CODEUR CELP A EXCITATION MIXTE

P.DYMARSKI* - N.MOREAU**

* Ecole Polytechnique de Varsovie, 15 rue Nowowiejska VARSOVIE
** ENST Dépt Signal 46 rue Barrault 75634 PARIS Cedex 13

RESUME

Dans la plupart des codeurs à débit réduit (autour de 10 Kb/s) du signal de parole, on trouve deux filtrages qui permettent d'enlever les redondances respectivement à court terme et à long terme (LTP) situées dans le signal et un troisième traitement qui approxime la partie non prédictible du signal par un vecteur représentable sur le minimum de bits possible.

Cette approximation peut être réalisée de plusieurs façons : par le placement de quelques impulsions (MP), par la gestion plus ou moins sophistiquée d'un dictionnaire stochastique ou créé sur un corpus suffisant de signal réel par des algorithmes de type LBG (CELP) ou de façon moins standard par la combinaison éventuelle de techniques MP suivies d'une analyse de type CELP, etc ... Généralement les débits associés aux traitements MP, CELP, LTP sont fixés à priori ainsi que l'ordre dans lequel s'effectue l'analyse (par exemple d'abord LTP puis MP etc ...).

On montre qu'il est possible de s'abstraire de cette double contrainte en adoptant une représentation unique pour ces trois traitements et en construisant un dictionnaire "mixte" où deux parties sont fixées une fois pour toutes (dictionnaire stochastique + dictionnaire impulsionnel) et où la troisième est actualisée chaque sous-fenêtre par l'excitation du filtre de synthèse y_n (dictionnaire "prédictif"). La détermination des no et des gains des vecteurs à transmettre est effectuée en minimisant un critère construit sur l'énergie du signal perceptuel moins le signal perceptuel modélisé. Le codeur optimise à chaque itération son choix parmi les trois types de vecteurs possibles. Le traitement est alors complètement homogène quelque soit le type de l'excitation.

SUMMARY

In most speech coders proposed nowadays for rates between 4.8 and 16 Kbit/sec, we find two filters which are respectively short term and long term predictors (LTP). The non-predictable part of the signal is approximated every frame by a sequence encoded with a minimum number of bits.

This approximation could be realized with a number of now standardized techniques : the localization of impulses (MP technique) or a more or less sophisticated search through a dictionnary (CELP) ... This dictionary could be either : a purely stochastic one (define a-priori) or designed from a large corpus of residuals using clustering technique (ie : LBG).

A number of variations of these algorithms have been published (for example MP followed by CELP ...) but in all cases the bit rate associated with the MP, CELP and LTP coders is fixed a-priori and the order in which processing is performed is also fixed (for example LTP before MP ...).

This paper demonstrates that no constraint needs to be placed on the order of processing. A unique representation can be obtained for these three coders using a "mixed" dictionnary made of three parts. The first two parts (stochastic and impulse dictionaries) are constructed a-priori. The third part is modified for each frame using the excitation sequence of the preceding frames ("predictive" dictionary). The search for the vector index and the associated gain factor to be transmitted is done using a least square criteria between the original and the predicted signal (both being passed through a perceptual filter). The coder optimizes a choice between the three types of vectors (stochastic, impulse, predictive). The processing is therefore homogeneous for the different types.

I. INTRODUCTION

La plupart des codeurs à débit réduit actuels font appel à des techniques de modélisation de type paramètrique : on cherche à construire, à l'analyse, un signal synthétique le plus ressemblant possible au signal de départ. Le système générateur de ce signal synthétique est un vecteur d'excitation choisi convenablement de façon qu'il soit caractérisable par un nombre faible de bits passant au travers d'un ou plusieurs filtres dont les coefficients sont réactualisés de façon régulière étant donné la non-stationnarité du signal. Le critère de ressemblance est presque toujours un critère de type quadratique (simplicité de la résolution) adapté au système auditif par introduction d'une pondération dans le domaine fréquentiel dépendant d'un facteur dit perceptuel γ. La minimisation de ce critère relativement aux paramètres caractérisant le vecteur d'excitation et aux coefficients des filtres est difficile à réaliser simultanément (problème de déconvolution). Cette optimisation est donc faite en plusieurs étapes.

1) On réalise d'abord une analyse LPC à l'ordre P sur des fenêtres de N échantillons en minimisant l'énergie de l'erreur de prédiction à court terme :

$$E_1 = \sum_n (s_n - \sum_{i=1}^{P} a_i \, s_{n-i})^2$$

puis on filtre le vecteur $s_0 \cdots s_{N-1}$ par le filtre d'analyse A(z) pour obtenir cette erreur de prédiction $r_0 \cdots r_{N-1}$.

(Un bon compromis adopté par la norme GSM à 13 Kb/s pour le futur radio téléphone numérique européen est le suivant : N = 160, P = 8, Méthode d'autocorrélation, Algorithme de Schur, Encodage des LAR sur 36 bits, Interpolation des K_i avec ceux de la fenêtre précédente dans les filtres en treillis).

2) Les méthodes multi-impulsionnelles (MP), introduites en 1982, consistent à s'imposer un vecteur d'excitation de la forme :

$$e_n = \sum_{k=1}^{K} A_{m(k)}\, \delta_{n,m(k)}$$

et à minimiser le critère :

$$E_2 = \sum_n ((r_n - e_n) * h'_n)^2$$

$$E_2 = \sum_n (p_n - \sum_{k=1}^{K} A_{m(k)}\, \delta_{n,m(k)} * h'_n)^2$$

relativement aux paramètres inconnus : les positions et les amplitudes des impulsions.

p_n est le signal perceptuel, h'_n la réponse impulsionnelle du filtre perceptuel $1/A(z/\gamma)$.

La minimisation globale étant difficile, on réalise des itérations.

3) La plupart des codeurs utilisent aussi un prédicteur à long terme (LTP).

On peut adopter une structure directe où le filtre B(z) est placé en amont ou en aval du filtre A(z) ($B(z) = 1 - bz^{-Q}$ pour un filtre à un coefficient mais on peut généraliser à un nombre quelconque de coefficients). La détermination des deux paramètres b et Q s'effectue par minimisation de l'énergie de l'erreur de prédiction à long terme :

$$E_3 = \sum_n (r_n - b\, r_{n-Q})^2 \quad \text{(cas aval)}$$

Ce critère est la copie conforme du critère E_1 et sa minimisation s'effectue exactement comme celle du critère E_2.

La structure bouclée, largement adoptée actuellement car elle donne de bien meilleurs résultats, est une généralisation directe, au cas vectoriel, du cas scalaire étudié il y a de nombreuses années pour le MIC différentiel. Les paramètres b et Q peuvent être calculés en minimisant : directement le critère E_3, de façon théoriquement plus correcte, le critère (norme GSM) :

$$E_4 = \sum_n (r_n - b\, y_{n-Q})^2$$

de façon encore plus cohérente avec les techniques multi-impulsionnelles, le critère [1]:

$$E_5 = \sum_n ((r_n - b\, y_{n-Q}) * h'_n)^2$$

$$E_5 = \sum_n (p_n - b\, y_{n-Q} * h'_n)^2$$

4) A peu prés simultanément est apparu l'utilisation de dictionnaire stochastique ou créé sur un corpus suffisant de signal réel par des algorithmes de type LBG (CELP). Le dictionnaire est constitué par L vecteurs d'excitation $c_0^j \cdots c_{N-1}^j$ pré-determinés. Les techniques CELP consistent à determiner le no et le gain des K vecteurs minimisant :

$$E_6 = \sum_n ((r_n - \sum_{k=1}^{K} g_{j(k)}\, c_n^{j(k)}) * h'_n)^2$$

$$E_6 = \sum_n (p_n - \sum_{k=1}^{K} g_{j(k)}\, c_n^{j(k)} * h'_n)^2$$

Ici aussi, la minimisation globale étant difficile, on réalise des itérations.

5) Finalement, le schéma de principe de la plupart des codeurs à déb_it réduit est le suivant :

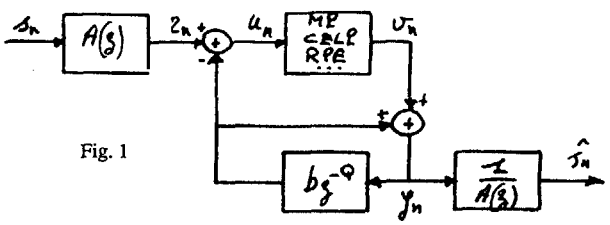

Fig. 1

Les différentes minimisations sont réalisées de façon doublement itérative : on détermine d'abord le prédicteur à court terme en minimisant le critère E_1, ensuite le prédicteur à long terme en minimisant E_4 ou E_5, enfin on réalise une approximation de la partie non prédictible du signal en minimisant le critère :

$$E_7 = \sum_n ((u_n - v_n) * h'_n)^2 = \sum_n (p_n - v_n * h'_n)^2$$

par des techniques MP ou CELP, cette minimisation étant elle-même itérative (la norme GSM approxime le vecteur résiduel de façon différente : il s'agit plutôt d'un sous échantillonnage à phase adaptative).

6) De nombreuses variantes ont été publiées récemment : plusieurs prédicteurs à long terme, plusieurs étages CELP [2], des combinaisons MP CELP [3], etc ...

II. GENERALISATION

II.1 Modélisation du perceptuel

Les critères E_2, E_5 après une 1ere généralisation (on peut prévoir la caractérisation de plusieurs prédicteurs à long terme successifs en rajoutant une sommation sur k) et E_6 ont exactement la même forme. Ils ont également la même interprétation : on cherche à modéliser le vecteur perceptuel $p_0 \cdots p_{N-1}$, dans la fenêtre d'analyse courante, comme le résultat de la convolution d'un vecteur d'excitation $y_{-\infty} \cdots y_{-1}\, y_0 \cdots y_{N-1}$ et de la réponse impulsionnelle h'_n du filtre perceptuel $1/A(z/\gamma)$. Ce modèle du perceptuel dans la fenêtre courante se décompose en deux parties :

$$\hat{p}_n = \sum_{i=-\infty}^{+\infty} y_i\, h'_{n-i} = \sum_{i=-\infty}^{-1} y_i\, h'_{n-i} + \sum_{i=0}^{n} y_i\, h'_{n-i}$$

Le 1er terme fait intervenir les valeurs déjà déterminées de y_n. Il traduit l'influence des fenêtres précédentes. Dans le 2eme terme il faut paramétrer l'expression du vecteur d'excitation $y_0 \cdots y_{N-1}$ de façon à pouvoir optimiser le critère. On le modélise comme une combinaison linéaire des vecteurs $c_0^j \cdots c_{N-1}^j$ d'un dictionnaire que l'on précisera ultérieurement :

$$y_n = \sum_{k=1}^{K} g_{j(k)}\, c_n^{j(k)}$$

où K est l'ordre de modélisation du perceptuel.

On peut donc écrire :

$$\hat{p}_n = \sum_{i=-\infty}^{-1} y_i\, h'_{n-i} + \sum_{k=1}^{K} g_{j(k)} \sum_{i=0}^{n} c_i^{j(k)}\, h'_{n-i}$$

$$\hat{p}_n = \sum_{i=-\infty}^{-1} y_i\, h'_{n-i} + \sum_{k=1}^{K} g_{j(k)}\, f_n^{j(k)}$$

Le vecteur $f_0^j \cdots f_{N-1}^j$ est simplement le vecteur $c_0^j \cdots c_{N-1}^j$ filtré par le filtre perceptuel partant de conditions initiales nulles. Le critère à minimiser s'écrit :

$$E = \sum_n (p_n - \hat{p}_n)^2$$

$$E = \sum_n ((p_n - \sum_{i=-\infty}^{-1} y_i\, h'_{n-i}) - \sum_{k=1}^{K} g_{j(k)}\, f_n^{j(k)})^2$$

où les paramètres à optimiser sont le no du vecteur j et le gain correspondant g_j. La minimisation globale étant difficile, on effectue des itérations suivant la procédure habituelle en enlevant la contribution de la modélisation à l'ordre k-1 au signal perceptuel avant de procéder à la modélisation d'ordre k. Le terme traduisant l'influence des fenêtres précédentes s'interprète comme une modélisation d'ordre 0.

II.2 Construction du dictionnaire

Le dictionnaire des vecteurs d'excitation se décompose en plusieurs parties :

- le dictionnaire stochastique constitué de séquences gaussiennes ou de séquences calculées,
- le dictionnaire multi-impulsionnel directement construit par : $c_n^j = \delta_{n,j}$ pour j = 0 ... N-1,
- le dictionnaire "prédictif" utilisant les vecteurs d'excitation passés : $c_n^j = y_{n-j}$ pour j = N ... Q_{max}.

Le dictionnaire prédictif doit être actualisé chaque fenêtre d'analyse alors que les deux précédents sont pré-déterminés.

De même que l'on subdivise la fenêtre d'analyse LPC en N/N' sous-fenêtres lorsque l'on utilise un prédicteur à long terme, de même ici on utilise un dictionnaire dont les vecteurs peuvent avoir une dimension inférieure à N. On peut même avoir des dimensions distinctes pour chacune des parties, notées N^C, N^M et N^P. On introduit donc des fenêtres d'analyse, sous-fenêtres, sous-sous-fenêtres, ...

La 1ere partie du traitement consiste donc, à partir de ce dictionnaire d'excitation, à calculer le dictionnaire filtré et l'énergie F_j de chaque vecteur filtré. Pour pouvoir regarder l'influence de la fenêtre (sous-fenêtre) courante sur la fenêtre (sous-fenêtre) suivante, il faut prolonger les vecteurs : $f^j_0 \cdots f^j_{N-1} \cdots f^j_{N-M-2}$ où M est l'indice à partir duquel la réponse impulsionnelle du filtre perceptuel peut être considérée comme nulle.

II.3 Gestion des index

On peut considérer le dictionnaire des vecteurs d'excitation comme composé de parties distinctes et s'imposer la recherche du vecteur d'excitation optimal dans chacune de ces parties successivement. Il est plus logique d'imaginer les différentes parties mises bout à bout et d'utiliser le dictionnaire globalement sans contraintes en laissant le codeur choisir. Le traitement est alors complètement homogène quelque soit le type de l'excitation.

On obtient le schéma de principe du codeur (cf Fig. 2).

II.4 Remarques

1) Cette forme permet de retrouver la grande majorité des codeurs actuels en s'imposant des contraintes sur l'index j en fonction de l'indice k : MP, CELP, CELP à plusieurs étages, LTP puis CELP, LTP puis MP puis CELP, ...

De façon moins standard, on obtient le codeur "Self Excited" lorsque le nombre de vecteurs prédictifs sélectionnés est supérieur au nombre des vecteurs gaussiens.

2) On peut prolonger le dictionnaire prédictif vers les faibles valeurs de Q en périodisant les échantillons disponibles de l'excitation passée selon le principe proposé dans [4].

3) On peut rajouter d'autres parties à ce dictionnaire : salves d'impulsions, modèle de l'onde glottale, ...

III. COMPORTEMENT DU CODEUR

III.1 Définition du corpus de test

On a utilisé quatre phrases phonétiquement équilibrées, prononcées par deux locuteurs masculins et deux locuteurs féminins et enregistrées dans des conditions semi-réelles (légères saturations, léger bruit ambiant). L'ensemble représente environ 80000 échantillons représentés sur 12 bits.

III.2 Conditions de simulation

Le choix des paramètres est le suivant : N = 160, $N^C = N^M = N^P = 40$, P = 8, $\gamma = 0.875$, M = 20, $Q_{min} = 20$, $Q_{max} = 167$, L = 32, pas de pré-actentuation.

On ne se pose pas ici le problème de l'encodage des coefficients du filtre A(z). Toutes les simulations ont été faites en utilisant un encodage sur 27 bits [1] ou en recopiant l'encodage proposé par la norme GSM sur 36 bits, les différences n'étant pas significatives dans cette étude. Toutes les simulations sont réalisées en virgule flottante.

III.3 Histogrammes des index et des gains

La figure 3 donne l'histogramme des index choisis pour modéliser le perceptuel au 1er ordre puis au 2eme. On remarque la sur-représentation du dictionnaire prédictif lors du premier choix et la relative uniformité lors du deuxième choix. On remarque également un pic accusé aux alentours de Q = 40 (voix féminines), plusieurs pics voisins de Q = 70 (voix masculines) et l'utilisation non négligeable du dictionnaire prédictif "périodisé" (pour Q < 40) ce qui justifie à postériori le choix de $Q_{min} = 20$.

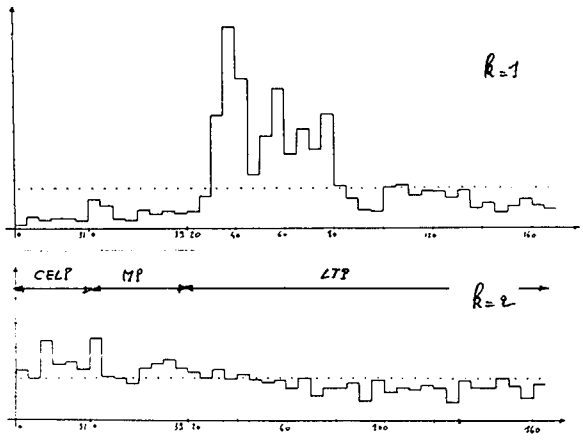

Fig. 3

Pour définir la procédure de quantification des gains, on trace l'histogramme des gains correspondant à k = 1 (en distinguant lorsque l'index appartient au dictionnaire prédictif ou pas) et pour k = 2 et 3 (Fig. 4). Ceci suggère de réaliser une quantification régulière entre -1.0 et +2.0 lorsque k = 1 et d'utiliser la caractéristique de compression "7 segments" entre -1.0 et +1.0 puis de faire une quantification régulière lorsque k > 1. Le pas de quantification dépend du débit visé. On a choisi d'utiliser 32 niveaux quelque soit l'ordre de modélisation du perceptuel.

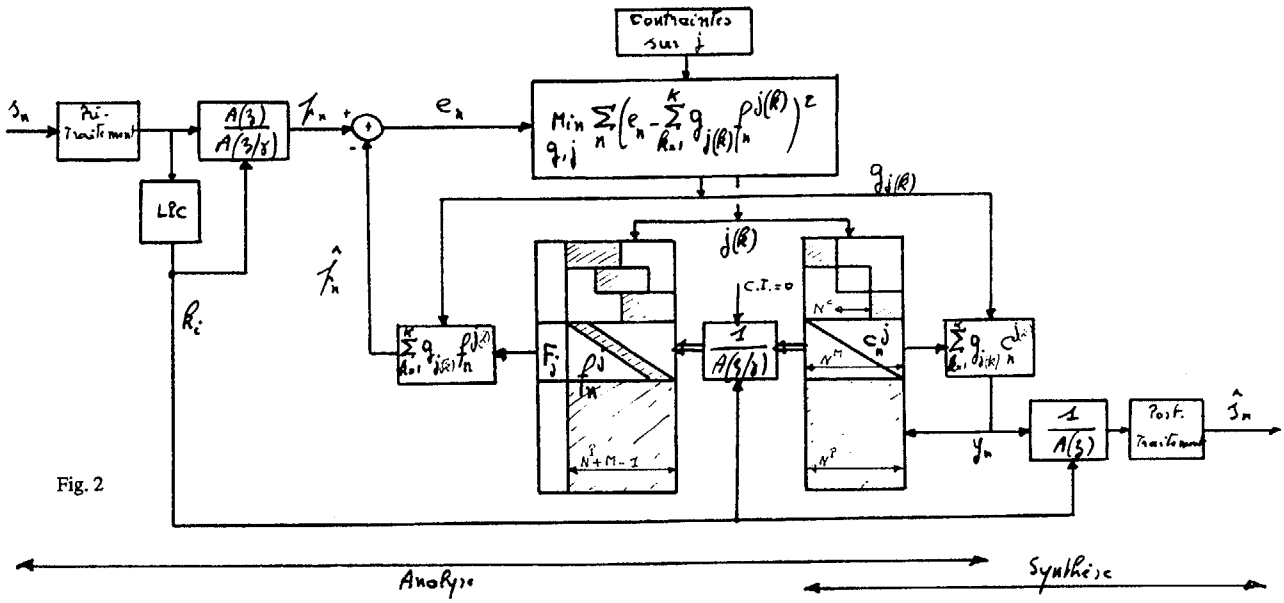

Fig. 2

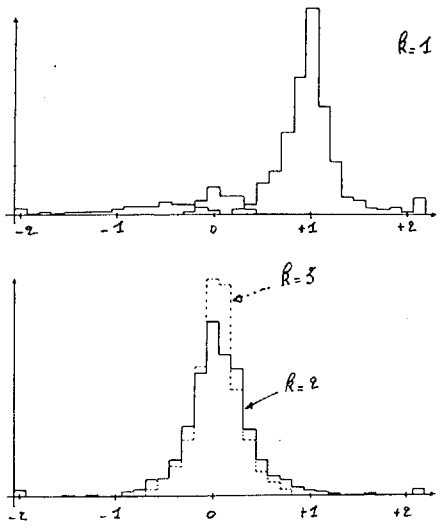

Fig. 4

III.4 Comportement

La figure 5 affiche le choix du codeur, pour chacune des sous-fenêtres lorsque K = 3, entre le dictionnaire CELP (=> C), le dictionnaire MP (=> M) et le dictionnaire prédictif (=> P). Comme on pouvait s'y attendre, le dictionnaire prédictif est peu utilisé lors de la transition. Son intérêt devient évident en cours de phonème. Il est alors utilisé plusieurs fois de suite.

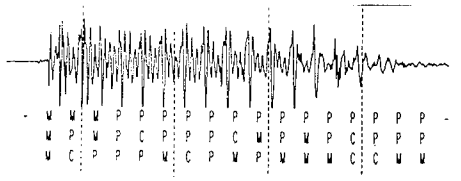

Fig. 5

La figure 6 montre plus en détail le comportement du codeur lorsque K = 2. La 1^{ere} modélisation est très efficace, la 2eme apparemment nettement moins.

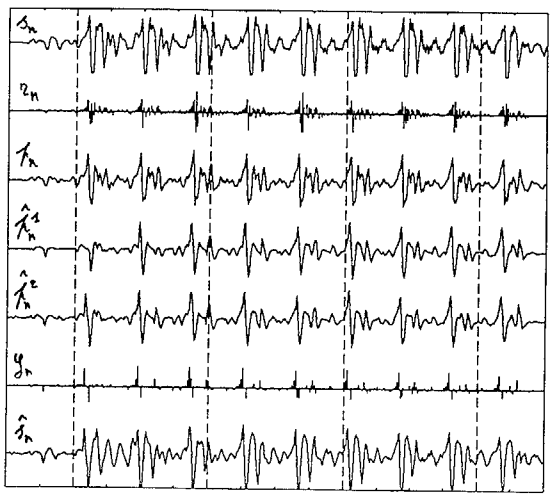

Fig. 6

III.5 Famille de codeurs

On remarque qu'il suffit, en 1^{ere} approximation, de jouer sur un seul paramètre, l'ordre de modélisation K du perceptuel, pour faire varier le débit. On choisit 36 bits pour coder les coefficients du filtre de synthèse toutes les 20 ms et 13 bits pour coder un vecteur d'excitation toutes les 5 ms (5 bits pour le gain et $\log_2 (L + N^M + Q_{max} - Q_{min} + 1)$ < 8 bits pour l'index). On obtient donc, pour K variant de 1 à 4, respectivement 4.4, 7, 9.6 et 12.2 Kb/s.

Les RSB segmentaux sur l'ensemble du corpus de test sont respectivement égaux à 4.5, 8.4, 10.3 et 11.8 pour les débits 4.4, 7, 9.6 et 12.2 Kb/s. A titre de comparaison, on a réalisé la même évaluation en utilisant un programme qui simule "au bit près" le codeur GSM. On a obtenu 10.0.

A l'écoute, le codeur GSM parait le meilleur mais la différence ne semble pas très significative avec le codeur à 12.2 Kb/s et même avec celui à 9.6 Kb/s. La dégradation auditive étant régulière lorsque le débit passe de 12.2 à 4.4 Kb/s, on a ainsi une famille de codeurs en jouant sur un seul paramètre.

Cette dégradation est clairement mise en évidence par le traitement de l'enregistrement d'une note (environ à 500 Hz), frappée sur un piano. Des tracés comparables à celui de la figure 6 montrent une très bonne modélisation de ce signal, spécialement lors de l'évanouissement de la note. Le RSB est assez élevé. Pourtant, l'écoute du signal codé est désagréable, même avec le codeur GSM (on a, en fait, très bien modélisé les basses fréquences). Le signal étant apparemment quasi-stationnaire sur une longue durée, on peut calculer des périodogrammes avec un nombre assez élevé d'échantillons ce qui permet d'avoir une bonne résolution fréquentielle. On observe que le signal original a une structure harmonique très riche jusqu'à 5 KHz (f_e = 10 KHz pour ce signal). Pour le signal codé, cette structure harmonique disparait pour des fréquences supérieures à 2 KHz environ, au profit d'un bruit important.

IV. IMPLANTATION DANS UN µPTS OU UN CIRCUIT ASIC

IV.1 Complexité de l'algorithme

Le nombre de multiplication/accumulations par seconde est essentiellement du au filtrage des dictionnaires, au calcul des énergies F_j et aux inter-corrélations, le reste étant négligeable. On trouve, pour le filtrage des dictionnaire CELP, MP et prédictif (une fois par fenêtre) : $(8000/N)*(L + N' + Q_{max} - Q_{min} + 1)*(N' + M - 1)*(\alpha P + 1)$ soit 16 Mflops (α = 3 puique, à priori, on utilise un filtre en treillis et qu'il y a un facteur perceptuel), pour le filtrage du dictionnaire prédictif (N/N' - 1 fois par fenêtre) : $(8000/N)*(N/N' - 1)*(2N' - Q_{min})*(N' + M - 1)*(\alpha P + 1)$ soit 13 Mflops, pour l'inter-corrélation (K fois par sous-fenêtre) : $(8000/N')*(L + N' + Q_{max} - Q_{min} + 1)*(N' + 1)*K$ soit 1.8*K Mflops, ce qui fait une masse de calcul comprise entre 31 et 36 Mflops.

La mémorisation des dictionnaires peut également poser quelques difficultés. On trouve : $(L + N' + Q_{max} - Q_{min} + 1)*(2*N' + M)$ soit 22 Kmots.

IV.2 Implantation dans un µPTS

Pour implanter cet algorithme dans un µPTS, il faut évidemment réduire cette complexité. On peut déjà se limiter au filtrage de N' échantillons (on a besoin de $f_N \cdots f_{N+M-2}$ uniquement pour calculer le perceptuel modélisé) et fixer α = 1 (en adoptant une structure transversale avec des coefficients pondérés par γ). La masse de calcul est ramenée à 9-14 Mflops et la mémoire nécessaire à 18 Kmots.

On peut aussi éviter de filtrer le dictionnaire multi-impulsionnel (gain de 0.7 Mflops) et utiliser l'algorithme proposé dans [4] pour le filtrage du dictionnaire prédictif et même du dictionnaire stochastique.

Si on accepte une dégradation des performances, on peut encore réduire l'ordre du filtrage des dictionnaires (de moitié par exemple) et diminuer Q_{max}. On arrive ainsi à une masse de calcul raisonnable.

IV.3 Implantation dans un ASIC

Si l'on désire réaliser un circuit spécialisé, le nombre initial de multiplications/accumulations est important mais il n'est pas rédibitoire car l'algorithme est très régulier, l'essentiel du traitement étant constitué par le filtrage du dictionnaire et le calcul des inter-corrélations. Il est possible de définir simplement quelques opérateurs spécialisés ayant ce débit de calcul [5] et de les connecter à un coeur de µprocesseur. Il ne faut pas chercher à réduire le nombre de multiplications car on diminue alors la régularité et on complexifie l'implantation.

V. REFERENCES

1. "Codeur multi-impulsionnel avec prédiction vectorielle à long terme" N.MOREAU, P.DYMARSKI, J.G.FRITSCH - GRETSI 87
2. "Multiple-stage vector excitation coding of speech waveforms" DAVIDSON, GERSHO - ICASSP 88
3. "Etude et simulation d'une famille de codeurs hybrides temporels offrant des débits de 6 à 12 Kb/s pour des applications de qualité sub-téléphonique" J.G.FRITSCH - Thèse - NANCY 1988
4. "Improved speech quality and efficient vector quantization in SELP" W.KLEIJN, D.KRASINSKI, R.KECHTUM - ICASSP 88
5. "VLSI architecture for a real-time LPC-based feature extractor" H.BARRAL, N.MOREAU - ICASSP 86