Quantification Vectorielle en Réseau de Points D_5 pour un Codeur Audio Bas Délai en Sous-Bandes

K. Hay[(1)], **L. Mainard**[(2)] **et S. Saoudi**[(1)]

[(1)] ENST-Br, Dépt. SC., BP.832., 29285 Brest cedex, France

email: Karine.Hay@enst-bretagne.fr / Samir.Saoudi@enst-bretagne.fr

[(2)] CCETT, Dépt. RSA/SDA, rue du Clos Courtel, BP 59, 35512 Cesson Sévigné Cedex, France.

email: lmainard@ccett.fr

RÉSUMÉ

Une nouvelle méthode de codage des signaux audio génériques est présentée, à un débit de 64 kbit/s dans la bande de fréquence 20-15000 Hz, avec un faible délai. Cette méthode se base sur un codage en sous-bandes associé au LD-CELP ainsi qu'à des bancs de filtres cascadés. Des travaux initiaux [1] montrent que, lorsqu'un débit de 16 kbit/s est alloué à chaque sous-bande, la qualité audio n'est pas satisfaisante. Nous proposons un nouvel algorithme basé sur la quantification en réseau de points afin de pallier la complexité de la quantification vectorielle statistique. La souplesse du système permet d'effectuer une allocation binaire dynamique dans chaque sous-bande. Des résultats expérimentaux sont présentés et évaluent la validité de la méthode proposée.

ABSTRACT

A new method for coding generic audio signals at 64 kbit/s in the 20-15000 Hz bandwidth with a low delay is presented. It combines subband coding, Low Delay CELP algorithm and cascaded filterbanks. Our earlier works [1] show that, when using a 16 kbit/s bit rate on each subband, the resulting audio quality was not appropriate. We suggest a new technique based on lattice quantization to avoid the search complexity of the statistical vector quantization. It allows an adaptive bit rate allocation in each subband. Experimental results assessing the validity of the proposed method are presented.

1 Introduction

Durant ces dernières années, les techniques de codage du son haute qualité ont énormément progressé, et ont conduit l'ISO (Organisation Internationale de Normalisation) à normaliser des algorithmes de codage audionumérique sous la référence MPEG (Moving Pictures Expert Group). Ces algorithmes permettent de reconstruire le signal contenu dans une bande de fréquence allant de 10 à 20000 Hz avec une qualité proche de la transparence, le débit étant compris entre 64 et 128 kbit/s par voie monophonique. Cette qualité de codage à débit relativement faible est atteinte au prix d'une complexité croissante avec l'efficacité des algorithmes normalisés. Le délai de codage-décodage est un autre inconvénient de ces algorithmes, car trop élevé pour effectuer une voie de retour sur un locuteur.

Pour cette dernière raison, nous nous sommes interessés à la technique de codage LD-CELP (Low Delay Code Excited Linear Prediction) recommandé par le CCITT en 1992 sous la norme G728 [2, 3]. Ce codeur permet d'obtenir une bonne qualité pour les signaux de parole à un débit de 16 kbit/s et un délai inférieur à 2 ms. Murgia et Feng [4, 5] ont proposé récemment une extension directe du LD-CELP à la bande de fréquence 20 Hz-15000 Hz.

Dans cette article, nous proposons une nouvelle méthode de codage de signaux audio à un débit de 64 kbit/s dans la bande de fréquence 20-15000 Hz. Cette méthode associe un codage en sous-bandes avec la technique du LD-CELP. Nous décrirons brièvement le principe de notre codeur, et la manière d'intégrer des méthodes de codage perceptuel afin d'effectuer une allocation binaire dynamique sur chaque sous-bande. Enfin, nous introduirons la quantification en réseau de points D_5 qui permet de contourner la complexité de mise en œuvre d'une quantification statistique du signal résiduel.

2 Codage en sous-bandes associé au codeur LD-CELP

2.1 Description générale du codeur

Le signal est échantillonné à la fréquence de 32 kHz, ce qui permet d'obtenir une qualité satisfaisante pour des signaux audio. Le signal d'entrée est séparé en 4 sous-bandes de largeurs égales, à l'aide d'un banc de filtres PQF (Polyphase Quadrature Filter) [6]. Chaque sous-bande est codée par un codeur proche du LD-CELP [2, 3]. Les seules informations envoyées au décodeur sont les 4 meilleurs indices des dictionnaires de chaque sous-bande. Un modèle psychoacoustique permet d'effectuer une allocation binaire dynamique sur chaque sous-bande. Le délai d'un banc de filtres de 96 coefficients est de 3 ms à la fréquence d'échantillonnage retenue. Le délai total de notre codeur est donc de 5 ms. Un schéma du codeur est présenté en Figure 1.

Dans une première approche de notre codeur, nous avons optimisé le gain de prédiction sur chaque sous-bande [1]. Chacune de ces sous-bandes était codée avec un débit de 16 kbit/s. Nous avons facilement établi qu'il existait un bruit de quantification très audible sur le signal reconstruit. Le débit doit donc être adaptatif au signal pour chaque sous-bande.

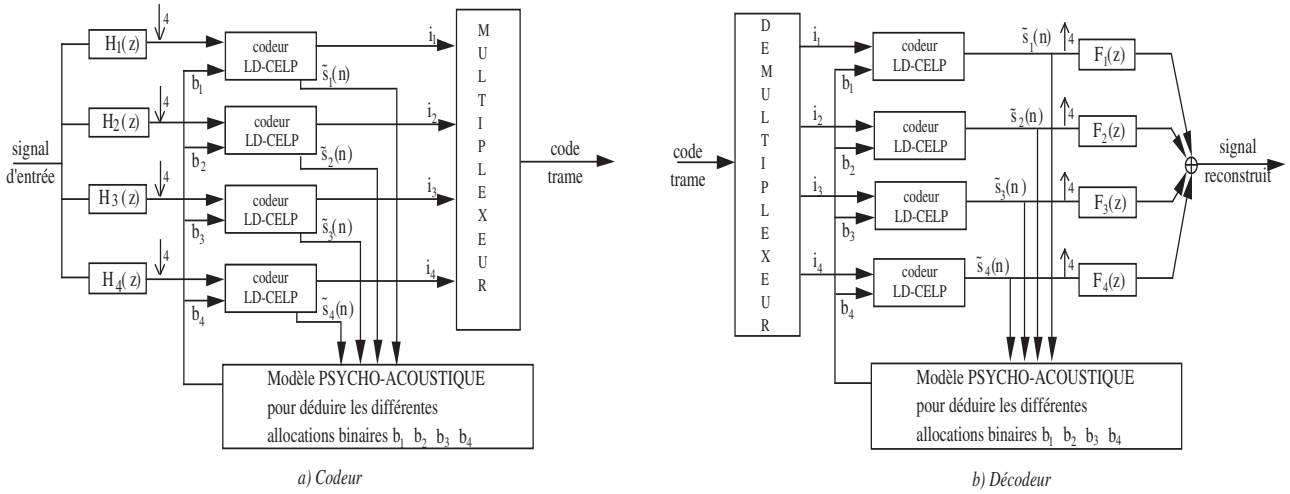

FIG. 1 — Codage en sous-bandes associé au codeur LD-CELP

2.2 Codage en sous-bande et contraintes de délai

Le codage en sous-bande utilisant un banc de filtres permet d'effectuer une modélisation du bruit de quantification sur toute la bande de fréquence tout en conservant l'avantage du gain de codage du banc de filtres. Pour la quantification en sous-bandes, il est préférable d'utiliser un banc de filtres produisant un nombre de sous-bandes élevé afin d'obtenir un meilleur gain de codage et une répartition plus fine en fréquence du bruit de quantification. En revanche, le délai du banc de filtres est lié au nombre de sous-bandes et, dans le cas de la MDCT, le délai total correspond à deux fois le nombre de sous-bandes. Par conséquent, une MDCT remplissant la contrainte de délai présenterait un faible nombre de sous-bandes (inférieur à 100), donc un mauvais taux de réjection et une mauvaise localisation en fréquence du bruit de quantification.

Pour ces raisons, nous avons donc choisi d'utiliser un banc de filtres PQF avec un nombre restreint de sous-bandes (4 sous-bandes), mais avec une réjection fréquentielle élevée. Afin de compenser le faible gain de codage, nous faisons intervenir la prédiction arrière [7].

2.3 Le codeur LD-CELP

Nous allons décrire rapidement les bases de fonctionnement du LD-CELP (voir Figure 2).

Cette technique de codage est basée sur la recherche dans un dictionnaire par analyse par synthèse. Elle a l'avantage de ne transmettre au décodeur que l'indice de la séquence d'excitation, les autres paramètres étant adaptés de façon "arrière". Cela signifie que les coefficients du filtre de prédiction linéaire (LPC) ainsi que le gain d'excitation ne sont pas calculés à partir du signal d'entrée, mais à partir des échantillons précédents du signal synthétisé. Le décodeur procède de même, il n'y a donc pas besoin d'envoyer ces paramètres.

Un signal d'excitation, appelé vecteur candidat, est stocké dans un dictionnaire et est appliqué au filtre de synthèse afin de produire le signal reconstruit. Le meilleur vecteur, au sens de la minimisation de l'erreur quadratique moyenne, est sélectionné et son indice est transmis au décodeur. Étant donné que la taille du buffer d'entrée est de 5 échantillons (0.625 ms à la fréquence d'échantillonnage de 8 kKz), le délai de codage-décodage est très faible (3 fois le buffer d'entrée : inférieur à 2 ms).

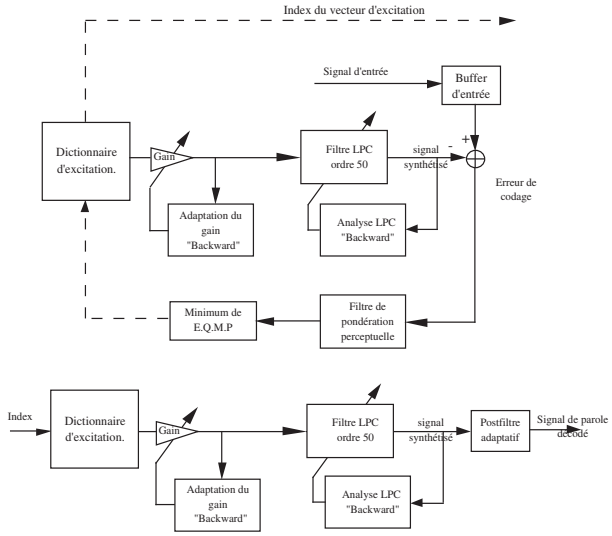

FIG. 2 — (a) Codeur low-delay CELP à 16 kbit/s. (b) Décodeur low-delay CELP à 16 kbit/s .

2.4 Introduction d'un modèle psychoacoustique

Adapté au fait que le niveau de bruit injectable maximum délivré par le modèle psychoacoustique est généralement plus faible pour les basses fréquences, un codage en sous-bandes permet d'effectuer une allocation binaire variable entre les sous-bandes, tout en conservant un débit global de 64 kbit/s. La taille du dictionnaire de formes d'ondes contenu dans le LD-CELP fixe le débit de la sous-bande. Pour chaque sous-bande, nous pouvons stocker des dictionnaires de différentes tailles correspondant à des précisions de codage données. En fonction du Rapport Signal à Masque (RSM) calculé par le modèle psychoacoustique, nous allouons plus ou moins de bits sur chaque sous-bande en sélectionnant le dictionnaire approprié.

À partir du signal synthétisé de chaque sous-bande, nous calculons un modèle psychoacoustique sur toute la bande de fréquence à l'aide d'une DFT incluant un traitement anti-aliasing [8]. Le modèle étant calculé sur des échantillons passés, il n'y a pas de perte de temps en stockage dans un buffer. Le déco-

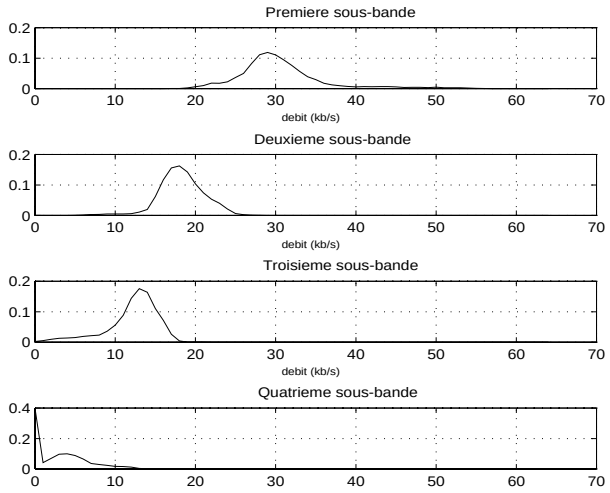

FIG. 3 — Densité de probabilité de l'allocation binaire.

deur peut effectuer la même opération, il n'est donc pas nécessaire d'envoyer d'informations.

Le RSM est calculé de la façon suivante : une courbe de masquage est calculée à partir des signaux synthétisés disponibles sur le LD-CELP avec un algorithme standard (Modèle 1 proposé par la norme MPEG) [9]. Nous rafraîchissons la courbe de masquage et l'allocation binaire toutes les 4 ms, une fois que le décodeur de chaque sous-bande a décodé 256 échantillons. La Figure 3 représente l'histogramme de l'allocation binaire calculé sur une base de données comprenant des signaux de tests audio couramment employés au sein de l'ISO. Le débit moyen nécessaire pour coder la première sous-bande est de 31 kbit/s, tandis que la deuxième ne requiert que 18 kbit/s, la troisième 12 kbit/s et la quatrième 3 kbit/s.

3 La quantification en réseau de points dans le LD-CELP

3.1 Limites de complexité de la quantification vectorielle statistique

Étant donné que le maximum de bruit injectable, délivré par le modèle psychoacoustique, est variable selon les sous-bandes, chaque sous-bande nécessite un débit approprié calculé à partir de la sortie du module psychoacoustique.

Une première approche consiste à stocker des dictionnaires de tailles différentes pour chaque sous-bande. Les dictionnaires sont optimisés de façon à ce que chaque sous-bande possède un certain nombre de dictionnaires correspondant à des précisions données. Les dictionnaires sont calculés à l'aide de l'algorithme de la K-moyenne [10, 11]. Le débit requis dans la première sous-bande est généralement trop élevé pour le schéma de quantification du LD-CELP classique. Dans le prochain paragraphe, nous présentons un nouvel algorithme qui satisfait aux contraintes de débit, donc de complexité.

3.2 Description de la méthode

Lorsque le débit proposé par le module d'allocation binaire est très élevé, la quantification vectorielle statistique devient impossible en temps réel. Nous proposons donc une nouvelle méthode de quantification du signal résiduel. Dans le LD-CELP, chaque vecteur contenu dans le dictionnaire est passé à travers un filtre de synthèse et comparé avec le signal original. Le meilleur vecteur du dictionnaire, r_q, est sélectionné selon le critère de l'erreur quadratique moyenne pondérée.

L'idée est de passer le vecteur du signal d'entrée à travers la chaîne de synthèse inverse, afin de trouver le signal résiduel, r. Pour quantifier ce résidu, nous utilisons un réseau de points [12]. Les vecteurs à quantifier sont de taille 5, nous utilisons donc un réseau D_5 défini comme suit :

$$D_5 = \{(x_1, \ldots, x_5) \in Z^5 : x_1 + \cdots + x_5 = m, \quad m \quad pair\},$$

Ces quantificateurs ont pour avantages de posséder des algorithmes de calcul rapides et de ne pas nécessiter de stockage de dictionnaire. Dans un premier temps, nous quantifions r dans le réseau de points. Ensuite, nous créons un dictionnaire temporaire autour du vecteur quantifié r_q. La taille maximum de ce dictionnaire est 50. Nous procédons à partir de ce dictionnaire à une analyse par synthèse comme dans le LD-CELP classique, afin d'obtenir le vecteur r_f qui minimise l'erreur quadratique moyenne pondérée.

Étant donné que le décodeur calcule aussi le débit, il peut en déduire la taille du réseau à utiliser. Seules les informations concernant r_f sont envoyées au décodeur.

3.3 Ajustement des paramètres

Pour un débit donné, le nombre de mots de code est fixé. Dans une première approche, nous avons décidé de fixer le débit de la première sous-bande à 32 kbit/s afin de pouvoir tester l'algorithme. Nous possèdons donc 20 bits pour coder 5 échantillons. Ces 20 bits doivent être judicieusement répartis entre le réseau de points et le dictionnaire de gain. Nous avons défini le gain comme un facteur d'échelle :

$$scf = \sum_{i=1}^{5} \mid x_i \mid$$

À l'aide de notre base de données, nous avons constitué une base de données de facteurs d'échelle en passant le signal à travers le filtre de synthèse inverse. Un histogramme de ces facteurs d'échelle est présenté en Figure 4 :

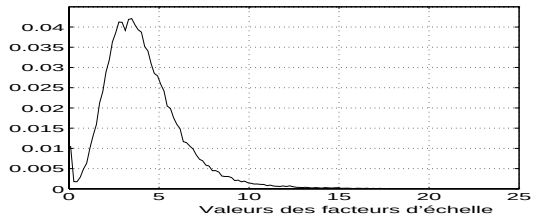

FIG. 4 — Histogramme des facteurs d'échelle

L'histogramme des facteurs d'échelle n'est pas uniformément réparti, ce qui nous a amené à constituer un dictionnaire de gain stochastique à l'aide de l'algorithme de la K-moyenne [10, 11].

Dans [13], une méthode de calcul du nombre optimum n de bits pour la quantification du facteur d'échelle est proposée dans le cas d'une quantification logarithmique.

$$n = log_2(\frac{A}{40} N ln 10) \qquad (1)$$

avec $A \approx 120dB$ et $N = 5$ (la taille du vecteur à quantifier). D'après cette méthode, la valeur théorique est autour de 4. Nous

avons calculé un Rapport Signal à Bruit (RSB) entre le signal original et le signal codé avec la quantification par réseau de point afin d'optimiser l'allocation binaire entre le facteur d'échelle et le réseau de points. Les résultats sont fournis dans le Tableau 1 et sont relatifs à la valeur théorique calculée dans 1.

Débit pour les facteurs d'échelle	débit pour le réseau de points	RSB relatif dB
3 bits	17 bits	-0.13
4 bits	16 bits	0
5 bits	15 bits	-0.54
6 bits	14 bits	-2.13

TAB. 1 — RSB en fonction de l'allocation binaire

En considérant les résultats du Tableau 1 ainsi que la valeur théorique entière calculée par l'équation 1, nous avons choisi d'allouer 4 bits pour les facteurs d'échelle et 16 bits pour le réseau de points, ce qui donne le meilleur RSB.

Étant donné que la restriction du réseau D_5 relative à la norme L_1 est loin d'être de la taille d'un entier puissance de 2, les bits alloués aux facteurs d'échelle et ceux alloués au réseau sont groupés. Ce qui laisse 20 valeurs possibles pour le facteur d'échelle et 50973 valeurs pour les quantificateurs, correspondants à tous les points ayant une norme L_1 inférieure ou égale à 12.

3.4 Comparaison avec une recherche exhaustive

Afin de valider notre méthode, nous avons comparé la recherche du meilleur quantificateur dans le dictionnaire réduit avec une recherche exhaustive dans tout le réseau de points. La recherche exhaustive consiste à passer chaque point de D_5 à travers le filtre de synthèse pour trouver le meilleur vecteur. Le Tableau 2 montre que la recherche sur un maximum de 50 points autour du signal résiduel quantifié donne des résultats proche d'une recherche exhaustive sur 50973 points.

Signal	Recherche exhaustive - Recherche restreinte SNR (dB)
German Speech	1.49
Suzanne Vega	0.30
Harpsichord	0.75
Coleman	1.18

TAB. 2 — Comparaison entre une recherche exhaustive et une recherche sur un dictionnaire réduit

4 Conclusions

Nous avons développé une nouvelle méthode de codage audio à faible délai. Cette méthode est basée sur un codage en sous-bandes associé au LD-CELP ainsi qu'à des bancs de filtres cascadés. Nos précédents travaux ont permis de valider notre méthode en terme de gain de prédiction [1]. Nous avons introduit un modèle psychoacoustique afin d'effectuer une allocation binaire dynamique sur les sous-bandes et une nouvelle technique de quantification en réseau de points a été développée pour subvenir aux demandes de débit trop élevées pour une recherche stochastique. Des résultats objectifs sur des signaux audio ont montré que la réduction de complexité n'affecte pas les performances globales du codeur. Cette méthode semble être prometteuse dans le sens où elle améliore la qualité des signaux audio dans le domaine du codage audio faible délai. Des travaux à venir s'appuieront sur une meilleure adaptativité du quantificateur aux débits alloués ; des tests formels seront effectués en accord avec les codeurs standards existants.

Références

[1] K. Hay, L. Mainard, and S. Saoudi. A low delay subband audio coder (20hz-15khz) at 64 kbit/s. In *Proc IEEE-SP on Time-Frequency and Time-Scale Analysis*, pages 265–268, june 1996.

[2] J.H. Chen, R.V. Cox, Y.C. Lin, N. Jayant, and M.J. Melchner. A Low-Delay CELP Coder for the CCITT 16 kb/s speech coding standard. *IEEE J. Select. Areas Commun.*, 10(5) :830–849, june 1992.

[3] *Coding of speech at 16 Kbit/s using Low-Delay Code Excited Linear Prediction*, sept. 1992. Recommendation G.728, CCITT.

[4] C. Murgia, G. Feng, C. Quinquis, and A. Le Guyader. Very low delay and high quality coding of 20 hz-15 khz speech at 64 kbit/s. In *4th Europ. Conf. on Speech Comm. and Technol.*, volume 1, pages 37–40, sept 1995.

[5] C. Murgia, G. Feng, C. Quinquis, and A. Le Guyader. Very low delay and high quality coding of 20 hz-15 khz speech at 64 kbit/s. In *Int. Conf. on Spoken Lang. Process.*, pages 302–305, oct. 1996.

[6] K. Akaigiri. Detailed technical description for MPEG2 audio NBC. Technical report, ISO/IEC JTC1/SC29/WG11, 1995.

[7] H. Fuchs. Improving MPEG Audio Coding by Backward Adaptive Linear Stereo Prediction. In *99th AES Convention*, volume preprint 4086, New York, 1996.

[8] B. Tang, A. Shen, G. Pottie, and A. Alwan. Spectral Analysis of Subband Filtered Signals. In *Proc IEEE Int. Conf. Acoust., Speech, Signal Process.*, pages 1324–1327, may 1995.

[9] *ISO/MPEG-Audio Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mb/s*, 1992.

[10] Y. Linde, A. Buzo, and R.M. Gray. An algorithm for vector quantizer design. *IEEE Trans.*, COM-28 :84–95, jan 1980.

[11] B. Kövesi, S. Saoudi, J.M. Boucher, and Z. Reguly. A fast robust stochastic algorithm for vector quantizer design for nonstationary channels. In *Proc IEEE Int. Conf. Acoust. and Speech, Signal Process.*, pages 269–272, may 1995.

[12] J. H. Conway and N. J. A. Sloane. *Sphere packing, Lattice and groups*. Springer Verlag, New York, 1988.

[13] L.Mainard. A low delay encoding scheme. In *100th AES Convention*, volume preprint 4199, Copenhague, 1996.