CODAGE DES SIGNAUX DE PAROLE (20 Hz A 15 kHz) A TRES FAIBLE RETARD AU DEBIT DE 64 KBIT/S

C. Murgia[1], G. Feng[1], A. Le Guyader[2] & C. Quinquis[2]

murgia@icp.grenet.fr

[1] Institut de la Communication Parlée, URA CNRS n° 368, INPG/ENSERG, Université Stendhal, GRENOBLE, FRANCE
[2] CNET Lannion A, TSS/CMC, 22301 LANNION, FRANCE

RÉSUMÉ

Dans cette communication, nous proposons un algorithme de compression de haute qualité des signaux de la bande 20 Hz - 15 kHz avec un débit de 64 kbit/s et à très faible retard (trame de 0,16 ms). Pour atteindre une qualité proche de la transparence, nous proposons une adaptation de la technique de codage Low-Delay CELP à la bande 20 Hz - 15 kHz et nous introduisons une nouvelle technique de mise en forme du bruit de quantification. De cette manière, nous associons les avantages des techniques de codage par prédiction linéaire aux propriétés de masquage fréquentiel du système auditif humain.

ABSTRACT

In this paper, we propose an algorithm for coding 20 Hz - 15 kHz signals at 64 kbit/s with a very low delay (frame 0.16 ms). To achieve a quality near to transparency, we propose to adapt the Low-Delay CELP coder to the 15 kHz bandwidth. A new noise shaping method is suggested. In this way we take advantage of linear predictive coding and masking properties of the human hearing.

1. INTRODUCTION

La compression des signaux de parole de haute qualité joue un rôle de plus en plus important dans les systèmes modernes de télécommunication, particulièrement dans les systèmes de communication de groupes. Par rapport à la bande téléphonique (300 Hz - 3400 Hz), une bande passante de 20 Hz - 15 kHz permet d'accroître significativement la richesse et le côté naturel des signaux codés ainsi que la sensation de présence des locuteurs lointains. De plus, dans ces systèmes, il est essentiel d'assurer une bonne interactivité en situation de conversation. Compte tenu du retard dû aux différents traitements, à la prise de son et à la transmission, le retard maximal admissible codeur-décodeur doit être limité à des valeurs de l'ordre de 10 ms. Dans cet article, nous présentons un algorithme de codage de très haute qualité au débit de 64 kbit/s pour les signaux de la bande 20 Hz - 15 kHz. Nous utilisons, comme base, un système de codage prédictif excité par code (Low-Delay CELP), originellement développé pour la bande téléphonique [1]. L'adaptation de l'algorithme à la bande 20 Hz - 15 kHz pose de nombreux problèmes (stabilité des filtres, optimisation des paramètres, etc.), dont la résolution est indispensable pour réaliser un codage de haute qualité.

Dans cette communication, nous proposons plusieurs solutions pour optimiser les paramètres, ainsi qu'une nouvelle technique de mise en forme du bruit de codage basée sur un modèle psycho-acoustique du système auditif humain.

2. DESCRIPTION GENERALE DE L'ALGORITHME

L'algorithme de codage est basé sur le codeur CELP à adaptation " backward " [1]. Rappelons brièvement le principe de cette technique. Les coefficients LPC sont déterminés au codeur (décodeur local) et au décodeur par prédiction linéaire " backward " sur le signal reconstruit. Seul le signal d'excitation est transmis au décodeur. Constitué d'une forme d'onde de 5 échantillons issue d'un dictionnaire de 128 vecteurs, ajustée en amplitude par un facteur de gain, il est déterminé au codeur par une méthode d'analyse par synthèse. Grâce à cette technique " backward ", on obtient un très faible retard, la longueur de la trame étant de 5 échantillons.

Lors de l'application directe de la technique LD-CELP à la compression des signaux échantillonnés à 32 kHz, nous avons constaté de nombreux problèmes. Les signaux reconstitués présentent notamment un bruit de codage assez fort, répandu sur toutes les fréquences de la bande 20 Hz - 15 kHz. Une analyse détaillée a montré la nécessité de réoptimiser plusieurs paramètres et de reconsidérer le rôle de certains éléments du codeur.

3. ORDRE DU FILTRE DE SYNTHESE

Un paramètre essentiel est l'ordre du filtre de synthèse. Celui-ci a été fixé à 50 dans le codeur LD-CELP [1], afin de modéliser, en plus de l'enveloppe spectrale, la structure harmonique des signaux à fréquence fondamentale élevée. Pour maintenir le même gain de prédiction, l'ordre du filtre devrait être égal à 200, la fréquence d'échantillonnage étant de 32 kHz. Outre la complexité impliquée par un ordre si élevé, se pose le problème de la stabilité du filtre. Nous avons mené une première étude qui consiste à mesurer le gain de prédiction en fonction de l'ordre du filtre (figure 1). Les résultats montrent qu'une certaine saturation se produit au delà de l'ordre 60 pour les voix masculines. Pour les voix féminines, le phénomène de saturation est moins net.

Par la suite, en nous basant sur ces résultats, nous avons simulé une version du codeur pour quatre valeurs de l'ordre du filtre de synthèse. La figure 2 représente l'évolution du rapport signal sur bruit (RSB) segmental pour les signaux restitués en fonction de l'ordre du filtre. Au delà de l'ordre 64, le RSB n'augmente plus pour les signaux de musique et nous constatons même une légère diminution pour les signaux de parole.

Nous avons donc fixé l'ordre à 64, ce qui nous semble un judicieux compromis entre l'exigence d'une bonne résolution spectrale de l'enveloppe du signal et la complexité de l'algorithme.

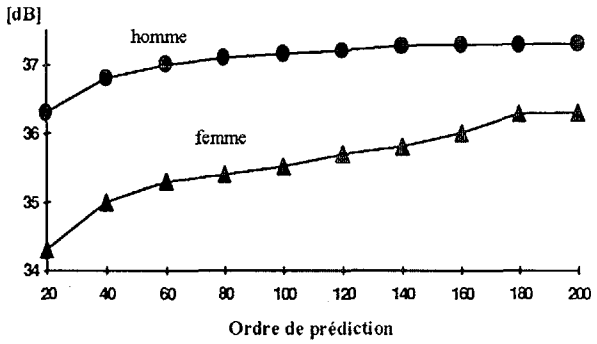

Figure 1. Exemple d'évolution du gain de prédiction en fonction de l'ordre du filtre de synthèse pour une voix d'homme et une voix de femme.

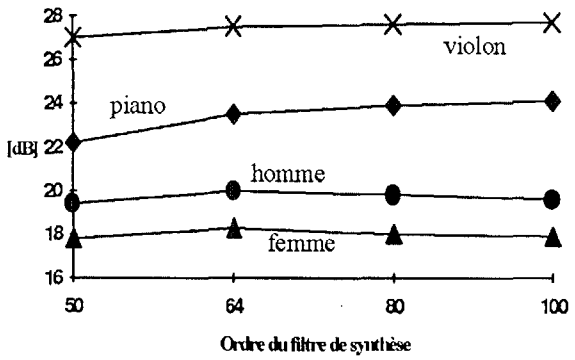

Figure 2. Evolution du rapport signal sur bruit segmental en fonction de l'ordre du filtre de synthèse pour des signaux de parole et de musique.

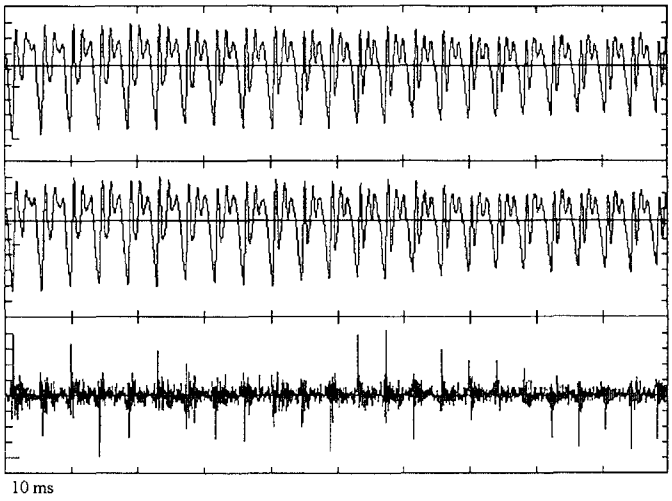

Figure 3. Modélisation de l'excitation du filtre de synthèse. De haut en bas : le signal original ($F_0 \approx 280$ Hz), le signal restitué et l'excitation court terme du filtre de synthèse.

4. MODELISATION DE LA FREQUENCE FONDAMENTALE

Le taille du vecteur d'excitation étant de 5 échantillons, elle correspond à une résolution temporelle de 0,16 ms pour une fréquence d'échantillonnage de 32 kHz. Ceci permet, par l'adaptation du gain et le choix des vecteurs d'excitation de reproduire la richesse et la périodicité du résidu sans nécessiter un filtre d'ordre excessivement élevé, et ce pour l'ensemble des fréquences fondamentales de la parole. La figure 3 illustre ce phénomène pour un segment de voix de femme : nous observons que le signal d'excitation modélisé présente des pics synchrones au *pitch* du signal original.

Dans notre codeur, le filtre de synthèse assure principalement la modélisation de l'enveloppe spectrale, tandis que la modélisation de la fréquence fondamentale est essentiellement effectuée par la détermination des vecteurs optimaux d'excitation et de leurs gains.

5. DYNAMIQUE SPECTRALE DES SIGNAUX AUDIO ET STABILITE DES FILTRES

Dans les systèmes à prédiction " backward ", les coefficients de prédiction linéaire sont obtenus par l'algorithme de Levinson-Durbin appliqué aux échantillons du signal précédemment codés. Lorsque les signaux sont très prédictibles, la dynamique spectrale est importante. La matrice d'autocorrélation peut être alors mal conditionnée faisant ainsi apparaître des instabilités du filtre de synthèse. Un moyen de résoudre ce problème est de multiplier la diagonale de la matrice d'autocorrélation par un coefficient $(1 + \varepsilon)$ [2]. Le choix de ε n'est pas simple. En effet, l'ajout d'un bruit blanc de forte puissance peut dégrader sensiblement le gain de prédiction. Dans notre étude, nous avons déterminé le niveau de bruit (-52 dB par rapport à l'énergie du signal) qui constitue un bon compromis entre le gain de prédiction et la stabilité. Cette optimisation est nécessaire pour le codage des

signaux à forte dynamique spectrale, notamment les signaux musicaux.

6. FILTRE PERCEPTUEL BASE SUR LE MASQUAGE FREQUENTIEL

L'utilisation d'un filtre perceptuel dans un codeur CELP permet la mise en forme du bruit de codage. Ce filtre est souvent exprimé sous la forme suivante :

$$W(z)=\frac{A(z/\gamma_1)}{A(z/\gamma_2)}$$

où 1/A(z) est le filtre de synthèse. Un bon masquage du bruit peut être obtenu en choisissant soigneusement γ_1 et γ_2, pour que la fonction de transfert du filtre perceptuel se trouve au-dessous de celle du filtre de synthèse. Cependant, en faisant varier γ_1 et γ_2, on ne peut pas contrôler à la fois la forme et la pente de sa fonction de transfert sur une large bande de fréquence lorsque la dynamique spectrale est importante. Pour cette raison, il a été proposé de l'associer à un filtre autorégressif dit " correcteur de pente ", dont le rôle est de modéliser la pente du spectre du signal à coder [3, 4]. De plus, issu d'un modèle de production de parole, ce type de filtre perceptuel n'arrive à prendre en compte que de façon limitée les caractéristiques perceptives du système auditif humain.

Nous proposons une nouvelle méthode de mise en forme du bruit, basée sur le modèle psycho-acoustique décrit en [7] et [8].

A chaque nouvelle trame LPC (soit toutes les 2,5 ms), une courbe de masquage est calculée sur les 1024 derniers échantillons du signal d'entrée, taille nécessaire pour assurer une définition spectrale suffisante. Ce signal est alors pondéré par une fenêtre mixte constituée d'une partie exponentielle pour les échantillons les plus anciens et d'une partie en cosinus pour les plus récents. Cette fenêtre permet d'augmenter, lors de la définition du filtre de pondération, le poids des échantillons les plus récents par rapport aux plus anciens. De plus, une telle fenêtre possède des caractéristiques spectrales comparables à celles de la fenêtre de Hamming [5, 6]. A partir du spectre d'énergie du signal (calculé par une FFT), une courbe de masquage est déterminée par convolution avec la fonction d'étalement spectral de la membrane basilaire suivant l'échelle des barks [7, 8]. On obtient ensuite les coefficients d'autocorrélation par une FFT inverse. Enfin, un modèle AR de la courbe de masquage est déterminé à partir de ceux-ci par l'algorithme de Levinson-Durbin.

Il est important de souligner que la méthode proposée n'introduit aucun retard supplémentaire, car les coefficients du filtre sont déterminés à partir des échantillons passés du signal d'entrée.

La figure 4 montre la courbe de masquage obtenue par l'algorithme proposé en [7] et [8] pour un segment de parole. La figure 5 représente le spectre LPC du filtre de

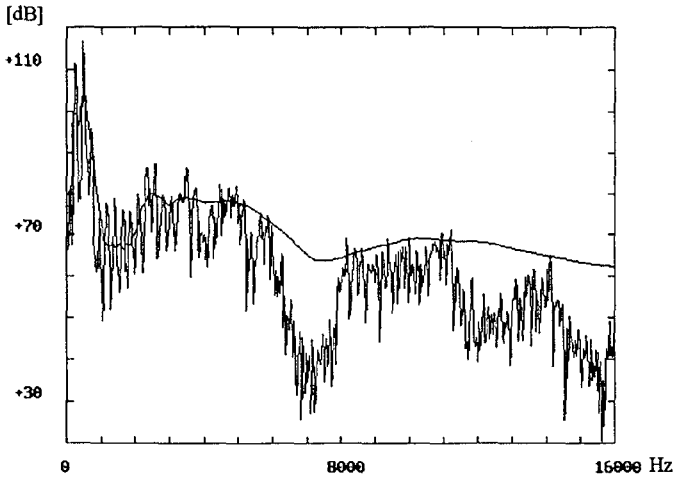

Figure 4. Spectre d'un signal de parole et courbe de masquage correspondante.

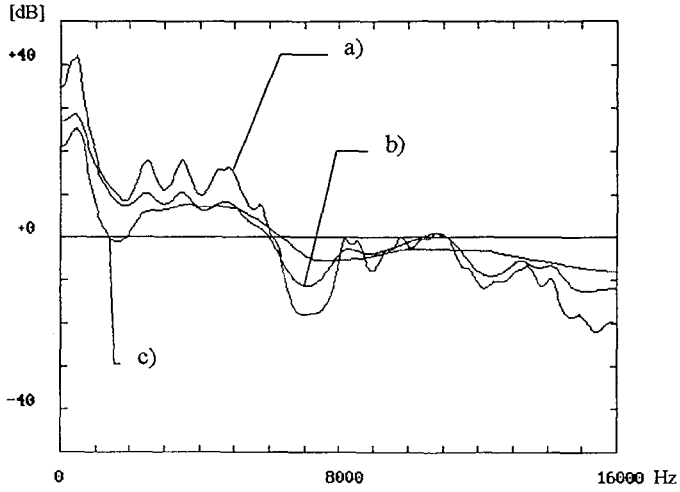

Figure 5. Allures des spectres LPC :
a) filtre de synthèse; b) filtre perceptuel calculé de façon classique, avec correction de pente; c) filtre perceptuel déduit de la courbe de masquage.

synthèse correspondant, et pour le même signal celui du filtre perceptuel calculé de manière classique avec correction de pente, et celui déduit à partir de la courbe de masquage. Nous observons que la courbe de masquage est bien modélisée par le nouveau filtre perceptuel. De plus, notons que la mise en forme obtenue par le filtre de pondération classique est loin d'être optimale au sens perceptuel du terme, si l'on considère que la mise en forme optimale est donnée par la courbe de masquage. En effet, en observant la zone de fréquence entre 1 et 3 kHz, il est clair que la réponse en fréquence du filtre classique dépasse la courbe de masquage, ce qui peut rendre le bruit de quantification plus facilement audible. D'autre part, entre 6 et 8 kHz et au delà de 11 kHz, sa réponse en fréquence est beaucoup plus faible que la courbe de masquage, ce qui a pour conséquence de mettre en forme le bruit de quantification largement au dessous du seuil d'audition. L'optimisation des coefficients γ_1 et γ_2 et des paramètres du correcteur de pente ne permet pas d'atteindre la mise en forme optimale.

7. ALLOCATION DES BITS ET OPTIMISATION DU DICTIONNAIRE

Puisque cet algorithme est entièrement " backward ", seuls les paramètres de l'excitation sont transmis au décodeur. L'indice de la forme d'onde est codé sur 7 bits et le facteur de gain sur 3 bits. La trame étant de 5 échantillons, le débit est de 64 kbit/s pour une fréquence d'échantillonnage de 32 kHz. Les vecteurs du dictionnaire de formes d'onde ont été optimisés par une procédure itérative [9]. La figure 6 montre les résultats de cette procédure d'optimisation. Le fait que la base de données lors des cycles d'optimisation ne soit pas fixe explique que la courbe n'est pas monotone décroissante. Bien que le gain, en terme de distorsion dans le domaine de l'erreur quadratique moyenne pondérée par le filtre perceptuel, soit seulement de l'ordre de 1,5 dB (le dictionnaire de départ étant déjà très performant), le gain en terme de qualité du signal reconstruit est très significatif. En effet, les résultats des écoutes informelles montrent que lorsque nous faisons suivre (dans un ordre aléatoire) le signal original par deux signaux codés, avec ou sans dictionnaire optimisé, les auditeurs trouvent dans 70% des cas que le plus proche de l'original est le signal codé par le dictionnaire optimisé.

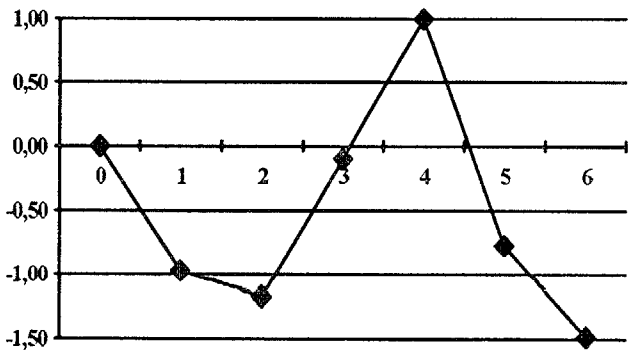

Figure 6. Optimisation du dictionnaire des formes d'onde : évolution de la distorsion totale dans le domaine de l'erreur quadratique moyenne pondérée par le filtre perceptuel, en fonction des cycles d'optimisation.

8. RESULTATS

Nous avons réalisé deux versions du codeur pour lesquelles tous les paramètres (ordre du filtre de synthèse, taille des vecteurs d'excitation, dictionnaire des formes d'onde,...) ont été optimisés. L'une utilise le filtre perceptuel classique avec correction de pente, l'autre le filtre perceptuel déduit de la courbe de masquage. Le tableau I résume les mesures du RSB segmental sur un corpus de signaux de parole et de musique, et ce pour les signaux restitués par les deux codeurs. Nous observons que l'utilisation du nouveau filtre perceptuel permet un gain de 1,46 dB. Ce dernier semble lié à une meilleure prise en compte du bruit de quantification surtout en basse fréquence, comme l'illustrent les figures 4 et 5. Des écoutes informelles confirment l'amélioration de la qualité apportée par cette nouvelle technique de mise en forme du bruit. Nous avons notamment pu constater une amélioration significative de la clarté des signaux restitués.

	parole	musique	total
a)	23.41	24.34	23.78
b)	24.62	26.18	25.24
gain	+1.21	+1.84	+1.46

Tableau I. Mesure du RSB segmental :
a) filtre perceptuel classique avec correction de pente;
b) filtre perceptuel basé sur la courbe de masquage.

9. CONCLUSION

Nous avons proposé un algorithme de codage des signaux de la bande 20 Hz - 15 kHz conçu pour fonctionner avec un débit de 64 kbit/s et un très faible retard. La qualité obtenue par simulation est tout à fait satisfaisante. Nous avons proposé l'utilisation d'un nouveau filtre de mise en forme du bruit de codage basé sur un modèle psycho-acoustique du système auditif humain. Son utilisation permet une très sensible amélioration de la qualité des signaux restitués. Des écoutes informelles effectuées sur de nombreux signaux de parole et quelques morceaux de musique ont montré que la qualité est assez proche de la transparence.

REFERENCES

[1] J.-H. Chen, R. V. Cox, Y.-C. Lin, N. Jayant & M. J. Melchner, " A Low-Delay CELP Coder for the CCITT 16 kb/s Speech Coding Standard " - IEEE Jour. on Selec. Areas in Comm., vol. 10, n° 5, 1992, pp. 830-849.

[2] J.-H. Chen & R. V. Cox, " Convergence and Numerical Sensitivity of Backward-Adaptive LPC Prediction " - IEEE Workshop on Speech Coding, 1993, pp. 83-84.

[3] E. Ordentlich, Y. Shoham, " Low-Delay Code-Excited Linear-Predictive Coding of Wideband Speech at 32 kbps " - ICASSP, 1991, pp. 9-12.

[4] O. Gottesman & Y. Shoham, " Real -Time Implementation of High-Quality 32 kbps Wideband LD-CELP Coder " - Proc. of EUROSPEECH '93, pp. 1115-1118.

[5] J.-H. Chen, Y.-C. Lin & R. V. Cox, " A Fixed - Point 16 kb/s LD-CELP Algorihm " - ICASSP, 1991, pp. 21-24.

[6] T. P. Barnwell, " Recursive Windowing for Generating Autocorrelation Coefficients for LPC Analysis " - IEEE Trans. on Acous., Speech, and Sig. Proc., vol. 29, n° 5, 1981, pp. 1062-1066.

[7] Y. Mahieux & J. P. Petit, " Transform Coding of Audio Signals at 64 kbit/s " - Globecom, 1990, pp. 518-522.

[8] Y. Mahieux & J. P. Petit, " High-Quality Audio Transform Coding at 64 kbps " - IEEE Trans. on Comm., vol. 42, n° 11, 1994, pp. 3010-3019.

[9] G. Davidson, M. Yong & A. Gersho, " Real-Time Vector Excitation Coding of Speech at 4800 bps " - ICASSP, 1987, pp. 2189-2192.