celle du standard précédent LPC10e à 2 400 bit/s.

Les codeurs HSX ou Harmonic Stochastic eXcitation coders

Le codage HSX [34, 21] est très proche d'un point de vue conceptuel du codage MELP. La modélisation de l'excitation est plus élémentaire ce qui permet d'obtenir des débits plus bas et assure une complexité plus faible.

L'excitation synthétique d'un codeur HSX est la somme d'une composante harmonique et d'une composante stochastique. L'excitation est harmonique jusqu'à une fréquence limite f_{max} puis stochastique au-delà de cette fréquence. Le spectre de l'excitation est plat. Le codeur détermine la fréquence fondamentale, l'énergie, les paramètres LPC, l'intensité de voisement dans 4 bandes de fréquence adjacentes. L'intensité de voisement est contrainte à être une fonction décroissante de la fréquence. Le codeur détermine la fréquence f_{max} par analyse multibande. Le synthétiseur filtre l'excitation mixte par le filtre de synthèse LPC et par un filtre de renforcement des formants puis met à l'échelle le résultat en fonction de l'énergie de la trame originale.

Ce principe de codage permet d'obtenir des débits de l'ordre de 600 bit/s avec une qualité subjective très supérieure au standard LPC10.

Ce codeur est décrit en détail dans la section 3.

Le codeur MPEG-4 HVXC de Sony

La norme MPEG-4 propose un ensemble d'outils pour le codage des sons naturels (tels que la parole et la musique) et pour la synthèse de sons (sources musicales MIDI, synthèse à partir du texte, effets sonores tels que réverbération et spatialisation) [25]. Parmi ces outils, les techniques de codage par transformée AAC (Advanced Audio Coding) et TwinVQ sont recommandées pour le codage de l'audio au-dessus de 6 kbit/s ; le codage CELP est recommandé pour le codage de la parole (en bande étroite ou en bande élargie) entre 4 et 24 kbit/s ; enfin, le plus bas débit consacré au codage de la parole est pris en charge par le codeur HVXC (Harmonic Vector Excitation Coding) de Sony [44]. Ce codeur est hiérarchique (ou « graduel ») : son schéma de quantification génère deux flux binaires à 2 kbit/s et 4 kbit/s totalement imbriqués (le décodage à 2 kbit/s est possible en utilisant une partie seulement du flux à 4 kbit/s). Cette propriété du train binaire est particulièrement intéressante pour toutes les applications pour lesquelles la capacité du canal de transmission est variable dans le temps, que ce soit pour des raisons physiques (canal à évanouissement comme celui rencontré en HF) ou à cause des limitations du système (congestion du réseau) ; elle simplifie également le pro-

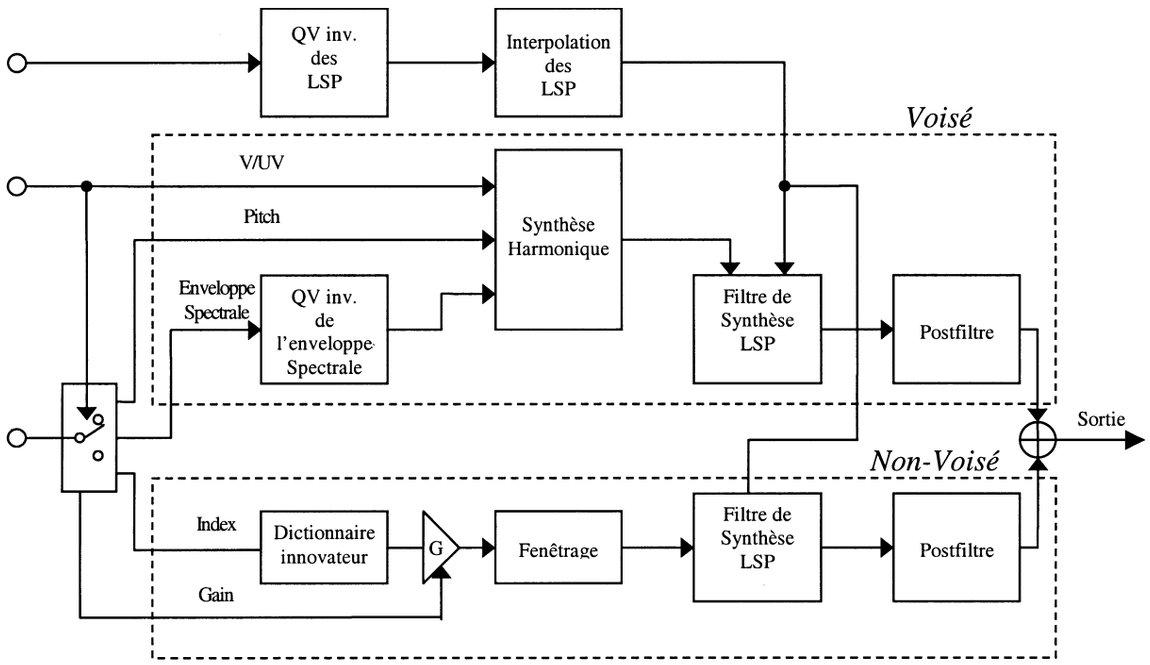

FIG. 7. — Synthèse HVXC.

Légende anglaise

blème du transcodage lorsque la capacité est variable le long du canal de transmission (rencontre d'un tronçon de plus faible capacité).

La figure 7 représente la partie synthèse du codeur HVXC. La principale caractéristique de ce codeur est qu'il met en œuvre deux schémas de codage très différents selon que le signal de parole est voisé ou non-voisé : une technique paramétrique est utilisée pour le codage des sons voisés ; une technique à analyse par synthèse de type CELP est utilisée pour le codage des sons non-voisés.

Le signal de parole est segmenté en trames de 20 ms. Le codeur réalise tout d'abord une analyse LPC d'ordre 10 sur des fenêtres de 256 échantillons. Les coefficients LPC ainsi obtenus sont convertis en LSP et codés par quantification vectorielle (quantification prédictive multi-étage). Le résidu de l'analyse LPC est calculé en utilisant les filtres de prédiction linéaire quantifiés et interpolés. Une première estimation du pitch est tout d'abord obtenue en boucle ouverte sur la base des maxima de l'autocorrélation du résidu. Une procédure de suivi de pitch exploite les valeurs successives du pitch et du voisement. Cette estimation est ensuite raffinée par une procédure d'estimation de la valeur fine du pitch et de l'enveloppe spectrale aux harmoniques (cette dernière procédure est similaire à celle mise en œuvre dans le codeur MBE). Le codeur effectue alors une décision de voisement sur la base du nombre de passages par zéro, de la structure harmonique et du maximum de l'autocorrélation du résidu de prédiction linéaire. Lorsqu'une trame est déclarée non-voisée, un codeur de type CELP est mis en œuvre. Ce codeur CELP ne comporte pas de prédicteur à long terme.

Les résultats des tests présentés en [44] montrent que le codeur HVXC présente aux deux débits une qualité significativement supérieure à celle du codeur CELP à 4.8 kbit/s (standard américain FS1016).

Codage Multitrame

Un codage multitrame peut être appliqué pour diminuer le débit des codeurs précédents. Le standard OTAN à 800 bit/s [43] correspond à un codage LPC10 dans lequel on code globalement les paramètres de 3 trames successives.

II.4. Les codeurs à très bas débits

Pour obtenir des débits inférieurs à quelques centaines de bits par seconde, il n'est plus possible de travailler sur des trames de longueur fixe. Une approche segmentale utilisant des segments de longueur variable est nécessaire[6, 8, 10, 14, 24, 27, 36, 46, 47, 50, 51, 52, 53, 54, 56, 57, 63, 66].

On peut considérer que les codeurs à très bas débit effectuent une reconnaissance de segments acoustiques dans la phase d'analyse et une synthèse de parole à partir d'une suite d'indices de segments dans le décodeur. Le codeur réalise une transcription symbolique du signal de parole à partir d'un dictionnaire d'unités élémentaires de taille variable qui peuvent être des unités linguistiques (comme des phonèmes, des transitions entre phonèmes,

des syllabes), on parle alors de vocodeurs phonétiques, ou bien des unités acoustiques obtenues automatiquement de manière non supervisée sur un corpus d'apprentissage, on utilisera par la suite l'expression vocodeurs pseudo-phonétiques pour désigner ces derniers codeurs.

On distingue 2 approches. La 1re segmente le signal de parole par différentes méthodes telles que l'identification de régions stables puis code les séquences de vecteurs spectraux de longueur variable par des techniques comme la quantification matricielle par exemple. Dans la 2e approche, la segmentation et la quantification sont effectuées simultanément, à l'aide de techniques de reconnaissance d'unités de longueur variables utilisant des modèles de Markov cachés HMM (*Hidden Markov Model*) ou une technique DTW (*Dynamic Time Warping*).

Dans la 1re approche, la segmentation de la séquence de vecteurs spectraux peut se faire en comparant à un seuil une approximation de la dérivée des vecteurs spectraux. Souvent, les segments vont du milieu d'une zone stable au milieu de la zone suivante. Deux techniques sont couramment utilisées pour le codage des séquences de vecteurs spectraux, la quantification matricielle et le codage VFR (*Variable Frame Rate*).

La quantification matricielle [54,5] code une suite de vecteurs spectraux de dimension p à l'aide d'un dictionnaire de matrices-codes, de dimension (N,p), formées de N vecteurs spectraux. Si la longueur des séquences de vecteurs à coder est variable, on peut effectuer un alignement temporel (par DTW par exemple) entre la séquence et les matrices du dictionnaire, aussi bien lors de l'apprentissage que lors de la classification. Il faut alors transmettre une information sur la durée réelle du segment. Dans [52], une contrainte est ajoutée sous la forme d'un réseau qui détermine quelles matrices-codes peuvent suivre une matrice-code donnée.

Dans la technique du codage VFR, on ne code et on ne transmet qu'un nombre réduit de vecteurs d'une séquence donnée. Au décodeur, les vecteurs manquants sont récupérés par interpolation à partir des vecteurs transmis [37, 30]. Le choix des vecteurs à coder est fait au codeur soit en boucle ouverte soit en boucle fermée. En boucle ouverte, on détermine les vecteurs en repérant ceux pour lesquels la dérivée des paramètres spectraux est la plus grande ou ceux qui présentent le plus grand écart par rapport à une interpolation effectuée sur les vecteurs adjacents. En boucle fermée, on choisit les vecteurs qui permettent d'obtenir la plus faible distorsion spectrale en synthèse, en testant toutes les possibilités.

La 2e approche, par segmentation et quantification conjointes, utilise soit des matrices de longueur fixe [57], soit des matrices de longueur variable [13, 3], on parle alors de VVVQ *Variable to Variable Vector Quantization*, ou bien des modèles HMM.

L'approche phonétique ou pseudo-phonétique qui reconnaît la suite des phonèmes ou des unités acoustiques constituant le signal original est une technique de segmentation et indexation conjointes.

Quelle que soit la méthode de codage utilisée, pour chaque segment, le codeur transmet le symbole correspondant à l'unité reconnue ainsi que des paramètres auxiliaires tels que les contours de fréquence fondamentale et d'énergie, et la longueur du segment. La synthèse se fait généralement par concaténation de représentants des unités élémentaires. Elle peut utiliser les techniques PSOLA (*Pitch Synchronous Overlap and Add*) ou HNM [60] (*Harmonic plus Noise Model*).

Le débit moyen nécessaire pour coder la séquence d'unités reconnues est compris entre 50 et 150 bit/s (soit un débit moyen de 12 segments par seconde et 50 à 2000 unités). À ce débit il faut ajouter le débit des paramètres auxiliaires qui est du même ordre de grandeur.

Le retard introduit par ces codeurs est grand, de l'ordre de quelques centaines de ms.

Le dictionnaire d'unités élémentaires peut contenir des séquences de vecteurs spectraux de longueur variable, des segments de parole, des modèles HMM décrivant les unités.

Les vocodeurs phonétiques nécessitent la transcription phonétique du corpus d'apprentissage, tâche lourde et sujette aux erreurs qui doit être effectuée pour chaque nouvelle langue. La détermination automatique d'unités acoustiques à partir d'un corpus de parole non étiqueté est donc une approche intéressante.

Les sections suivantes (3 et 4) présentent en détail nos travaux dans le domaine des codeurs à bas et très bas débit. La section 3 décrit un codeur HSX à bas débit (1 200 bit/s ou 600 bit/s) qui présente une qualité suffisante pour des applications commerciales. La section 4 présente une nouvelle approche segmentale pseudo-phonétique de codage à très bas débit s'appuyant sur un jeu d'unités obtenues automatiquement sur un corpus d'apprentissage non étiqueté phonétiquement.

III. LE CODEUR HSX DE THOMSON-CSF COMMUNICATIONS

Le codeur à bas débit présenté dans ce chapitre est basé sur le modèle paramétrique HSX (Harmonic Stochastic excitation) développé par Thomson-CSF Communications en collaboration avec l'Université de Sherbrooke (Canada) [34]. Ce codeur a été sélectionné récemment pour une application de radio messagerie vocale en Amérique du Nord (« pager » vocal Mobi-Darc® de la société Info Télécom [21]). Ce chapitre donne tout d'abord une brève description de la technique HSX, puis présente le procédé de codage et de quantification des paramètres pour un fonctionnement à 1 200 bit/s. Il récapitule les exigences en termes de puissance CPU et d'occupation mémoire pour une implémentation sur un processeur de traitement de signal opérant en virgule fixe (TI C54x) et en virgule flottante (TI C3x). Enfin, il donne ensuite quelques résultats d'évaluations subjectives des performances du codeur.

III.1. La technique hsx

III.1.1. Le modèle d'excitation mixte

Le vocodeur HSX (pour *Harmonic Stochastic eXcitation*) utilise un modèle d'excitation mixte simple, dans lequel le train d'impulsion périodique excite les fréquences basses et le bruit les fréquences hautes du filtre LPC de synthèse (figure 8). La fréquence de coupure f_c est variable dans le temps. Les deux filtres de mise en forme de l'excitation sont complémentaires et le gain g est ajusté de sorte que l'excitation mixte soit à spectre plat.

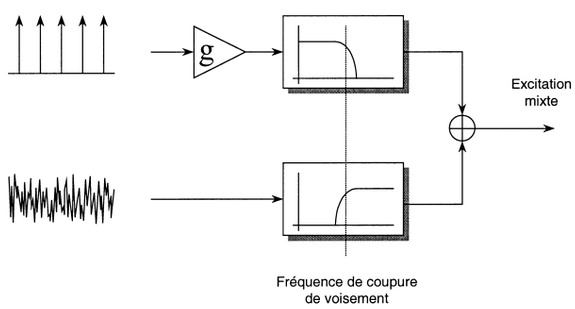

FIG. 8. — Modèle d'excitation mixte du vocodeur HSX.

Légende anglaise

III.1.2. Description de l'analyse

Le diagramme de la partie analyse du vocodeur HSX est représenté sur la figure 8. Le signal de parole échantillonné à 8 kHz est segmenté en trames de 180 échantillons (22.5 ms).

La première étape de l'analyse consiste à éliminer les composantes à très basse fréquence (en particulier les perturbations à 50 ou 60 Hz causées par le secteur). Cette opération est effectuée par un filtre IIR à deux pôles et deux zéros coupant approximativement à 200 Hz. Deux analyses LPC d'ordre 10 sont alors effectuées par la méthode de l'autocorrélation sur des fenêtres de Hamming de 240 échantillons (30 ms). Pour la première analyse, cette fenêtre est centrée sur le milieu de la trame. Pour la seconde, elle est centrée sur le dernier échantillon de la trame. Une fenêtre de pondération des retards est appliquée aux coefficients de corrélation (fenêtre gaussienne réalisant une expansion en fréquence de 50 Hz avec un plancher de bruit à 1.0E-04). Les coefficients de prédiction linéaire obtenus par l'algorithme de Durbin-Levinson sont alors convertis en paires de raies spectrales (LSF) en vue de leur quantification. Le signal de parole est ensuite passé dans un filtre semi-blanchisseur de la forme $A(z)/A(z/\gamma)$, avec $\gamma=0.75$ et où $A(z)$ est le filtre de résidu de prédiction linéaire. Cette opération diminue la structure formantique du signal de parole, sans toutefois l'éliminer (on conserverait le signal de parole d'origine pour $\gamma = 1.0$, et on obtiendrait le résidu de prédiction pour $\gamma = 0.0$) ; elle aide au fonctionnement du suiveur de pitch et de l'analyseur du voisement. Le signal semi-blanchi est filtré en 4 sous-bandes par un jeu de filtres FIR ayant respectivement pour bande passante 0 à 800 Hz, 700 à 1 700 Hz, 1 500 à 2 500 Hz et 2 300 à 3 300 Hz. Une première estimée grossière de la valeur du pitch est obtenue par une technique de corrélation sur le signal filtré dans la première sous-bande. Un effort particulier est fait pour éviter les doublements de pitch. La fréquence de coupure de voisement est ensuite déterminée grâce aux taux de voisement dans les quatre sous-bandes. Lors de cette opération, la valeur du pitch est raffinée avec une précision de l'ordre du quart d'échantillon. Les suiveurs de pitch et de voisement mettent en œuvre des logiques de décision complexes basées sur trois trames successives ; ils rendent leur résultat avec une trame de retard. Enfin,

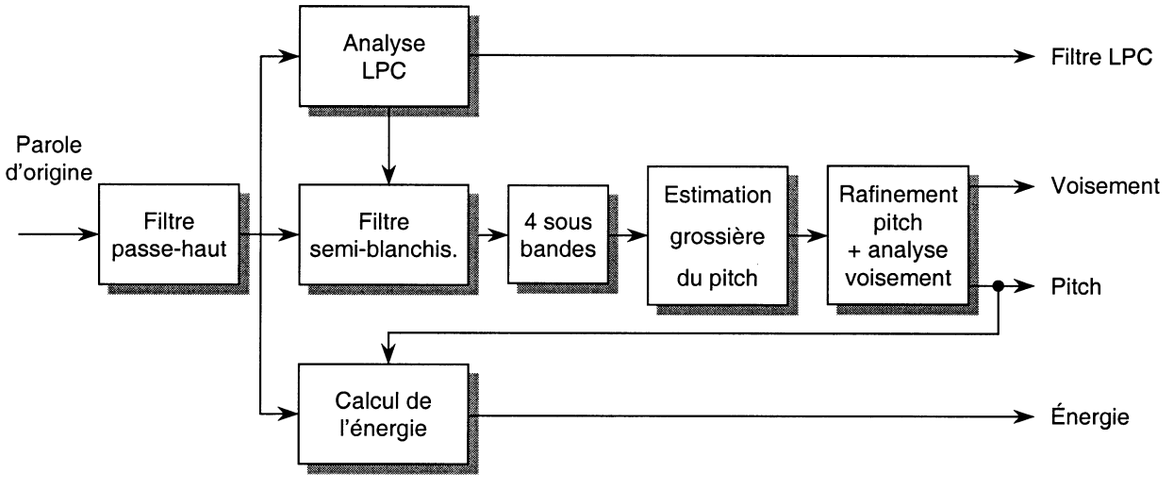

FIG. 9. — Diagramme de l'analyse *HSX*.

Légende anglaise

l'énergie du signal de parole est calculée quatre fois par trame. Durant les passages voisées, elle est calculée sur une fenêtre rectangulaire synchrone avec les impulsions de pitch. Pendant les passages non-voisées, elle est calculée sur des sous-trames fixes de 45 échantillons. L'énergie est exprimée en dB par échantillon en vue de sa quantification.

III.1.3. Description de la synthèse

Le diagramme de la partie synthèse du vocodeur HSX est représenté sur la figure 9. L'excitation harmonique est obtenue en juxtaposant – au rythme d'une réponse par période de pitch – des réponses impulsionnelles de filtres passe-bas pré-calculées. La valeur du pitch est interpolée pour chaque nouvelle impulsion avec une précision d'un demi-échantillon. La valeur de la fréquence de coupure de voisement est également interpolée, avec une précision de 250 Hz. Lorsque le signal n'est pas voisé, le pitch est fixé à 45 échantillons et aucune impulsion n'est synthétisée. La position du premier échantillon de chacune de ces impulsions est mémorisée ; par la suite, le filtre LPC et l'énergie seront systématiquement interpolés au niveau de cet échantillon. L'excitation stochastique est obtenue par une technique combinant des transformées de Fourier inverses sur 128 points et une addition avec recouvrement (overlap-add). À ce niveau, l'information de voisement est interpolée deux fois par trame. Le signal d'excitation mixte est la somme de l'excitation harmonique et de l'excitation stochastique. On considère maintenant chacune des périodes comprises entre le début d'une impulsion et le début de l'impulsion suivante. On applique à l'excitation mixte un gain égal à l'énergie interpolée moins le gain du filtre LPC de synthèse et celui du post-filtre. Cette excitation ajustée en gain est passée dans le filtre LPC de synthèse interpolé $1/A_i(z)$ puis dans le filtre perceptuel. Ce dernier est un filtre adaptatif basé sur les coefficients de prédiction linéaire qui permet une meilleure restitution des caractéristiques spectrales (en particulier nasales) du signal de parole d'origine. Il possède pour expression :

$$H(z)\cdot p(z),$$

avec

$$H(z)=\frac{A_i(z/\gamma_1)}{A_i(z/\gamma_2)},\quad avec\ \gamma_1=0.65\ \ et\ \gamma_2=0.80,$$

et avec une compensation au premier ordre de la pente introduite par H(*z*) :

$$p(z)=1-\mu z^{-1},\ avec\ \mu=0.80\,\frac{r_1}{r_0}.$$

Dans cette dernière équation, r_0 et r_1 sont les premier et second coefficients d'autocorrélation de la réponse impulsionnelle de $H(z)$. Un second contrôle de gain est appliqué à la sortie du filtre perceptuel. Enfin, un post-filtre fixe (filtre FIR passe-tout) permettant de rendre le signal plus naturel est appliqué au signal de synthèse.

III.2. Procédé de quantification à 1 200 bit/s

Un débit binaire aussi bas que 1 200 bit/s ne permet pas d'encoder parfaitement les paramètres pour toutes les trames de 22.5 ms. Il est alors nécessaire de regrouper N trames successives en une seule multitrame, de sorte que le procédé de codage et de quantification des paramètres puisse exploiter au maximum les périodes de stabilité du signal de parole. Comme dans la norme OTAN 4479 [43], la valeur N = 3 a été choisie parce qu'elle permet d'obtenir un bon compromis entre la réduction possible du débit binaire et le délai de bout en bout.

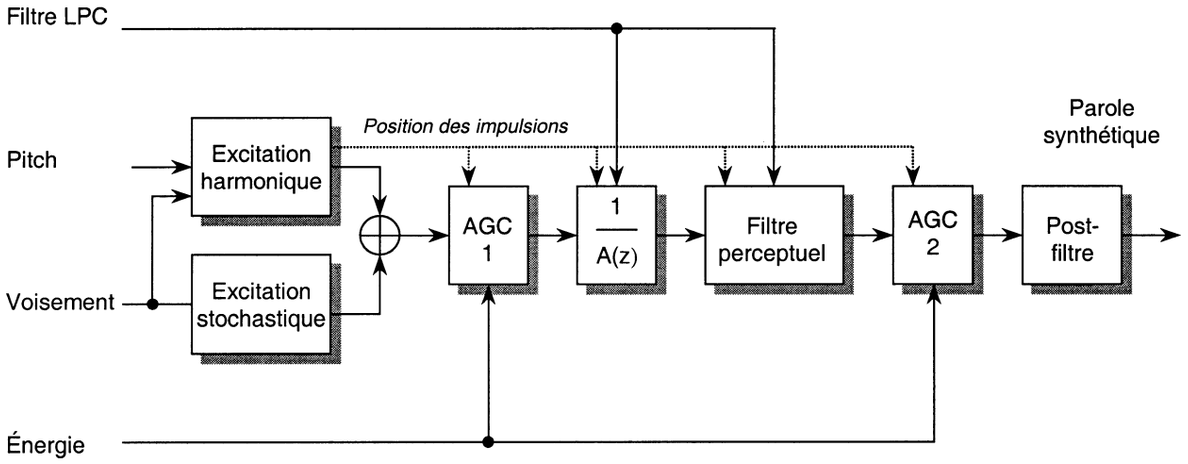

FIG. 10. — Diagramme de la synthèse *HSX*.

Légende anglaise

III.2.1. Encodage du voisement

La fréquence de transition de voisement peut être décrite de façon efficace en utilisant uniquement les quatre valeurs suivantes : 0, 750, 2 000 et 4 000 Hz. En théorie, 6 bits sont donc nécessaires pour transmettre exactement la configuration de voisement pour les trois trames. Toutefois, certaines configurations de voisement ne se présentent que très rarement, elles ne sont pas forcément caractéristiques de l'évolution d'un signal de parole normal, et elles ne semblent participer ni à l'intelligibilité, ni à la qualité de la parole restituée (par exemple, une trame voisée jusqu'à 4 000 Hz comprise entre deux trames totalement non voisées). La répartition des configurations de voisement sur trois trames successives, calculées sur une base de données de 123 158 multitrames de parole, est présentée dans le tableau I. Les 32 configurations les moins fréquentes comptent pour seulement 4 % de toutes les multitrames (partiellement ou totalement) voisées. La dégradation obtenue en remplaçant chacune de ces configurations par la plus proche (en termes d'erreur absolue) des 32 configurations les plus représentées est imperceptible. Ceci montre qu'il est possible d'économiser un bit en quantifiant vectoriellement la fréquence de transition de voisement sur une multitrame.

TABLE I. — Répartition des configurations de voisement sur 3 trames successives.

Légende anglaise

Configuration de voisement	**Nb. Multitrames**
Totalement non voisé (silences compris)	55 585
Totalement voisé	34 586
Partiellement voisé (30 plus fréquentes)	30 273
Partiellement voisé (32 moins fréquentes)	2 714
Nombre total de multitrames :	*123 158*

III.2.2. Encodage du pitch

On utilise un quantificateur scalaire sur 6 bits, avec un pas de quantification uniforme sur une échelle logarithmique, sur une échelle de 18 à 148 échantillons. Une seule valeur est transmise pour trois trames consécutives. Le calcul de la valeur à quantifier à partir des trois valeurs de pitch, et la procédure permettant de récupérer les trois valeurs de pitch à partir de la valeur quantifiée, diffèrent selon la valeur des fréquences de transition de voisement à l'analyse :

1. Lorsqu'aucune trame n'est voisée, les 6 bits sont positionnés à zéro, le pitch décodé est fixé à 45.0 (cette valeur correspond à un quart de la trame ; elle est également une valeur moyenne pour le pitch) pour chacune des trames de la multitrame.
2. Lorsque la dernière trame de la multitrame précédente et les trois trames de la multitrame courante sont voisées (fréquence de transition de voisement strictement supérieure à 0), on quantifie la valeur du pitch de la dernière trame de la multitrame courante (valeur cible). Au décodeur, la valeur du pitch pour la troisième trame de la multitrame courante est la valeur cible quantifiée, et les valeurs du pitch pour les deux premières trames de la multitrame courante sont récupérées par interpolation linéaire entre la valeur transmise pour la multitrame précédente et la valeur cible quantifiée.
3. Pour toutes les autres configurations de voisement, on quantifie la moyenne pondérée du pitch sur les trois trames de la multitrame courante (valeur moyenne pondérée). Le facteur de pondération est proportionnel à la fréquence de transition de voisement pour la trame considérée :

Valeur Moyenne Pondérée

$$= \frac{\sum_{i=1\ldots3} \text{Pitch}\,(i)*\text{Voisinement}(i)}{\sum_{i=1\ldots3} \text{Voisinement}(i)}$$

Au décodeur, la valeur du pitch utilisée pour les trois trames de la multitrame courante est égale à la valeur moyenne pondérée quantifiée.

De plus, dans les cas 2 et 3, on applique systématiquement un léger trémolo à la valeur du pitch utilisée en synthèse. Ceci permet d'améliorer le naturel de la parole restituée en évitant la génération de signaux trop longtemps périodiques.

L'utilisation d'un quantificateur scalaire limite le problème de propagation des erreurs sur le train binaire. De plus, les schémas de codage 2 et 3 sont suffisamment proches l'un de l'autre pour être insensibles aux mauvais décodages de la fréquence de voisement.

III.2.3. L'énergie

Douze valeurs de l'énergie (numérotées de 0 à 11) doivent être transmises pour chaque multitrame. On sélectionne 6 valeurs parmi les 12, on construit deux vecteurs de 3 valeurs, et on quantifie chacun des vecteurs sur 6 bits (quantificateur vectoriel prédictif en boucle fermée avec un coefficient de prédiction égal à 0.5). Deux bits sont utilisés pour transmettre le numéro du schéma de sélection utilisé. Au niveau du décodeur, les valeurs de l'énergie qui n'ont pas été quantifiées sont récupérées par interpolation.

TABLE II. — Liste de schémas de sélection et d'interpolation pour l'encodage de l'énergie.

Légende anglaise

Nom du schéma	**Vecteur 1**	**Vecteur 2**	**Valeurs interpolées**
Stable	1, 3, 5	7, 9, 11	0, 2, 4, 6, 8, 10
Trame 1	0, 1, 2	3, 7, 11	4, 5, 6, 8, 9, 10
Trame 2	1, 4, 5	6,7,11	0, 2, 3, 8, 9, 10
Trame 3	2, 5, 8	9,10,11	0, 1, 3, 4, 6, 7

Seuls 4 schémas de sélection sont autorisés et sont décrits dans le tableau II. Ces schémas ont été optimisés afin d'encoder au mieux soit les vecteurs de 12 énergies stables, soit ceux pour lesquels l'énergie varie rapidement au cours des trames 1, 2 ou 3. On encode le vecteur d'énergie selon chacun des quatre schémas, et le schéma effectivement transmis est celui qui minimise l'erreur quadratique totale.

Les bits donnant le numéro du schéma ne peuvent pas être considérés comme « sensibles », puisqu'une erreur sur leur valeur ne fait qu'altérer légèrement l'évolution temporelle de la valeur de l'énergie. De plus, la table de quantification vectorielle des énergies est organisée de sorte que l'erreur quadratique moyenne produite par une erreur sur un bit d'adressage soit minimale.

III.2.4. Les filtres de prédiction linéaire

Six filtres LPC (numérotés de 0 à 5) à 10 coefficients doivent être transmis pour chaque multitrame. Ces six vecteurs de 10 coefficients LPC sont transformés en six vecteurs de 10 LSF (« paires de raies spectrales » exprimées en Hz). Ils sont alors encodés par une technique similaire à celle utilisée pour l'énergie : on sélectionne trois filtres LPC, on quantifie chacun de ces vecteurs sur 18 bits (quantificateur vectoriel prédictif en boucle ouverte, avec un coefficient de prédiction égal à 0.6, de type SPLIT-VQ portant sur deux sous-paquets de 5 LSF consécutives à chacun desquels on alloue 9 bits). Deux bits sont utilisés pour transmettre le numéro du schéma de sélection utilisé. Au niveau du décodeur, lorsqu'un filtre LPC n'est pas quantifié, sa valeur est estimée à partir de celle des filtres LPC quantifiés par interpolation (par exemple, interpolation linéaire) ou par extrapolation (par exemple, duplication du filtre LPC précédent).

Seuls 4 schémas de sélection sont autorisés et sont décrits dans le tableau III. Ces schémas ont été optimisés afin d'encoder au mieux, soit les zones pour lesquelles l'enveloppe spectrale est stable, soit les zones pour lesquels l'enveloppe spectrale varie rapidement au cours des trames 1, 2 ou 3. On encode l'ensemble des filtres LPC selon chacun des quatre schémas, et le schéma effectivement transmis est celui qui minimise l'erreur quadratique totale.

Comme pour l'encodage de l'énergie, les bits donnant le numéro du schéma ne peuvent pas être considérés comme « sensibles » aux erreurs de transmission, puisqu'une erreur sur leur valeur ne fait qu'altérer légèrement l'évolution temporelle des filtres LPC. De plus, les tables de quantification vectorielle des LSF sont organisées de sorte que l'erreur quadratique moyenne produite par une erreur sur un bit d'adressage soit minimale.

TABLE III. — Liste de schémas de sélection et d'interpolation /extrapolation pour l'encodage de filtres LPC.

Légende anglaise

Nom du schéma	LPC quantité	LPC interpolé	Valeurs extrapolé
Stable	1, 3, 5	0, 2, 4	–
Trame 1	0, 1, 4	2, 3	5
Trame 2	2, 3, 5	0, 1, 4	–
Trame 3	1, 4, 5	0, 2, 3	–

III.2.5. Allocation des bits

Ce codeur opère à 1 200 bits par seconde avec un codage des paramètres toutes les 67.5 ms ; 81 bits sont donc disponibles à chaque multitrame pour encoder les paramètres du signal. Ces bits sont alloués aux différents paramètres comme il est indiqué dans le tableau IV.

TABLE IV. — Allocation des bits pour le vocodeur HSX à 1 200 bit/s.

Légende anglaise

Paramètre	NB. Bits
LSFs	54
Schéma de décimation (LSFs)	2
Energie	2*6
Schéma de décimation (énergie)	2
Pitch	6
Voisement	5
Total bits / 67.5 ms	81

Le délai système minimal pour une application utilisant le codeur HSX à 1 200 bit/s est la somme du délai algorithmique, du délai de traitement et du délai pour une transmission à 1 200 bits par seconde. Le délai algorithmique lié uniquement à l'analyse et à la synthèse paramétrique est égal à 127.5 ms. Le délai de traitement pour une implémentation en temps réel peut être de 45 ms (une trame de 22.5 ms au codeur et une trame de 22.5 ms au décodeur). Le délai de transmission à 1 200 bits par seconde de 81 bits toutes les 67.5 ms est bien évidemment égal à 67.5 ms. On peut donc dire que le délai système minimal pour une mise en œuvre du codeur HSX à 1 200 bit/s est égal à 240 ms. L'implantation finale pourra donc faire apparaître quelques délais supplémentaires, liés par exemple à l'utilisation d'un entrelaceur, au temps de propagation de l'onde porteuse, ou encore à la construction et l'acheminement de paquets sur un réseau de type IP.

III.3. Exigences d'implémentation

Ce paragraphe présente les exigences pour une implémentation du codeur HSX à 1 200 bit/s sur un processeur de traitement du signal opérant en virgule fixe (TI C54x) et en virgule flottante (TI C3x). Dans l'application de *pager* vocal, seule la partie décodeur du codeur HSX devait être implémentée en temps réel sur le proces-

seur TI C54x ; la partie codage était effectuée en temps différé sur un serveur.

III.3.1. Implémentation en virgule fixe

Le tableau V donne les exigences en termes de mémoire et de puissance de calcul pour une mise en œuvre du codeur sur un processeur Texas Instruments de la famille C54x. A partir de la simulation en virgule flottante, nous avons tout d'abord établi une description algorithmique complète en virgule fixe utilisant les opérateurs de base du C-ETSI[10] (EFR-GSM, G.729, G.723, …). Le codeur et le décodeur ont ensuite été entièrement réécrits et optimisés pour l'assembleur C54x. L'occupation mémoire est donnée en kilo-mots (kw[11]). Pour le processeur TI C54x, un mot est formé de 16 bits.

TABLE V. — Exigences d'implémentation du HSX 1 200 (DSP TI C54x).

Légende anglaise

Exigences	Full-Duplex	Codeur	Décodeur
Mémoire programme	7 kw	4.5 kw	3 kw
Tables de données	9.2 kw	8.7 kw	9.2 kw
RAM de travail	5.5 kw	4 kw	2.5 kw
Complexité	22 MIPS	18 MIPS	4 MIPS

On notera que l'existence d'une description algorithmique complète en virgule fixe est susceptible de faciliter le l'implantation du codeur sur tout autre processeur en virgule fixe.

III.3.2. Implémentation en virgule flottante

Le tableau VI donne les exigences en termes de mémoire et de puissance de calcul pour une mise en œuvre du codeur sur un processeur Texas Instruments de la famille C3x. Ces valeurs ont été mesurées sur une implémentation C30 en temps réel. Le codeur écrit en langage C a été compilé et optimisé avec les outils de Texas Instruments. Seules quelques routines, les plus gourmandes en puissance de calcul, ont été optimisées en assembleur (filtres FIR, calcul des corrélations, quantificateurs). Pour le processeur TI C3x, un mot est formé de 32 bits.

TABLE VI. — Exigences d'implémentation du HSX 1 200 (DSP TI C3x).
Légende anglaise

Exigences	Full-Duplex	Codeur	Décodeur
Mémoire programme	17.7 kw	11.7 kw	11.7 kw
Mémoire donnée	9.2 kw	8.7 kw	3 kw
Complexité	25 MIPS	19 MIPS	6 MIPS

III.4. Résultats d'évaluation

Les performances en termes d'intelligibilité du codeur HSX à 1 200 bit/s ont été établies à l'aide d'un test de rimes simplifié. Ce test a été conçu en 1983 par l'Institut de Phonétique de l'Université d'Aix-Marseille (IPAM) pour Thomson-CSF. La procédure de test est la suivante : on présente aux auditeurs une liste de 56 mots ; pour chaque mot, les auditeurs doivent identifier parmi deux propositions qui ne diffèrent que par leur consonne initiale celui qui a été prononcé. Ce test ne porte que sur deux traits acoustiques : grave et compact[12]. Une formule de régression linéaire donne le score d'intelligibilité moyen pour le français à partir des scores d'intelligibilité pour ces deux traits. L'IPAM a montré que ce test de rimes simplifié est statistiquement en accord avec le test de rimes complet pour le français (test DRT classique).

Ce test a été mis en œuvre sur 8 auditeurs avec 4 séquences de test différentes. Deux codeurs étaient considérés : le codeur HSX à 1 200 bit/s et le codeur classique LPC 10-E à 2 400 bit/s (version 52). Deux séquences de test différentes traitées par deux codeurs différents étaient présentées à chacun des auditeurs. Les résultats des tests présentés dans le tableau VII montrent que le taux d'intelligibilité du codeur HSX à 1 200 bit/s est en moyenne 2.5 points supérieur à celui du codeur LPC10-E au double du débit.

TABLE VII. — Scores d'intelligibilité des codeurs.

Légende anglaise

Codeur	Score	Écart Type
HSX à 1 200 bit/s	96.51	1.07
LPC10-E à 2 400 bit/s	94.04	1.91

Nous n'avons pas conduit d'évaluation formelle de la qualité du codeur HSX à 1 200 bit/s. Toutefois, tous les auditeurs ont convenu que la parole reproduite par ce codeur était plus naturelle que celle produite par le codeur LPC10-E, même en présence de locuteurs rapides.

Nous avons également évalué de façon informelle les performances du codeur HSX à 1 200 bit/s dans de nom-

10. Le C-ETSI est un ensemble d'opérateurs de base (multiplication, addition, multiplication-accumulation, décalage,…) simulant une unité de calcul en virgule fixe. Ces opérateurs ont été définis à l'origine pour la norme ETSI TETRA.
11. kw = kilo-words = kilo-mots.
12. Ces traits sont 2 des 7 caractéristiques utilisées en « analyse binaire en traits acoustiques du système consonnantique du français ». Ces 7 traits étant les suivants : nasal, vocalique, interrompu, continu, compact, aigu, voisé. Chaque consonne du français possède ou ne possède pas chacune de ces caractéristiques ; une paire minimale est un ensemble de 2 consonnes qui ne diffèrent que par un trait acoustique.

breuses conditions opérationnelles de prise de sons et de transmission. Nous avons constaté que le codeur HSX est naturellement robuste au bruit de fond. La qualité de la parole reproduite reste tout à fait acceptable, avec relativement peu d'artefacts notables, pour un taux d'erreurs binaires jusqu'à 1 %. Le codeur est également résistant aux pertes de trames. De plus, il comporte une procédure de récupération de trames effacées par extrapolation qui peut être mise en œuvre lorsque le décodage de canal échoue et délivre un indicateur de trame effacée.

III.5. Conclusion

Nous avons décrit dans ce chapitre un codeur HSX à 1 200 bit/s qui a été sélectionné récemment pour une application de radio messagerie vocale en Amérique du nord (« pager » vocal MobiDarc® de la société Info Télécom). Nous avons montré que ce codeur est très intelligible, présente une qualité acceptable même pour des applications grand public, est robuste au bruit de fond et aux erreurs de transmission, et présente une complexité raisonnable. Il est donc idéalement taillé pour toute autre application nécessitant le codage de la parole à bas débit.

Il existe maintenant toute une famille de codeurs HSX à différents débits (notamment à 2 400 bit/s et 3 200 bit/s). La version à 2 400 bit/s présente des performances en termes d'intelligibilité et de qualité comparables à celles du codeur CELP à 4800 bit/s (standard américain FS1016).

Le problème de la sensibilité au bruit de fond des codeurs paramétriques est bien connu. Pour remédier à ce problème dans des environnements très fortement bruités, une procédure de réduction du bruit de fond optimisée pour le codeur HSX a été développée [23].

IV. CODEUR À TRÈS BAS DÉBIT ALISP

Ce codeur a été développé dans le cadre de la thèse[13] de Jan Cernock_ [8] portant sur la recherche non-supervisée d'unités pour le traitement automatique de la parole. L'approche utilisée s'inscrit dans le domaine des codeurs segmentaux « pseudo-phonétiques » décrits dans la sous-section 2.4. Son schéma général (Figure 11) montre que le noyau du codeur est un système de reconnaissance, qui sectionne la parole d'entrée en une chaîne de segments, et qui attribue à chaque segment l'unité de codage (UC) qui lui est la plus proche. Comme ces unités sont représentées par des *modèles*, nous avons également défini des unités de synthèse (US) et des représentants - ceux-ci, choisis dans le corpus d'apprentissage, servent à la synthèse de la parole dans le décodeur. Trois types d'information transitent alors du codeur au décodeur : une chaîne d'indices des UC, une information sur les représentants, et une information supplémentaire sur la prosodie (pitch, énergie, voisement).

Contrairement à Ribeiro et Trancoso [50], Ismail et Ponting [24] et Tokuda et al. [63], qui utilisent à l'unanimité des *phonèmes* pour les unités de base, nous avons tenté de minimiser l'intervention humaine lors de l'apprentissage du codeur. La transcription phonétique [20] des bases de données (BD) de la parole est l'étape la plus coûteuse et la plus sujette aux erreurs humaines. Notre

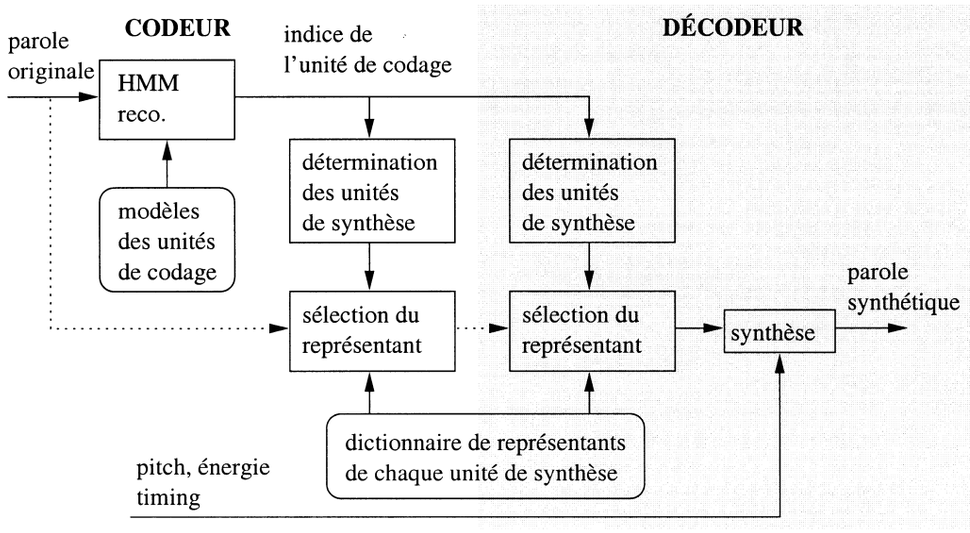

FIG. 11. — Codage et décodage de la parole: unités de codage, unités de synthèse et les représentants.

Légende anglaise

13. Ce travail a été partiellement financé par le Ministère de l'Éducation de la République Tchèque, sous le projet n° VS97060.

schéma fait appel à des techniques regroupées sous le nom générique ALISP (Traitement Automatique de la Parole, Indépendant de la Langue, Automatic, Language Independent Speech Processing) [11]. Ces techniques se basent sur les *données* et tentent de limiter au minimum les connaissances a priori nécessaires. Recherchant un équilibre entre la précision de la description et son économie, ces techniques détectent des régularités dans le signal (ou sa paramétrisation) pour en faire émerger sa structure.

IV.1. Recherche des unités dans un corpus d'apprentissage

Sur un corpus de parole donné, la détermination des unités s'effectue en deux étapes principales : dans la première, nous définissons le jeu d'unités et nous recherchons une segmentation initiale du corpus. Dans la deuxième, ces unités sont modélisées par des modèles stochastiques. Le système est ainsi *appris* et peut traiter un signal de parole inconnu.

Nous appelons les techniques utilisées pour cette extraction et cette modélisation des « outils » (voir la chaîne de traitement Figure 12). Certains parmi eux sont utilisés largement en traitement de la parole (paramétrisation, modèles de Markov cachés), les autres (décomposition temporelle, multigrammes) sont plus spécifiques aux approches ALISP. Ces outils sont hautement modulaires, et la position de certains d'entre eux dans la chaîne de traitement peut changer (c'est le cas pour les multigrammes).

4.1.1. Décomposition temporelle

Après une paramétrisation LPC-cepstrale classique sur des trames de longueur fixe, on applique la *décomposition temporelle (DT)* sur la matrice de coefficients LPCC[14]. La DT, introduite par Atal [2] et perfectionnée par Bimbot [4], approche une telle matrice par des vecteurs-cibles et des fonctions d'interpolation (FI). Techniquement, la recherche des cibles et des fonctions d'interpolation de la DT se fait par une *décomposition en valeurs singulières (SVD)* à court terme d'une sous-matrice Y de la matrice des coefficients cepstraux X.

$$Y^T = U^T D V$$

On assemble ensuite les lignes de la matrice U pour trouver une FI concentrée sur une fenêtre rectangulaire. La ré-estimation de la FI et l'adaptation de la fenêtre sont itérées pour obtenir une compacité maximale de la FI. Le post-traitement des FI contient un lissage, une dé-corrélation, et une normalisation. Dans l'étape suivante, le calcul des cibles est effectué en utilisant la pseudo-inverse : $A = X\Phi^{\#}$. Enfin, les cibles et FIs sont affinées localement.

Les FIs, déterminant ainsi des parties quasi-stationnaires du signal, définissent une première *segmentation* de la parole.

IV.1.2. Quantification vectorielle

Les segments trouvés subissent une classification non-supervisée. Il existe plusieurs méthodes [18], de recherche de classes en fonction de la proximité des vecteurs de paramètres dans un espace à P dimensions : Quantification Vectorielle (QV), Modèles de Markov

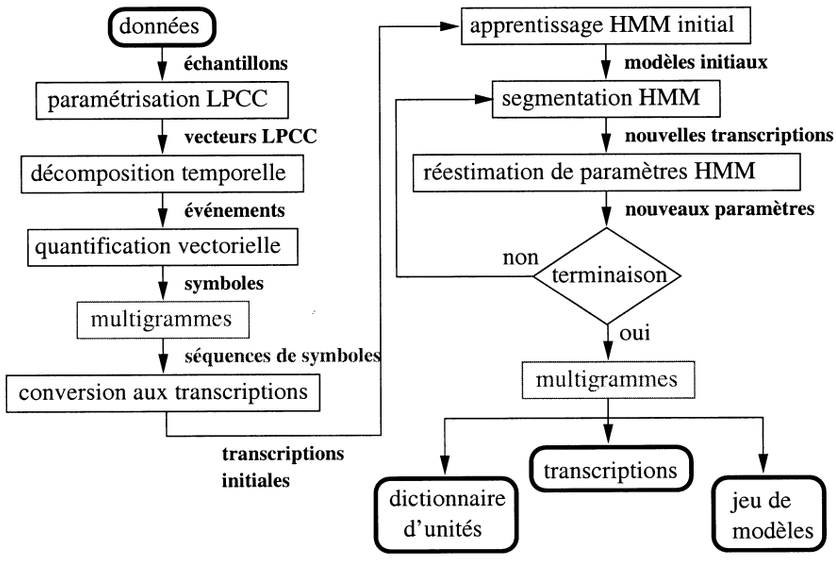

FIG. 12. — Outils utilisés dans la recherche des unités pour le traitement de la parole.

Légende anglaise

14. LPCC : Linear Prediction Cepstral Coefficient.

cachés ergodiques (EHMM), Self-Organizing Maps (SOM) de Kohonnen, et autres. Nous avons choisi la QV pour sa simplicité : les segments sont représentés par un dictionnaire de vecteurs-codes (nous allons utiliser le terme anglais *codebook* dans la suite, pour ne pas confondre ce dictionnaire avec le dictionnaire des unités ALISP) : $Y=\{y_i, 1\le i\le L\}$, où L est le nombre de classes. Ce dictionnaire est appris par l'algorithme LBG [35] avec des éclatements successifs du dictionnaire : *L = 1, 2, 4,…* L'ensemble d'apprentissage est constitué des vecteurs cepstraux originaux situés aux centres de gravité des FI.

Une fois le dictionnaire appris, nous pouvons procéder à une *quantification* : dans cette étape, on attribue à chaque événement de la DT le numéro (étiquette) de la classe qui lui est la plus proche. Pour cette quantification, nous avons utilisé tous les vecteurs d'un segment prédéterminé par la DT en utilisant une distance cumulée : $d_c[X_n, Y]=\sum_{t=bb_n}^{ee_n} d(x(t), y)$ où bb_n et ee_n sont respectivement le début et la fin du n^{ème} segment, les vecteurs $x(t)$ sont les vecteurs à coder et y est le vecteur-code.

La décomposition temporelle avec la quantification vectorielle effectuent ainsi une *transcription initiale* (bornes temporelles et labels) de la base de données de parole.

IV.1.3. Multigrammes

Il se peut que nous ayons besoin d'unités plus longues que celles déterminées par une combinaison DT+QV. Bien que nous travaillions avec des unités déterminées automatiquement, nous pouvons nous approcher ainsi des techniques syllabiques ou diphoniques utilisées dans les traitements classiques. Ce séquencement a de nombreux avantages : en codage par exemple, nous pouvons ainsi limiter le débit binaire (le dictionnaire d'unités devient plus grand, mais le nombre d'unités à transmettre par seconde décroît) et nous pouvons de plus, en limitant ainsi le nombre de transitions entre unités, atténuer les effets indésirables dus à la concaténation de segments courts. On appelle *« multigramme » (MG)* une séquence formée d'un nombre variable de symboles, et *n*-multigrammes les MG, dont la longueur est limitée à *n*. La technique utilisée pour ce séquencement est appelée décomposition en multigrammes [15]. Cette méthode, dont nous connaissons plusieurs variantes – discrètes ou continues – permet de détecter des *séquences caractéristiques* d'unités dans le corpus d'apprentissage.

Nous supposons que les événements de la DT ont déjà été étiquetés par la QV (nous avons donc une chaîne de symboles – Figure 12). Pour un dictionnaire de multigrammes $\{x_i\}$ donné, la segmentation d'une chaîne d'observations discrètes et sa transcription en multigrammes se font en maximisant la vraisemblance de la segmentation et de l'étiquetage :

$$(S^*, X^*)=\arg\max_{\forall(S,X)} L(O, S, X \mid \{x_i\}), \quad (4)$$

où O est la chaîne d'observations, S est sa segmentation et X l'attribution des multigrammes. Pour les MGs discrets, le dictionnaire contient les différentes séquences appelées multigrammes x_i ainsi que leurs probabilités π_i. Nous pouvons écrire :

$$s_i = x_i \text{ et } L(O, X \mid \{x_i\}) = P(x_{i_1})\,P(x_{i_2})\ldots P(x_{i_q}).$$

Le dictionnaire de MG n'est pas connu a priori et doit être appris sur une base de données de symboles. Cet apprentissage commence par une *initialisation*. On initialise les valeurs des probabilités π_i de toutes les séquences possibles de longueur 1 à n par le nombre d'occurrences de ces séquences dans la base de données d'apprentissage. Après cette initialisation, on réitère plusieurs étapes de segmentation au sens du maximum de vraisemblance (Éq. 4). À l'étape n, on effectue la segmentation en utilisant le dictionnaire déterminé à l'étape $n-1$, puis on met à jour les probabilités π_i des multigrammes à partir de la nouvelle segmentation. Durant ces itérations, le dictionnaire est élagué des MGs rares en imposant un nombre d'occurrences minimal.

On peut appliquer la méthode des multigrammes à 2 niveaux différents de la chaîne de traitement. On peut l'utiliser sur :

- *Les événements de la DT quantifiés par QV*. Les MG servent ici à initialiser des HMM (voir la sous-section suivante) avec des nombres d'états variables.
- *Les symboles générés par une segmentation par les HMM*. Les MG aident ici à la création d'unités plus longues.

Les MG constituent ainsi un module dont la position peut varier dans le schéma de la Figure 11.

IV.1.4. Modèles de Markov cachés

Dans la deuxième étape de traitement, les unités trouvées par la combinaison DT+QV ou DT+QV+MG sont *modélisées* par les *Modèles de Markov Cachés (HMM)* [49,65]. Ce formalisme, utilisé largement en reconnaissance de parole, ne sert pas seulement à produire des modèles, mais contribue lui-même à un affinement du jeu

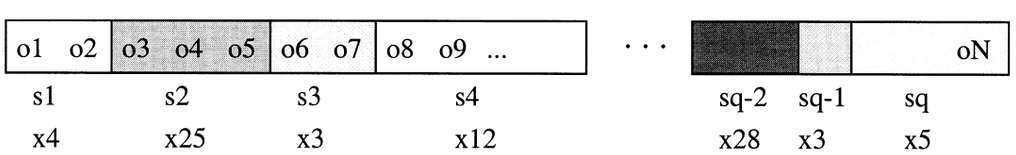

FIG. 13. — Séquencement des symboles par les multigrammes.

Légende anglaise

d'unités par des itérations de segmentation du corpus (un alignement des HMM avec les données) et de ré-estimation des paramètres des modèles.

La reconnaissance de parole à l'aide des HMM est basée sur la maximisation de la vraisemblance de l'observation et des modèles:

$$\arg\max_{\{M_1^N\}} L(O|\, M_1^N)\, L(M_1^N),$$

Où O est une chaîne d'observations (vectorielles cette fois-ci), et M_1^N une séquence de modèles. La vraisemblance $L(O|\, M_1^N)$ dite « acoustique » quantifie la correspondance entre les données et les modèles, quant à la vraisemblance $L(M_1^N)$ (modèle de langage), elle donne une plausibilité a priori de la séquence de modèles M_1^N.

Un choix important est celui de *l'architecture* des HMM. Nous avons choisi l'architecture la plus simple gauche-droite. Le *nombre de modèles* est déterminé par la taille L du dictionnaire de quantification vectorielle ou par la taille Z du dictionnaire des MG. Le nombre d'états-émetteurs des HMM est défini comme $2i+1$, où i est le nombre d'unités dans un multigramme. Au cas, où l'on ne travaille pas avec les MG, ce nombre est $2 \times 1 + 1 = 3$. Dans la plupart de nos travaux, la notion du modèle de langage n'a pas été utilisée et nous avons attribué la même probabilité a priori à tous les modèles.

L'apprentissage des HMM se fait sur le même corpus que celui utilisé pour apprendre la DT et la QV. *L'initialisation* des HMM prend en compte les transcriptions initiales T^0 obtenues par la combinaison DT+QV ou DT+QV+MG originales. Les modèles sont appris sans contexte et en contexte (apprentissage itéré) [65] pour aboutir à un jeu de paramètres initiaux Λ^0 :

$$\Lambda^0 = \{\lambda_i^0\} = \arg\max_{\forall\Lambda} L(O, \Lambda \,|\, T^0)$$

On répète ensuite, les étapes de segmentation à l'aide des modèles préalablement appris et de ré-estimation des paramètres de ces modèles :

– *Segmentation :* $T^{m+1} \arg\max_{M_1^N} L(O, M_1^N \,|\, \Lambda^m, LM^m)$
où T^{m+1} représente les nouvelles transcriptions obtenues à l'aide des anciens paramètres des modèles Λ^m et de l'ancien modèle de langage LM^m.

– *Ré-estimation des paramètres HMM* : $\Lambda^{m+1} = \arg\max_{\Lambda} L(O, \Lambda \,|\, T^{m+1})$ où Λ^{m+1} représente les nouveaux paramètres des modèles obtenus par une ré-estimation avec les transcriptions T^{m+1}. Dans cette étape, on peut aussi ré-estimer le modèle de langage LM^{m+1}.

– **Terminaison :** on arrête si l'augmentation de la vraisemblance n'est plus significative, ou si le nombre d'itérations est plus grand qu'un seuil donné. Sinon, retour à la segmentation.

Nous avons trouvé que l'utilisation de cette technique d'affinement améliore la cohérence des modèles avec les données (au sens d'une augmentation de la vraisemblance) ainsi que la cohérence des segments acoustiques dans des différentes classes (la ressemblance des segments dans une classe devient meilleure).

Les techniques utilisées fournissent donc 3 types de résultats : un *dictionnaire d'unités*, déterminé sur le corpus d'apprentissage, une *transcription* du corpus d'apprentissage utilisant ces unités et un *jeu de modèles HMM*.

IV.2. Expériences – Boston University Radio Speech Corpus

Nous avons effectué plusieurs jeux d'expériences en mode dépendant du locuteur en français [8], anglais américain [8] et tchèque [9]. Nous allons présenter ici les résultats obtenus sur la base de données américaine « Boston University Radio Speech Corpus ». Les données de ce corpus distribué par LDC[15] sont de qualité « Hi-Fi » (fréquence d'échantillonnage 16 kHz). Le corpus contient la parole de 7 présentateurs professionnels de la station WBUR. Nous avons utilisé les données d'un locuteur masculin (M2B) – 78 minutes et un locuteur féminin (F2B) – 83 minutes. Selon la provenance des enregistrements, les données ont été divisées en un corpus d'apprentissage (celles enregistrées de la radio) et de test (données enregistrées au studio de Boston University).

Nous avons effectué une paramétrisation avec 16 coefficients LPC-cepstraux en trames de 20 ms (recouvrement 10 ms). La soustraction de la moyenne cepstrale (CMS) a été faite pour chaque appel. Nous avons ensuite appliqué la DT, ajustée afin de produire 15 cibles par seconde en moyenne. Sur les segments obtenus, nous avons appris un dictionnaire de QV à 64 vecteurs-codes. Les HMM étaient appris directement sur les transcriptions DT+QV (sans pré-traitement par les multigrammes). Leur nombre réduit (64) a permis un affinement avec 5 itérations de segmentation et de ré-estimation. Nous avons vérifié que la vraisemblance d'alignement des données avec les modèles augmentait. Nous avons ensuite testé une application des MG sur la dernière segmentation HMM, et nous avons obtenu des dictionnaires de séquences de longueur variable 1 à 6, de tailles 722 pour le locuteur féminin et 972 pour le sujet masculin.

Pour le décodage, nous avons utilisé des unités de synthèse équivalentes à celles de codage, et nous avons disposé de 8 représentants pour chacune. Ici, nous avons testé une synthèse LPC et nous n'avons pas considéré le codage de la prosodie, les contours de F_0 et de l'énergie originaux étant introduits directement dans le synthétiseur.

15. Linguistic Data Consortium – University of Pennsylvania, http://www.ldc.upenn.edu/ $R_u = \dfrac{\log_2(Z)\sum_{i=1}^{Z} c(M_i)}{T_f \sum_{i=1}^{Z} c(M_i) l(M_i)}$

Dans l'évaluation du *débit binaire* nécessaire pour la transmission de l'information sur les unités, nous n'avons pas considéré les probabilités a priori des unités (codage entropique [18]), mais nous avons calculé le nombre de bits nécessaire pour la transmission de chaque unité M_i par $\log_2(Z)$, où Z est la taille du dictionnaire. Le débit binaire moyen est ainsi défini :

où $c(M_i)$ est le nombre d'occurrences de M_i dans la chaîne encodée, $l(M_i)$ est la longueur de M_i et T_f est le décalage entre les trames acoustiques en secondes. La *qualité* de la parole après codage-décodage a été évaluée subjectivement par des tests informels.

Les débits binaires obtenus sont donnés dans le tableau VIII. En évaluant la qualité de la parole obtenue, nous l'avons jugée intelligible, avec une meilleure qualité pour les multigrammes (moins de distorsions sur les transitions).

TABLE VIII. — Débits binaires obtenus sur le « BU radio speech corpus » (seulement pour le codage des unités et incluant les 3 bits nécessaires pour le codage du choix de représentant).

Légende anglaise

locuteur	F2B		M2B	
débit binaire sur : l'ensemble de	appren-tissage	test	appren-tissage	test
HMM 6-ème génération	189.27	190.28	189.75	195.51
HMM 6-ème génération + MG	135.91	145.09	141.86	156.02

IV.3. Codage ALISP – Conclusions

L'application des unités « ALISP » dans le codage à très bas débit nous a permis d'obtenir un signal de parole intelligible avec des débits moyens de 120 bit/s pour le codage des unités acoustiques, sans avoir à faire appel à une base de données transcrite. De nombreuses améliorations restent à apporter au codeur : nous avons choisi une méthode de synthèse par concaténation très rudimentaire, qui devrait être remplacée par un schéma de meilleure qualité (PSOLA, HNM). La détermination des unités acoustiques à utiliser en synthèse et le lissage à leurs bornes n'ont pas été entièrement résolus. Pour l'application dans des systèmes réels, l'algorithme proposé doit être complété par une adaptation au locuteur et éventuellement par un module de modification de la voix dans le synthétiseur.

V. CONCLUSIONS

Nous avons tenté de rendre compte de l'état de l'art en codage de la parole à bas et très bas débit. Plus le débit de codage diminue, plus le délai introduit par le codeur augmente. On peut difficilement concevoir un codeur de parole à 100 bit/s pour une communication en full-duplex. Le codeur HSX a 1 200 bit/s introduit un délai de 240 ms, ce qui semble être la limite tolérable pour une application grand public. Les codeurs segmentaux peuvent introduire des délais 2 à 10 fois plus élevés. Leur domaine d'application est plutôt la diffusion de messages et la restitution de parole. Le codage segmental par indexation d'unités est un domaine prometteur qui trouve aussi des applications en reconnaissance et synthèse de la parole, en vérification du locuteur et en identification de la langue [12].

Manuscrit reçu le 14 janvier 2000
accepté le 12 juin 2000

BIBLIOGRAPHIE

[1] ATAL (B.S.), HANAUER (S.L.) Speech Analysis and Synthesis by linear Prediction of the speech Wave, J. *Acoust. Soc. Amer.,* **50** n° 2 p. 637-657, 1971.

[2] ATAL (B.S.), Efficient coding of LPC parameters by temporal decomposition, In *Proceedings IEEE ICASSP* **83**, pp. 1-84, 1983.

[3] BAUDOIN (G.), CERNOCK (J.), CHOLLET (G.), Quantization of spectral sequences using variable length spectral segments for speech coding at very low bit rate, *Proceedings Eurospeech-97*, pp. 1295-1298, Rhodes, 1997.

[4] BIMBOT (F.) An evaluation of temporal decomposition, Technical report, Acoustic research department AT&T Bell Labs, 1990.

[5] BRUHN (S.), Matrix Product Vector Quantization for Very Low bit Rate Speech Coding, *Proceedings ICASSP-95*, p. 724-727, 1995.

[6] CERNOCK (J.), BAUDOIN (G.), CHOLLET (G.), Segmental vocoder - going beyond the phonetic approach, *Proceedings ICASSP98*, pp. 605-608, Seattle, 1998.

[7] CERNOCK (J.), BAUDOIN (G.) and CHOLLET (G.) The use of ALISP for automatic acoustic-phonetic transcription, *Proceedings SPoS-ESCA Workshop on Sound Patterns of Spontaneous Speech*, pp. 149-152, Aix en Provence, 1998.

[8] CERNOCK (J.), *Speech Processing Using Automatically Derived Segmental Units: Applications to Very Low Rate Coding and Speaker Verification*, PhD thesis, Université Paris XI Orsay, 1998.

[9] CERNOCK (J.), I. KOPECEK, BAUDOIN (G.), and CHOLLET (G.), Very low bit rate speech coding: comparison of data-driven units with syllable segments, In *Proceedings of Workshop on Text Speech and Dialogue (TSD'99)*, Lecture notes in computer science, Mariánské Lázne, Czech Republic, September 1999. Springer Verlag.

[10] CHENG (Y.M.), O'SHAUGHNESSY (D.), A 450 BPS Vocoder with natural sounding Speech. *Proceedings ICASSP-90*, p. 649-652, 1990.

[11] CHOLLET (G.), CERNOCK (J.), CONSTANTINESCU, DELIGNE (S.), and BIMBOT (F.). *Computational models of speech pattern processing*, chapter Towards ALISP: a proposal for Automatic Language Independent Speech Processing, pp. 375-388. NATO ASI Series. Springer Verlag, 1999.

[12] CHOLLET (G.), CERNOCK (J.), GRAVIER (G.), HENNEBERT (J.), PETROVSKA (D.), YVON (F.), Toward Fully Automatic Speech Processing Techniques for Interactive Voice Servers, in *Speech Processing, Recognition and Artificial Neural Networks*, CHOLLET (G.), BENEDETTO (M-G), ESPOSITO (A.), MARINO (M.) eds, Springer Verlag, 1999.

[13] CHOU (P.A.), LOOKABAUGH (T.), Variable dimension vector quantization of linear predictive coefficients of speech. *Proceedings ICASSP-94*. pp. I-505-508, Adélaïde, 1994.

[14] CROSMER (J.R.), BARNWELL (T.P.), A Low Bit Rate Segment Vocoder Based on Line Spectrum Pairs, *Proceedings ICASSP-85* pp. 240-243, 1985.

[15] DELIGNE (S.), *Modèles de séquences de longueurs variables: Application au traitement du langage écrit et de la parole*, PhD thesis, École nationale supérieure des télécommunications (ENST), Paris, 1996.

[16] FETTE (B.), JASKIE (C.), A 600 bps LPC Voice Coder, *Proceedings MILCOM-91*, pp. 1215-1219, 91.

[17] FLANAGAN (J.-L.), Springer Verlag. *Speech Analysis, Synthesis and Perception* New York, 1965, 2nd ed. 1972.

[18] GERSHO (A.), *Vector Quantization and Signal Compression* Kluwer Academic Publisher 1996.

[19] GERSHO (A.), Advances in speech and audio compression, *Proceedings IEEE*, 82(6):900-918, June 1994.

[20] GIBBON (D.), MOORE (R.), and WINSKI (R.), editors, *EAGLES Handbook on Spoken Language Systems*, Mouton de Gruyter, 1997.

[21] GOURNAY (P.), CHARTIER (F.), A 1200 bps HSX speech coder for very low bit rate communications, *IEEE Workshop on Signal Processing System SiPS'98*, Boston, 1998.

[22] GRIFFIN (D.W.) and LIM (J.S.), « Multiband Excitation Vocoders » *IEEE Trans. on Acoustics, Speech, and Signal Processing*, **36**, n° 8, pp. 1223-1235, 1988.

[23] GUILMIN (G.), LE BOUQUIN–JEANNÈS (R.) et GOURNAY (P.), Study of the influence of noise pre-processing on the performance of a low bit rate parametric speech coder, *Eurospeech'99*, **5**, pp. 2367-2370, Budapest 1999.

[24] ISMAIL (M.) and PONTING (K.), Between recognition and synthesis 300 bps speech coding. In *Proceedings Eurospeech-97*, pp. 441-444, Rhodos, 1997.

[25] ISO/IEC JTC1/SC29/WG11 N2503-sub2, « Final Draft International Standard of ISO/IEC 14496-3 Subpart 2 », octobre 1998.

[26] JASKIE (C.), FETTE (B.), A survey of low bit rate vocoders. *DSP & Multimedia Technology*, p 26-40, apr. 94.

[27] JEANRENAUD (P.), PETERSON (P.), Segment Vocoder Based on Reconstruction with Natural Segment *Proceedings ICASSP-91*, pp. 605-608, 1991.

[28] JELINEK (M.), BAUDOIN (G.), Excitation Construction for the robust CELP coder, In *Speech Recognition and Coding, new advances and trends*, Springer Verlag, NATO ASI Serie F., Ed. par A. Rubio & J.-M. Lopez, pp. 439-443, 1995.

[29] KANG (G.S.), FRANSEN (I.J.), Application of Line Spectrum Pairs to Low-Bit Rate Speech Encoders, *Proceedings ICASSP-85*. pp. 244-247, 85.

[30] KEMP (D.P.), COLLURA (J.S.), TREMAIN (T.E.), Multiframe Coding of LPC Parameters at 600-800 bps, *Proceedings ICASSP-91*, pp. 609-612, 91.

[31] KLEIJN (W.) Encoding Speech Using Prototype Waveforms. *IEEE Trans. Speech Audio Processing*, **1**, n° 4, pp. 386-399, 1993.

[32] KLEIJN (W.B.), HAAGEN (J.), A Speech Coder based on Decomposition of Characteristic Waveforms, *Proceedings ICASSP-95*, pp. 508-511, 1995.

[33] KLEIJN (W.B.), HAAGEN (J.), « Waveform Interpolation for Coding and Synthesis », in *Speech Coding and Synthesis*, edited by KLEIJN (W.B.) and PALIWAL (K.K.), Elsevier 1995.

[34] LAFLAMME (C.), SALAMI (R.), MATMTI (R.), and ADOUL (J.-P.), « Harmonic Stochastic Excitation (HSX) speech coding below 4 kbps », *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Atlanta, May 1996, pp. 204-207.

[35] LINDE (Y.), BUZO (A.), GRAY (R.M.), Algorithm for Vector Quantization Design, *IEEE trans. on communications*, **28**, p 84-95, Jan. 1980.

[36] LIU (Y.J.), ROTHWEILER (J.), A High Quality Speech Coder at 400 BPS, *Proceedings ICASSP-89*, pp. 204-206, 1989.

[37] LOPEZ-SOLER (E.), FAVARDIN (N.), A combined quantization-Interpolation scheme for Very Low bit rate coding of speech LSP parameters, *Proceedings ICASSP-93*, p.II-21-24, 1993.

[38] MCAULAY (R.), QUATIERI (T.), Speech Analysis/Synthesis based on a sinusoïdal representation of speech. *IEEE trans. ASSP-34*, n° 4, pp. 744, 1985.

[39] MCAULAY (R.), CHAMPION (T.), Improved Interoperable 2.4 kbps LPC Using Sinusoïdal Transform Coder techniques, *Proceedings ICASSP-90*, pp. 641-643, 1990.

[40] MCAULAY (R.), QUATIERI (T.), Multirate Sinusoïdal Transform Coding at Rates from 2.4 kbps to 8kbps, *Proceedings ICASSP-87*, Dallas, 1987.

[41] MCAULAY (R.), QUATIERI (T.), Sine-Wave Phase Coding at Low Data Rates, *Proceedings ICASSP-91*, pp. 577-580, 1991.

[42] MCCREE (A.), TRUONG (K.), GEORGE (E.B.), BARNWELL (T.P.), VISWANATHAN (V.), A 2,4 Kbits/s MELP Coder Candidate for the New U.S. Federal Standard, *Proceedings ICASSP-96*, pp. 200-203, 1996.

[43] MOUY (B.), DE LA NOUE (P.) and GOUDEZEUNE (G.), « NATO STANAG 4479: A standard for an 800 bps vocoder and channel coding in HF-ECCM system », *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Detroit, May 1995, pp. 480-483.

[44] NISHIGUCHI (M.), INOUE (A.), MAEDA (Y.), MATSUMOTO (J.), Parametric Speech Coding – HVXC at 2.0-4.0 kbps, *Proc IEEE Workshop on Speech Coding*, 1999.

[45] « Parameters and coding characteristics that must be common to assure interoperability of 2400 bps linear predictive encoded speech », NATO Standard STANAG-4198-Ed1, 13 February 1984.

[46] PETERSON (P.), JEANRENAUD (P.), VANDEGRIFT (J.), Improving Intelligibility at 300bps Segment Vocoder, *Proceedings ICASSP-90*, pp. 653-656, 1990.

[47] PICONE, DODDINGTON (G.R.), A phonetic Vocoder, *Proceedings ICASSP-89*, pp. 580-583, 1989.

[48] POTAGE (J.), ROCHETTE (D.), MATHEVON (G.), Speech Encoding Techniques for Low Bit Rate Coding Applicable to Naval Communications, *Rev. Tech. Thomson-CSF*, **18**, n° 1 pp. 171-205, Mar. 86.

[49] RABINER (L.) and JUANG (B.H.) *Fundamentals of speech recognition*, Signal Processing. Prentice Hall, Engelwood Cliffs, NJ, 1993.

[50] RIBEIRO (C.) and TRANCOSO (M.), Phonetic vocoding with speaker adaptation, In *Proceedings Eurospeech-97*, pp. 1291-1294, Rhodes, 1997.

[51] ROTHWEILER (J.), Performances of a real time Low Rate Voice Coder. *Proceedings ICASSP-86*, pp. 3039-3042, 1986.

[52] ROUCOS (S.), SCHWARZ (R.), MAKHOUL (J.), A segment vocoder at 150 bps, *Proceedings ICASSP-83*, pp. 61-64, 1983.

[53] ROUCOS (S.), WILGUS (A.M.), The Waveform Segment Vocoder: A New Approach for Very Low Rate Speech Coding, *Proceedings ICASSP-85*, pp.236-239, 1985.

[54] ROUCOS (S.), SCHWARZ (R.), MAKHOUL (J.), Segment Quantization for very-low rate speech coding, *Proceedings ICASSP-82*.

[55] SCHROEDER (M.R.), ATAL (B.), Code-Excited Linear Prediction (CELP): High Quality Speech at Very Low Bit Rates, *Proceedings IEEE ICASSP-85*, pp. 937-940, Tamp, 1985.

[56] SCHWARTZ (R.M.), ROUCOS (R.M.), A Comparison of Methods for 300-400 B/S Vocoders, *Proceedings ICASSP-83*, 83.

[57] SHIRAKI (Y.), HONDA (M.), LPC speech coding based on Variable Length Segment Quantization, *IEEE trans. on ASSP, vol.36*, n° 9, pp. 1437-1444, sept. 1988, pp. 1565-1568, 82.

[58] SHOHAM (Y.), « Very low complexity interpolative speech coding at 1.2 to 2.4 kbps », *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich, April 1997, pp. 1599-1602.

[59] SPANIAS, Speech coding: A Tutorial Review, *Proceedings IEEE, 82(10)* 1541-1582, Oct. 1994.

[60] STYLIANOU (Y.), DUTOIT (T.), SCHROETER (J.), Diphone concatenation using a Harmonic plus Noise Model of Speech, *Proceedings Eurospeech-97*, Rhodes, sept. 1997.

[61] SUPPLEE (L.M.), COHN (R.P.), COLLURA (J.S.), MCCREE (A.V.), « MELP : The new federal standard at 2400 bps », *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich, April 1997, pp. 1591-1594.

[62] Specifications for the Analog to Digital Conversion of Voice by 2,400 Bit /Second Mixed Excitation Linear Prediction. *Federal Information Processing Standards Publication (FOPS PUB) Draft-May 1998*.

[63] TOKUDA (K.), MASUKO (T.), HIROI (J.), KOBAYASHI (T.), KITAMARA (T.), A very low bit rate speech coder using hmm-based speech recognition/synthesis techniques, In *Proceedings ICASSP-98*, pp. 609-612, 1998.

[64] TREMAIN (T.E.), The government standard Linear Predictive Coding Algorithm: LPC10. *Speech Technology,* **1**, n° 2, pp. 40-49, Apr. 1982.

[65] YOUNG (S.), JANSEN (J.), ODELL (J.), OLLASON (D.), WOODLAND (P.), *The HTK book*, Entropics Cambridge Research Lab., Cambridge, UK, 1996.

[66] WONG (D.Y.), JUANG (B.H.), CHENG (D.Y.), Very Low Data Rate Speech compression using LPC Vector and Matrix Quantization, *Proceedings ICASSP-83*, pp. I-65-68, 83.

[67] Le test de diagnostic par paires minimales, adaptation au français du *Diagnostic rythm test* de W.D. Voiers, *Revue d'acoustiques*, n° 27, 1973.