COLLOQUE NATIONAL SUR LE TRAITEMENT DU SIGNAL
ET SES APPLICATIONS

NICE du 26 au 30 AVRIL 1977

SYSTEME DE CODAGE DU SIGNAL DE PAROLE PAR DECOMPOSITION SPECTRALE

C. Galand, D. Esteban

CER IBM La Gaude

J. Menez, D. Mauduit

Laboratoire 190 CNRS
Université de Nice

RESUME

Cet article décrit un système de codage numérique du signal vocal par quantification indépendante de ses composantes situées dans des sous-bandes de fréquences déterminées. Le signal est tout d'abord numérisé, puis décomposé en sous-bandes par décimation des échantillons à l'aide de filtres demi-bande. Après quantification des différentes sous-bandes, le signal est reconstitué par interpolation à l'aide des mêmes filtres. L'utilisation de filtres en quadrature permet d'éliminer les repliements de spectres intervenant lors de la décimation.

Cette technique est par ailleurs appliquée à la quantification de l'erreur de prédiction d'un codeur de type transversal, et permet d'obtenir une qualité téléphonique pour un taux d'information de l'ordre de 16 KBps.

SUMMARY

This paper describes a split band voice coding scheme. The speech signal is sampled and split into sub-bands by a tree arrangement of half-band filters, using decimation techniques. The different sub-bands are then quantized separately and the speech signal is reconstructed by interpolation with the same filters. Using quadrature mirror filters enables to eliminate the aliasing effects due to decimation.

This technique is then applied to the quantization of the residual signal of a transversal predictive coder, resulting in a telephonic quality for a bit rate of 16 KBps.

SYSTEME DE CODAGE DU SIGNAL DE PAROLE PAR DECOMPOSITION SPECTRALE

Introduction

Le principe du codage du signal vocal par décomposition spectrale, proposé récemment par R. Crochiere et al /1/, est le suivant : le signal de parole est tout d'abord filtré par un banc de filtres passe-bande adjacents couvrant toute la bande téléphonique. Les signaux résultants sont ramenés en bande de base par modulation et échantillonnés à leur fréquence de Nyquist, la quantité totale d'information restant constante. Puis chaque signal (ou sous-bande) est quantifié séparément. La reconstitution du signal vocal se fait en démodulant chaque sous-bande et en filtrant les signaux démodulés par le même banc de filtres qu'à la décomposition.

Ce type de codage possède, par rapport aux procédés classiques de codage direct de l'onde temporelle, deux avantages principaux :

- le bruit de quantification produit dans chaque sous-bande reste localisé dans la bande de fréquence correspondante. Par conséquent, le bruit résultant de la quantification des sous-bandes de forte énergie ne masque pas le codage des sous-bandes de faible énergie; il en résulte une amélioration de la qualité subjective du codage.

- les sous-bandes peuvent être quantifiées de façon non uniforme. L'énergie du signal vocal étant, en moyenne, concentrée dans les basses fréquences, les sous-bandes correspondantes sont quantifiées avec un plus grand nombre de bits que les sous-bandes situées en haute fréquence.
  Cette propriété permet, pour un même taux d'information, d'augmenter le rapport signal sur bruit de façon très sensible.

Dans ce type de codage par décomposition spectrale, l'échantillonnage des sous-bandes à leur fréquence de Nyquist est capital, car il permet de conserver la quantité d'information constante. D'autre part, les filtres utilisés pour le découpage en sous-bandes doivent se couper à -3dB pour assurer une réponse globale unitaire dans toute la bande téléphonique. Il s'ensuit que l'échantillonnage provoque des repliements de spectres d'une bande à l'autre. Pour éliminer ces repliements auxquels l'oreille est très sensible, on peut utiliser un banc de filtres très raide se coupant à environ -20dB, mais alors, il apparaît un phénomène d'écho très désagréable dû au fait que la réponse globale du banc de filtres comporte des trous.

Le système de codage que nous proposons ici utilise un banc de filtres permettant de supprimer les repliements de spectres survenant lors de l'échantillonnage des sous-bandes /2/. Le banc de filtres est synthétisé à partir de filtres demi-bande en quadrature.

Cette technique, à titre d'exemple, est alors appliquée à la quantification du signal d'erreur résultant de la prédiction linéaire à court terme d'un codeur de type transversal /3/.

Il est observé que le codage en sous-bandes peut être associé à la prédiction linéaire et permet d'obtenir un système de compression de la voix de qualité téléphonique pour un taux d'information binaire de 16 KBps.

Principe de décomposition spectrale à l'aide de filtres en quadrature

1) Décomposition en deux sous-bandes

Considérons le système schématisé sur la Fig. 1 qui permet de décomposer un signal en deux sous-bandes d'égales largeurs, puis de le reconstituer. Le signal de parole x(t) est filtré par un filtre analogique passe-bas, puis échantillonné à la fréquence $f_e = 1/T$.

Le signal x(nT), noté par la suite x(n), est filtré par le filtre passe-bas demi-bande H_1 (dont la réponse impulsionnelle sera notée par la suite $h_1(n)$) et par le filtre passe-haut demi-bande H_2. Les signaux résultants $x_1(n)$ et $x_2(n)$ occupent respectivement, en fréquence, la moitié basse et la moitié haute du spectre de x(n). Par conséquent, la fréquence d'échantillonnage de chacun de ces signaux peut être divisée par deux en ne considérant qu'un échantillon sur deux (décimation). Il s'ensuit que la quantité totale d'information représentée par les signaux $y_1(n)$ et $y_2(n)$ est la même que celle contenue dans le signal x(n).

La reconstitution du signal de parole se fait de la façon suivante : la fréquence d'échantillonnage des signaux $y_1(n)$ et $y_2(n)$ est ramenée à f_e en intercalant un échantillon nul entre chaque échantillon et en filtrant les signaux obtenus respectivement par le filtre passe-bas demi-bande K_1 et par le filtre passe-haut demi-bande K_2, puis en additionnant les signaux obtenus.

Le problème qui se pose est de choisir les quatre filtres H_1, H_2, K_1, K_2, de façon à reconstituer parfaitement le signal x(n) sans repliement de spectre. Si X(z), $H_1(z)$ et $X_1(z)$ représentent respectivement les transformées en z de x(n), $h_1(n)$ et $x_1(n)$ (Fig. 1), on peut écrire :

$$X_1(z) = H_1(z)\, X(z) \qquad (1)$$

Les transformées en z du signal décimé $y_1(n)$ et du signal interpolé $u_1(n)$ sont données conformément à /4/ par :

$$Y_1(z) = \frac{1}{2}\left\{ X_1(z^{\frac{1}{2}}) + X_1(-z^{\frac{1}{2}}) \right\} \qquad (2)$$

$$U_1(z) = Y_1(z^2) \qquad (3)$$

D'autre part, la transformée en z de $t_1(n)$ est :

$$T_1(z) = K_1(z)\, U_1(z) \qquad (4)$$

où $K_1(z)$ représente la transformée en z de $k_1(n)$.

En combinant les relations (1)-(4) on obtient :

$$T_1(z) = \frac{1}{2}\{H_1(z)X(z) + H_1(-z)X(-z)\}\ K_1(z) \qquad (5)$$

De même, on peut calculer :

$$T_2(z) = \frac{1}{2}\{H_2(z)X(z) + H_2(-z)X(-z)\}\ K_2(z) \qquad (6)$$

La transformée en z du signal s(n) est obtenue par addition de (5) et (6) :

$$S(z) = A(z)X(z) + B(z)X(-z) \qquad (7)$$

avec :

$$A(z) = \frac{1}{2}\{H_1(z)K_1(z) + H_2(z)K_2(z)\} \qquad (8)$$

$$B(z) = \frac{1}{2}\{H_1(-z)K_1(z) + H_2(-z)K_2(z)\} \qquad (9)$$

Une reconstitution parfaite du signal équivaut à :

$$A(z) = \alpha\, z^{-q} \quad \text{avec } q \text{ entier et } \alpha \text{ constant} \qquad (10)$$

$$B(z) = 0 \qquad (11)$$

Considérons le cas particulier où le filtre est transversal symétrique à coefficients réels :

$$H_1(z) = \sum_{n=0}^{N-1} h_1(n) z^{-n} \quad \text{avec} \quad h_1(N-1-n) = h_1(n)$$

on obtient un filtre passe-haut demi-bande en alternant le signe des échantillons impairs de la réponse impulsionnelle de $h_1(n)$. La réponse harmonique de ce filtre passe-haut est alors exactement symétrique de celle du filtre $h_1(n)$ par rapport à la fréquence $f_e/4$.

Le polynôme B(z), qui représente les repliements de spectres, est identiquement nul si l'on choisit les filtres H_2, K_1, K_2 de la façon suivante :

$$\left.\begin{aligned} H_2(z) &= z^{-p}.H_1(-z) \\ K_1(z) &= z^{-p}.H_1(z) \\ K_2(z) &= (-1)^{p+1}.H_1(-z) \end{aligned}\right\} \quad (12)$$

avec p entier.

L'équation (7) s'écrit alors :

$$S(z) = A(z)X(z)$$

$$S(z) = \frac{1}{2}\{H_1^2(z) - (-1)^p H_1^2(-z)\}\ z^{-p}X(z) \quad (13)$$

Evaluons cette expression sur le cercle unité :

$$S(e^{j\omega T}) = \frac{1}{2}\{H_1^2(e^{j\omega T}) - (-1)^p H_1^2(e^{j(\omega+\frac{\omega_e}{2})T})\} \quad (14)$$
$$.e^{-jp\omega T}.X(e^{j\omega T})$$

or, la transformée de Fourier $H_1(e^{j\omega T})$ est égale à :

$$H_1(e^{j\omega T}) = H_1(\omega).e^{-j(N-1)\pi\frac{\omega}{\omega_e}} \quad (15)$$

avec $H_1(\omega)$ réel.

Après substitution de (15) dans (14) :

$$S(e^{j\omega T}) = \frac{1}{2}\{H_1^2(\omega) + (-1)^{N+p}.H_1^2(\omega+\frac{\omega_e}{2})\}. \quad (16)$$
$$e^{-j(N+p-1)\omega T}.X(e^{j\omega T})$$

Le filtre H_1 étant un filtre demi-bande, on peut écrire si $H_1^2(\frac{\omega_e}{4}) = \frac{1}{2}$ (-3dB) :

$$H_1^2(\omega) + H_1^2(\omega+\frac{\omega_e}{2}) = 1 + \varepsilon(\omega) \quad (17)$$

où le terme $\varepsilon(\omega)$, qui doit être petit par rapport à l'unité, représente l'imperfection des filtres.
Une reconstitution quasi-parfaite du signal est donc possible si (N+p) est pair; deux cas peuvent se présenter :

- le filtre H_1 a un nombre pair de coefficients, on peut alors choisir p=0, ce qui revient à échantillonner les signaux $x_1(n)$ et $x_2(n)$ en phase,
- le filtre H_1 a un nombre impair de coefficients, on peut choisir p=1, ce qui revient à échantillonner les signaux $x_1(n)$ et $x_2(n)$ alternativement.

Dans les deux cas, l'équation (16) s'écrit :

$$S(e^{j\omega T}) = \frac{1}{2}\{1+\varepsilon(\omega)\}.e^{-j(N+p-1)\omega T}.X(e^{j\omega T}) \quad (18)$$

Par transformation de Fourier inverse, on obtient :

$$s(n) \approx \frac{1}{2}\,x(n-N-p+1) \quad (19)$$

Le signal est reconstitué de façon quasi parfaite avec un gain 1/2 et un retard de (N+p-1) échantillons.

2) Décomposition en arbre

Le processus de décomposition en deux sous-bandes peut se généraliser immédiatement à un nombre supérieur de sous-bandes. En effet, chacun des deux signaux $y_1(n)$ et $y_2(n)$ (voir Fig. 1) peut à son tour être décomposé de la même manière en deux autres signaux. On dispose alors de quatre signaux échantillonnés à la fréquence $f_e/4$ qui représentent les quatre sous-bandes du signal initial. En répétant le processus m fois, le signal est décomposé en 2^m signaux échantillonnés à la fréquence $f_e/2^m$. Les filtres sont donc disposés suivant une structure en arbre, l'étage "i" contenant 2^i filtres décimateurs. La recomposition se fait à l'aide de filtres interpolateurs disposés suivant une structure en arbre symétrique de la première. La Fig. 2 montre un exemple de découpage en 4 sous-bandes.

Application au codage du signal vocal

1) Codage direct de l'onde temporelle

La technique de découpage en canaux peut être directement appliquée au codage du signal vocal. Le signal, échantillonné à 8 KHz est décomposé en huit sous-bandes par la structure de décimation en arbre exposée en détail ci-dessus. Les signaux résultants sont alors quantifiés par blocs selon le principe suivant /5/ : pour chaque bloc d'échantillons on définit un facteur d'échelle fonction de la dynamique du signal et du nombre de bits alloués. Les échantillons sont alors quantifiés par rapport à cette échelle et on transmet à la fois les valeurs codées et la caractéristique.

Ce codeur a été simulé pour un débit binaire de 16 KBps correspondant aux paramètres suivants :

- durée des blocs : 20 ms
- nombre de bits/canal : 3 3 3 1 1 1 1 1
- nombre de bits pour coder les facteurs d'échelle : 40 par bloc.

Il a été observé que ce procédé de découpage en canaux permet d'améliorer de 3 dB le rapport signal sur bruit par rapport au simple codage par blocs. La qualité subjective du signal reconstitué étant très nettement meilleure.

2) Codeur prédictif à excitation résiduelle

a - Prédiction linéaire

La présence de formants dans le spectre de puissance du signal vocal est une des causes de sa redondance. Un codeur efficace doit chercher à diminuer cette redondance et donc extraire l'information qui est contenue dans le signal en la caractérisant par un petit nombre de paramètres. Les méthodes de prédiction linéaire permettent d'identifier la fonction de transfert du conduit vocal et donc de reconstituer le signal d'excitation par filtrage inverse.

Ces méthodes sont fondées sur l'hypothèse qu'un $n^{ième}$ échantillon s(n) du signal, correspondant au n instant d'échantillonnage peut être prédit à partir d'une combinaison linéaire des p échantillons précédents :

$$\hat{s}(n) = \sum_{i=1}^{p} a_i.s(n-i)$$

Si e(n) désigne l'erreur de prédiction :

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

$$e(n) = s(n) - \sum_{i=1}^{p} a_i.s(n-i)$$

On applique la transformée en z à l'équation précédente :

$$E(z) = S(z).(1-\sum_{i=1}^{p} a_i.z^{-i}) = S(z).A(z)$$

$$\frac{S(z)}{E(z)} = \frac{1}{A(z)}$$

On peut donc assimiler le signal s(n) à la sortie d'un filtre linéaire purement récursif excité par le signal e(n). Selon les travaux de Fant /6/ et Flanagan /7/, la fonction de transfert 1/A(z) représente bien le modèle du conduit vocal pour des sons sans nasalité.

Le problème de la prédiction linéaire consiste donc à déterminer un ensemble de paramètres a_i, $1 \leqslant i \leqslant p$ qui minimisent, au sens des moindres carrés, l'erreur

de prédiction e(n).

La méthode la plus couramment utilisée est la méthode d'autocorrélation partielle qui consiste à identifier l'autocorrélation R_0, R_1, ..., R_p du signal pour un bloc d'échantillons à l'autocorrélation d'un filtre qui n'a pas de pôles /8/. Les paramètres a_i recherchés vérifient l'équation de Yule-Walker :

$$\begin{bmatrix} R_0 & R_1 & R_2 & \cdots & R_{p-1} \\ R_1 & R_0 & R_1 & \cdots & R_{p-2} \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ R_{p-1} & & \cdots\cdots & & R_0 \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ a_p \end{bmatrix} = \begin{bmatrix} R_1 \\ R_2 \\ \cdot \\ \cdot \\ R_p \end{bmatrix}$$

Parmi les méthodes d'identification des modèles auto-régressifs, la méthode d'autocorrélation partielle présente deux avantages :

. la résolution rapide du système précédent est permise grâce à l'algorithme de Durbin /9/,

. le modèle obtenu 1/A(z) est stable, ce qui est capital lors de la reconstitution en temps réel.

b - Description du codeur transversal

Le codeur proposé (Fig. 3) s'apparente à un vocodeur à prédiction linéaire mais il en diffère par le fait que l'on code l'onde temporelle du signal d'excitation obtenu par filtrage inverse. Le traitement s'effectue par blocs d'échantillons et peut se décomposer, pour la phase d'analyse en trois parties :

. filtrage inverse

Le calcul des coefficients du filtre inverse par la méthode d'autocorrélation partielle ne s'effectue pas directement à partir des échantillons du signal original mais à partir d'un signal pré-emphasé ou différentiel :

$$x(n) = s(n) - \gamma\, s(n-1)$$

Le coefficient γ est obtenu par minimisation de l'erreur quadratique de x(n) pour chaque bloc d'échantillons. Cette pré-emphase est destinée à redresser la pente moyenne du spectre du signal et donc à faire ressortir les formants d'ordre supérieur.

. découpage en canaux

Le signal e(n) obtenu par filtrage inverse est alors décomposé en huit sous-bandes de fréquence selon la technique de filtrage décimation développée précédemment.

. codage par bloc

De même que dans le premier codeur proposé, le signal sortant de chaque canal est quantifié par blocs. Le taux de bits alloués à chaque canal est le même que précédemment.

Le schéma de principe du synthétiseur comprend donc deux phases : on reconstruit le signal d'excitation par une structure d'interpolation-filtrage puis on convolue le signal ainsi obtenu par le filtre formateur 1/A(z).

L'emploi de ce procédé de découpage en sous-bandes du signal d'excitation résiduelle, a permis en simulation, de reconstituer un signal de qualité téléphonique pour des débits d'information de l'ordre de 16 KBps.

Conclusion

Un système de codage numérique d'un signal par découpage de son spectre en sous-bandes a été décrit. Ce système présente, par rapport aux approches temporelles le double avantage de permettre une réduction considérable de la distorsion de quantification, et d'améliorer l'effet subjectif de cette distorsion en augmentant l'intercorrélation entre le signal d'erreur et le signal codé. D'autre part, il a été vérifié pour le signal vocal que ce système peut être associé à des techniques de prédiction linéaire qui permettent de reproduire une meilleure qualité téléphonique en conservant les propriétés spectrales du signal (formants).

Des enregistrements correspondant à des taux d'information de 16 KBps obtenus par simulation de ces algorithmes sur ordinateur seront présentés de façon à permettre une appréciation subjective de ces résultats.

Références

/1/ R.E. Crochiere, S.A. Webber, J.L. Flanagan
"Digital Coding of Speech in sub-bands"
1976 Int'l Conf. on ASSP, Philadelphia.

/2/ A. Croisier, D. Esteban, C. Galand
"Perfect channel splitting by use of interpolation/decimation/tree decomposition techniques"
1976 Int'l Conf. on Information Sciences and Systems, Patras.

/3/ D. Esteban, J. Menez
"Low bit rate transmission based on Transversal Block Coding"
91st ASA meeting, Washington, April 4-9, 1976
(This paper is available on request to the authors)

/4/ R. Schaffer, L. Rabiner
"A digital signal processing approach to interpolation"
Proc. IEEE, Vol. 61, pp. 692-702, June 1973.

/5/ A. Croisier
"Progress in PCM and Delta Modulation : block companded coding of speech signals"
1974 ZÜrich Seminar.

/6/ G. Fant
"Acoustic Theory of Speech Production"
Mouton and Co, The Hague (1960).

/7/ J.L. Flanagan
"Speech Analysis, Synthesis and Perception"
Springer-Verlay, Berlin (1965).

/8/ J.D. Markel, A.H. Gray
"On Autocorrelation equations as applied to speech analysis"
IEEE Trans. on ASSP-22 n° 2, April 1974.

/9/ J. Durbin
"The fitting of Time-Series Models"
Rev. Intern. Statist. Vol 28, pp. 233-244, 1960.

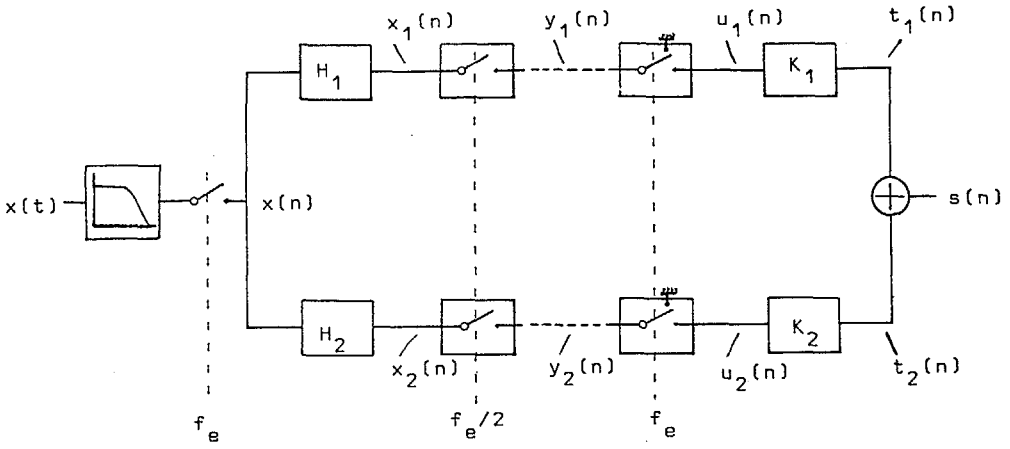


Fig 1 Découpage en deux sous-bandes à l'aide de filtres demi-bande.

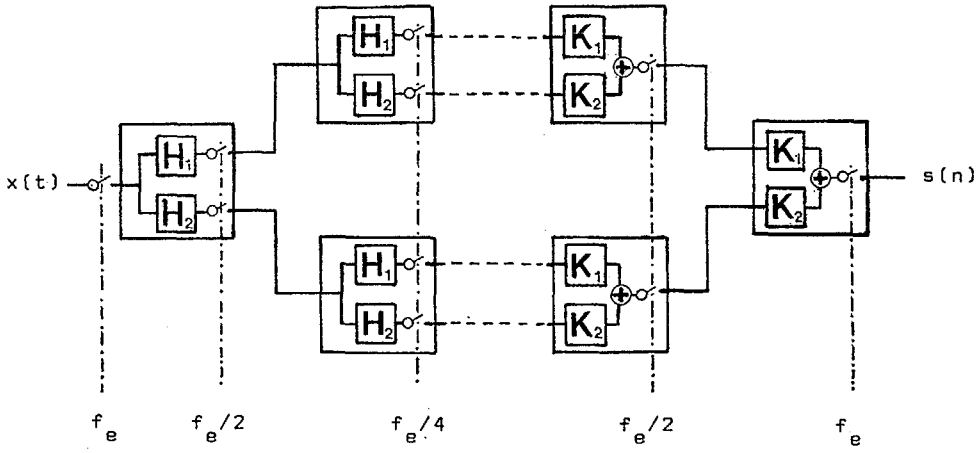
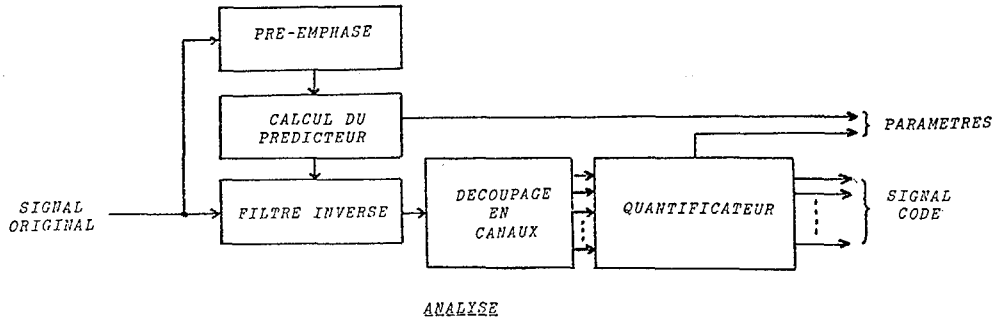


Fig 2 Décomposition en quatre sous-bandes.



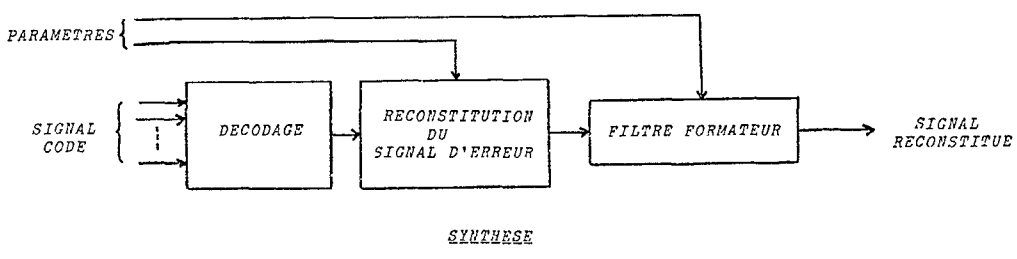


Fig 3 Codeur à excitation résiduelle et décomposition en sous-bandes.