SEPTIEME COLLOQUE SUR LE TRAITEMENT DU SIGNAL ET SES APPLICATIONS

NICE du 28 MAI au 2 JUIN 1979

CODAGE PREDICTIF DU SIGNAL DE PAROLE A 4800 BPS

C. Galand, D. Esteban — C.E.R IBM, 06610 La Gaude

D. Mauduit, J. Menez — E.R.A 835 C.N.R.S, Université de Nice

RESUME

Cet article décrit un système de compression numérique du signal de parole utilisant des techniques modernes de traitement du signal telles que la prédiction linéaire et le codage par décomposition spectrale, associées au concept d'excitation par bande de base.

L'architecture proposée permet d'obtenir un signal de qualité téléphonique pour un débit d'information de 4800 bps, et représente une solution intermédiaire pour les systèmes de compression à faible débit que l'on peut généralement diviser en deux groupes :

- Dans les systèmes du premier groupe, on transmet après codage le signal limité à la bande téléphonique 300-3400 Hz. Si l'on suppose une fréquence d'échantillonnage de 8 kHz, le taux d'information nécessaire pour obtenir un signal de qualité téléphonique est dans tous les cas supérieur à 10 kbps.

- Dans les sytèmes du second groupe, généralement désignés sous le nom de vocodeurs, on ne transmet pas le signal mais un certain nombre de ses caractéristiques, ce qui permet de réduire le taux d'information entre 2 et 4 kbps. Cette réduction se fait néammoins au détriment de la qualité du signal synthétisé qui peut dépendre du locuteur et laisser apparaitre certaines résonances peu naturelles.

Il est montré d'autre part que l'algorithme proposé peut être mis en oeuvre en temps réel à l'aide d'une architecture à base de microprocesseur rapide sans multiplieur cablé.

Des enregistrements de parole codée/décodée seront présentés à la conférence pour permettre d'en apprécier la qualité subjective.

SUMMARY

This paper describes a voice compression system based on some modern techniques such as linear prediction, sub-band coding, and base-band excitation.

The proposed architecture allows a telephone quality at a bit rate of 4800 bps, and represents a trade off for low bit rate compression systems which include coders and vocoders :

- In coders, the speech signal is band-limited (300-3400 Hz), coded, and transmitted. A telephone quality is obtained for transmission rates around 10 kbps, assuming an 8kHz sampling rate.

- In vocoder systems, several characteristics of the signal are extracted,transmitted and used at the receiver to reconstruct a synthetic signal. These techniques allow to decrease the bit rate around 2 kbps, but introduce some degradations in the decoded speech which sounds less natural.

In addition, it is shown that the proposed algorithm can be implemented on a fast microprocessor without hardwired multiplier.

Simulation results will be played at the conference.

1.0 INTRODUCTION ET GENERALITES.

La transmission numérique du signal de parole présente, par rapport à la transmission analogique, un certain nombre d'avantages parmi lesquels on peut citer par exemple la facilité de multiplexage temporel de plusieurs voies et l'élimination du bruit cumulatif dans le cas de retransmissions. D'autre part, le traitement numérique se prête bien à l'utilisation de codes détecteurs et correcteurs d'erreurs, et permet une plus grande sécurité dans le cas de communications chiffrées.

Néammoins, le principal obstacle à la transmission numérique de la parole sur le réseau téléphonique est l'étroitesse de la bande passante du canal (300-3400 Hz) qui limite, dans la pratique, le débit à un maximum de 9600 éléments binaires par seconde (ebps ou bps) alors que le standard établi par le CCITT est de 64 kbps.

Le rôle essentiel de tout système de codage consiste à réduire au mieux les redondances du signal à transmettre tout en permettant de restituer un signal analogique de qualité téléphonique.

Le signal de parole présente deux types distincts de redondances qui se traduisent, d'une part, par la présence de formants (fréquences de résonance du conduit vocal) dans son enveloppe spectrale et, d'autre part, par une quasi périodicité du signal (fréquence fondamentale de vibration des cordes vocales) pendant les sons voisés.

Parmi les codeurs les plus efficaces (opérant à des taux de transmission inférieurs à 16 kbps), on distingue les codeurs prédictifs dont le principe est d'éliminer les deux redondances du signal de parole.

1.1 LES CODEURS PREDICTIFS.

L'efficacité des codeurs prédictifs réside dans le fait qu'ils ne transmettent, après quantification, que la partie de chaque échantillon du signal qui ne peut être prédite de la connaissance des échantillons précédents.

Le traitement s'effectue, en général, en deux étapes:

- un filtre prédicteur à court terme blanchit l'enveloppe spectrale. L'utilisation d'un tel procédé se justifie par le fait que pour des sons voisés, le canal vocal peut etre assimilé à une cascade de résonateurs modélisés par un filtre linéaire purement récursif. L'erreur de prédiction s'interprète alors comme l'excitation du filtre modèle du conduit vocal.
- un filtre prédicteur à long terme permet ensuite de rendre compte d'une éventuelle périodicité du signal.

Si l'on s'intéresse plus particulièrement au problème de la prédiction à court terme, on remarque que deux types de codeurs prédictifs on été proposés à ce jour. Ils se distinguent par la structure de leur émetteur.

1.1.1 CODEUR A EMETTEUR RECURSIF.

Ce type de codeur, proposé par Atal et Schroeder /1/, est illustré à la figure 1. Le signal prédit $\tilde{S}r(n)$ est calculé à partir des échantillons du signal d'erreur de prédiction quantifié. Si l'on note Er(n) le signal d'erreur et Qr(n) le bruit additif qu'introduit le quantificateur, l'émetteur transmet Er(n)+Qr(n). En remarquant que le récepteur est inclu dans la boucle de prédiction de l'émetteur, on peut écrire la transformée en Z du signal de sortie en fonction des transformées en Z du signal d'entrée et du signal d'erreur :

$$(1)\quad Rr(Z) = \tilde{S}r(Z) + Er(Z) + Qr(Z)$$

d'où

$$(2)\quad Rr(Z) = S(Z) + Qr(Z)$$

Ce qui signifie que le bruit de transmission se réduit au bruit introduit par le quantificateur sur le signal d'erreur de prédiction.

1.1.2 CODEUR A EMETTEUR TRANSVERSAL.

Le quantificateur est maintenant placé à l'extérieur de la boucle de prédiction (Fig.2) /2/. Le signal prédit $\tilde{S}t(n)$ est obtenu par filtrage du signal d'entrée S(n). Le signal d'erreur de prédiction quantifié Et(n)+Qt(n) est envoyé au récepteur qui fournit le signal reconstitué Rt(n) dont la transformée en Z peut s'écrire :

$$(3)\quad Rt(Z) = S(Z) + \frac{Qt(Z)}{1 - P(Z)}$$

Le bruit de quantification est donc filtré par le filtre récursif:

$$(4)\quad (\,1 - P(Z)\,)^{-1}$$

correspondant au filtre modèle du conduit vocal.

1.1.3 CRITERE DE QUALITE.

La détermination du filtre prédicteur et du quantificateur pose le problème du choix d'un critère de qualité mesurable. Le critère communément admis est l'optimisation du rapport signal sur bruit, c'est à dire, pour les codeurs prédictifs, la minimisation de l'énergie du bruit de quantification. Dans le cas d'un quantificateur optimal, l'énergie du bruit de quantification est proportionnelle à l'énergie du signal d'erreur de prédiction /3/. C'est pourquoi le filtre prédicteur est choisi de façon à donner une erreur de variance minimale.

Le signal de parole ne possède pas de propriétés statistiques stationnaires ce qui oblige à calculer le filtre P(Z) par une méthode adaptative (on voit apparaître une première difficulté d'application du codeur à structure récursive qui oblige à optimiser globalement le prédicteur et le quantificateur). Dans la pratique, cependant, les paramètres du spectre varient lentement puisqu'ils dépendent des constantes mécaniques des muscles du conduit vocal. On considère donc en général, un modèle stationnaire pour des intervalles de temps de l'ordre de quelques millisecondes (10 à 30 ms par exemple) et pour chacun de ces intervalles, l'enveloppe spectrale du signal est approximée par le spectre du filtre modèle (4). Si l'ordre de ce filtre est bien choisi (en pratique 8, correspondant à la présence de quatre formants dans la bande de fréquence considérée), la redondance à court terme a été quasiment éliminée puisqu'alors l'enveloppe spectrale du signal d'erreur de prédiction est plate. Lorsque, de plus, le quantificateur possède un nombre de niveaux suffisant, l'erreur de quantification est décorrélée du signal et peut être assimilée à un bruit blanc. Ces remarques, ainsi que la relation (2), permettent d'expliquer qu'il résulte du codage de type récursif un bruit blanc (Fig.3) dont la nature granulaire est assez désagréable d'un point de vue subjectif. Le codeur transversal ne présente pas cet inconvénient car le bruit de quantification se trouve filtré par le filtre modèle du conduit vocal (relation (3), Fig.3). Il en résulte un effet subjectif nettement meilleur que dans le cas récursif, pour un même taux de transmission.

Lorsque l'on diminue le nombre de niveaux du quantificateur, l'effet de masque résultant de la structure transversale ne permet pas d'annuler toutes les imperfections. On perçoit en particulier un bruit sourd dans le cas du codage des sons voisés qui sont les plus importants du point de vue de l'appréciation subjective. Néammoins, cet inconvénient peut être corrigé en faisant subir au signal une pré-emphase adaptative lors du calcul du filtre prédicteur. En effet, si l'on calcule P(Z), non plus directement sur le signal vocal mais après filtrage par le filtre P'(Z) :

$$P'(Z) = 1 - a.Z^{-1} \quad (5)$$

où le coefficient 'a' représente le prédicteur optimal du premier ordre, le filtre modèle du conduit vocal n'est plus donné par la relation (4), mais par :

$$(1 - a.Z^{-1}).(1 - P(Z))^{-1} \quad (6)$$

Cette relation permet de remarquer que le filtre modèle possède les mêmes propriétés spectrales que le signal vocal (position des formants), mais que la pente moyenne de son spectre se trouve annulée par la pré-emphase. D'après (3), on voit donc que dans le cas du codeur transversal, le bruit de quantification demeure masqué par le signal mais avec un codage plus fin dans les parties subjectivement importantes du spectre (Fig.3).

1.2 LES VOCODEURS.

Si les codeurs prédictifs permettent d'obtenir un signal de bonne qualité pour un débit d'information de 10 à 16 kbps, ils ne peuvent être utilisés à des débits plus faibles. En effet, la quantification du signal résiduel nécessite au moins 1 élément binaire par échantillon, correspondant à un débit de 8 kbps, auxquels il faut rajouter le débit nécessaire au codage des coefficients du prédicteur, et aux paramètres du quantificateur.

Dans les vocodeurs, on ne cherche pas à transmettre une onde temporelle mais seulement certaines de ses caractéristiques permettant une reconstitution subjectivement satisfaisante. Pour modéliser l'enveloppe spectrale, on peut encore utiliser le modèle précédent (vocodeurs à prédiction linéaire). Dans ce cas, l'analyse du signal fournit un ensemble de paramètres qui sont : la décision de voisement, éventuellement la période du fondamental, les coefficients du prédicteur, et un facteur de gain. Ces paramètres sont transmis et utilisés à la réception pour générer un signal synthétique de la façon suivante : le signal d'excitation est assimilé soit à un train d'impulsions périodique dans le cas des sons voisés, soit à du bruit blanc pour les fricatives, ou encore de façon plus générale à une combinaison des deux (Fig.4).

Ce type de système permet de réduire le débit d'information entre 2 et 4 kbps, mais au détriment de la qualité du signal dont on perçoit nettement la nature synthétique.

Depuis quelques années on voit se développer des systèmes de codage intermédiaires pour des débits binaires inférieurs à 9600 bps. Leur principe repose sur le fait que la majeure partie de l'information contenue dans le signal d'excitation peut être extraite d'une bande étroite de son spectre.

2.0 LE CODEUR PROPOSE

On décrit ici un codeur à prédiction linéaire et à bande de base dont le schéma de principe est donné par la figure 5. On suppose que le signal de parole est limité à la bande de fréquences 300-3400 Hz, ce qui correspond à une transmission sur ligne téléphonique, puis échantillonné à 8 kHz et quantifié sur 12 bits.

2.1 L'EMETTEUR

L'émetteur travaille par blocs de 256 données (32 ms) pour lesquels on peut considérer que les caractéristiques spectrales du signal sont constantes. L'émetteur se compose de:

- un filtre prédicteur dont les coefficients sont calculés par la méthode d'autocorrélation partielle /4/ sur le signal pré-emphasé. Les coefficients PARCOR correspondant sont quantifiés avant d'être transmis au récepteur. Le signal de parole est ensuite filtré par son prédicteur pour donner le signal d'erreur.

Il est à noter que les coefficients du prédicteur sont interpolés toutes les 8 millisecondes de façon à éviter des effets secondaires dûs à la non stationnarité du signal dans le cas de transitions trop rapides des caractéristiques formantiques du signal.

- un dispositif d'extraction de la bande de base: un filtre passe-bas élimine toutes les fréquences du signal d'erreur de prédiction supérieures à 1000 Hz. La bande de base ainsi obtenue est sous-échantillonnée à 2000 Hz. La partie haute fréquence du signal d'erreur est caractérisée par la valeur de son énergie qui est transmise et utilisée au récepteur pour reconstituer le signal. En fait, on calcule cette énergie haute fréquence quatre fois par bloc, c'est à dire toutes les 8 ms.

- un codage indépendant des différentes composantes spectrales de la bande de base. Un banc de huit filtres passe-bande possèdant des propriétés de quadrature /5/ découpe le signal en composantes indépendantes. Du fait de la réduction de la bande passante, chaque composante est échantillonnée à 250 Hz. Un algorithme d'allocation des ressources binaires /5/ est utilisé pour quantifier de façon préférentielle les composantes les plus importantes d'un point de vue auditif, c'est à dire celles qui ont le plus d'énergie. Ce dispositif permet, de façon implicite, de suivre les évolutions de la mélodie du signal sans avoir à en extraire le fondamental.

  Du fait de la limitation en fréquence du signal analysé, on remarque que les deux premières bandes spectrales (0-250 Hz) ne nécessitent aucun codage. De plus, des expériences conduites sur un grand nombre de locuteurs ont prouvé que les deux dernières bandes (750-1000 Hz) pouvaient être éliminées sans pour cela dégrader la qualité subjective du signal reconstruit.

2.2 LE RECEPTEUR.

Le récepteur doit regénérer un pseudo signal d'excitation à partir de la bande de base réduite et des caractéristiques qu'il reçoit. Il se compose de :

- un dispositif de restitution de la bande de base. Pour cela, chaque composante spectrale est interpolée en insérant sept valeurs nulles entre chaque échantillon et en filtrant par le filtre passe-bande correspondant. La somme de ces différentes composantes permet d'obtenir la bande de base qui est alors interpolée par filtrage de façon à ramener la fréquence d'échantillonnage à la fréquence initiale de 8 kHz.

- un dispositif destiné à générer la bande-haute du signal d'excitation par distorsion non-linéaire de la bande de base. La méthode de distorsion utilisée est un redressement double alternance qui permet de générer un signal satisfaisant du point de vue de la perception subjective car la structure spectrale des harmoniques du fondamental est conservée. Cependant, lors des consonnes chuintantes, la bande de base ne contient pas suffisamment d'information. Pour pallier cet inconvénient, un bruit blanc de faible énergie est ajouté en permanence au signal distordu. Du fait de sa faible énergie, il se trouve automatiquement masqué dans le cas des voyelles. On élimine ensuite les basses fréquences du signal ainsi obtenu puis on le normalise à l'aide des valeurs d'énergie haute fréquence transmises par l'émetteur de façon à ce que son énergie soit identique à celle du signal initial. Ce signal est ensuite additionné à la bande de base pour donner un signal d'excitation.

- Le filtre modèle du conduit vocal restitue enfin un signal de parole à partir du signal d'excitation ainsi obtenu.

On remarque donc que le dispositif proposé se classe entre les codeurs prédictifs et les vocodeurs, en ce sens que l'analyse du signal fournit à la fois une onde temporelle (la bande de base) et des caractéristiques (énergie haute fréquence, enveloppe spectrale) de ce signal. Ce dispositif permet de générer un signal de qualité téléphonique pour un débit d'information de 4800 bps. Des résultats de simulation seront présentés à la conférence et permettront d'apprécier la qualité subjective. Ils correspondent aux allocations de ressources binaires données dans la Table 1.

2.3 MISE EN OEUVRE.

L'algorithme de compression numérique proposé ici a été mis en oeuvre en temps réel à l'aide d'une architecture à base d'un microprocesseur rapide /6/ (cycle 250 ns). La puissance de calcul nécessaire à l'exécution des différentes fonctions utilisées par l'algorithme est donnée dans la Table 2. On remarque que la puissance totale est inférieure à 4 MIPS (Million d'Instructions Par Seconde).

3.0 CONCLUSION

On a décrit un système de compression numérique du signal de parole fonctionnant à un débit binaire de 4800 bps permettant d'obtenir un signal décodé de qualité téléphonique. Ce disposif présente un compromis entre d'une part les codeurs prédictifs où le signal d'excitation est intégralement transmis, qui permettent d'obtenir une qualité téléphonique mais dont le débit d'information est toujours supérieur à 10 kbps, et d'autre part les vocodeurs où l'on ne transmet que quelques paramètres caractéristiques de cette excitation, qui permettent un débit d'information de l'ordre de 2 à 4 kbps, mais qui présentent l'inconvénient de fournir un signal décodé peu naturel.

On a montré de plus que ce système pouvait être mis en oeuvre à l'aide d'un microprocesseur sans multiplicateur, prouvant ainsi la relative simplicité de l'algorithme.

4.0 REFERENCES

/1/ B.S. Atal, M.R. Schroeder.
'Adaptive predictive coding of speech signals' B.S.T.J. vol 49, pp 1973-1986, 1970

/2/ D. Esteban, J. Menez.
'Low bit rate voice transmission based on transversal predictive block coding'. 91st ASA Meeting, Washington April 1976

/3/ L.D. Davisson.
'Rate-distortion Theory and Application' Proc. IEEE, vol 60, no 7, July 1972

/4/ J. Le Roux
'Optimisation du calcul des coefficients PARCOR' 7èmes Journées d'Etudes sur la Parole Nancy 1976

/5/ D. Esteban, C. Galand
'Application of Quadrature Mirror Filters to Split Band Voice Coding Schemes'. Proc. 1977 Int'l Conf on ASSP, Hartford, May 1977

/6/ J.P.Béraud, D.Esteban, C.Galand, D.Mauduit, M.Maurel, O.Orsini.
'Une nouvelle architecture de microprocesseur pour le traitement du signal de parole'. GRETSI Nice, 28 Mai-2 Juin 1979.

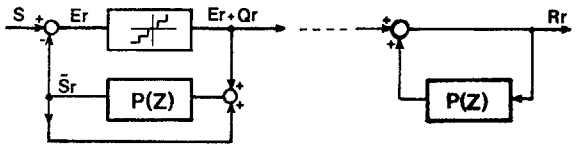


Fig 1 Codeur prédictif à émetteur récursif

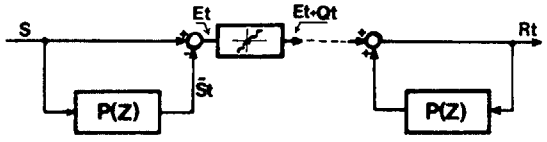
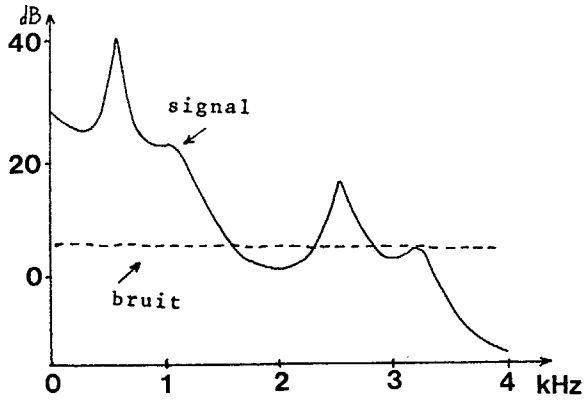
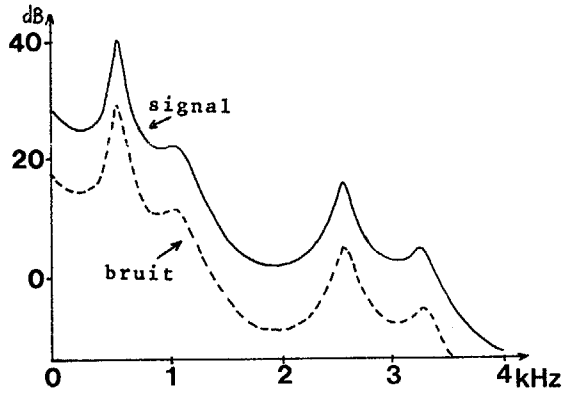


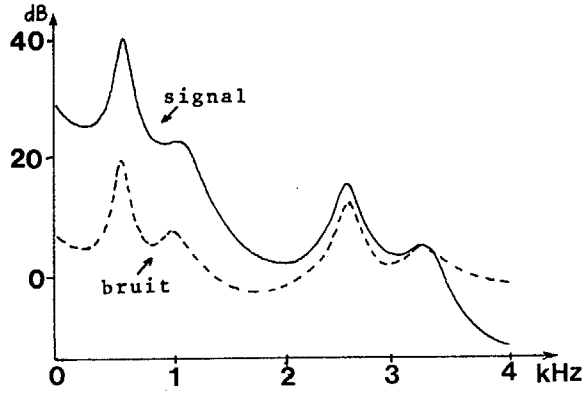
Fig 2 Codeur prédictif à émetteur transversal



A. Codeur récursif



B. Codeur transversal



C. Transversal et pré-emphase

Fig 3 Densité spectrale du signal et du bruit de quantification dans les codeurs prédictifs

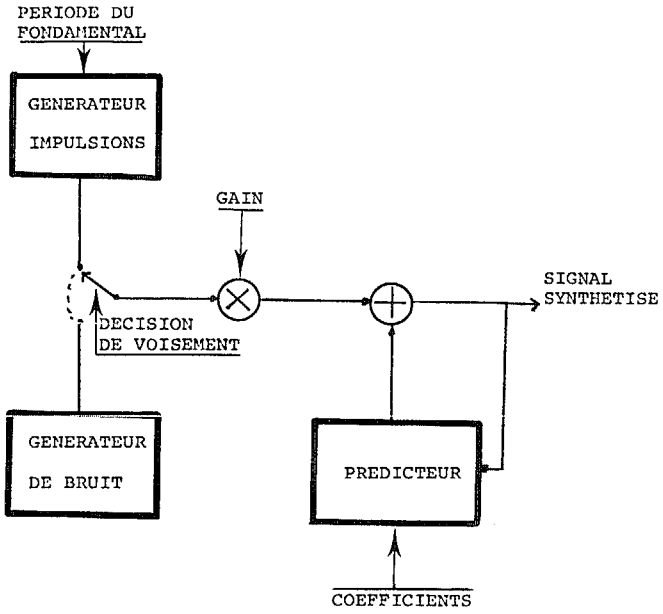


Fig 4 Récepteur d'un vocodeur à prédiction linéaire

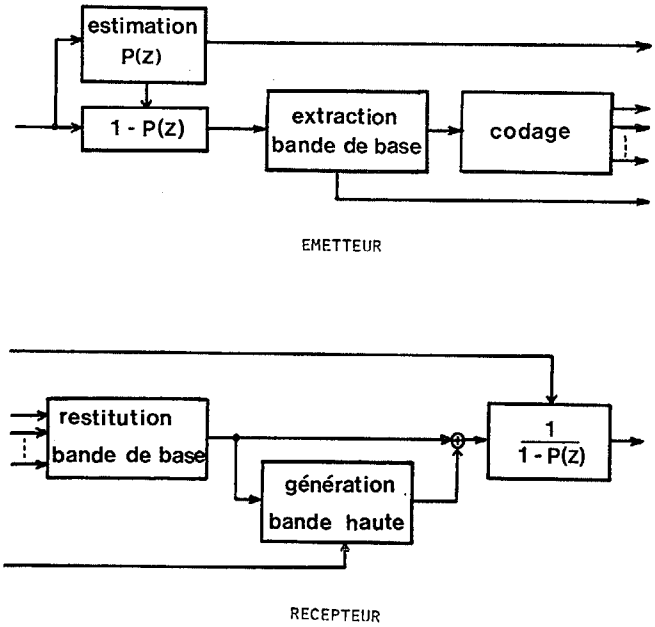


Fig 5 Codeur prédictif à excitation par la bande de base

TABLE 1 : ALLOCATIONS DES RESSOURCES	
FONCTION	Débit
Coefficients PARCOR	1.2 kbps
Bande de base quantifiée	3.2 kbps
Energie de la bande haute	0.4 kbps

TABLE 2 : MISE EN OEUVRE		
ANALYSE FONCTION	KCYCLES /BLOC	MIPS
Calcul du prédicteur	25.4	.794
Filtrage inverse	20.0	.624
Extraction de la bande de base	10.0	.313
Codage	6.0	.188
TOTAL	61.4	1.919
SYNTHESE FONCTION	KCYCLES /BLOC	MIPS
Restitution de la bande de base	16.8	.525
Génération de la bande haute	7.7	.240
Filtrage direct	20.0	.625
TOTAL	44.5	1.390
TOTAL (ANALYSE + SYNTHESE)	105.9	3.309