CODAGE PAR TRANSFORMEE DE LA PAROLE A BANDE ELARGIE (0 à 7 kHz).

Dia H.[1], Feng G.[1] & Mahieux Y.[2]

[1] Institut de la Communication Parlée, INPG/ENSERG Université Stendhal
URA CNRS N° 368, BP 25X, 38400 Grenoble Cedex, France
[2] CNET Lannion A, TSS/CMC, 22301 Lannion, France

RÉSUMÉ

La transmission de la parole à bande élargie (0 à 7 kHz) occupe une place de plus en plus importante dans les systèmes de télécommunications modernes tels le visiophone et les systèmes de téléconférence. La qualité de transmission atteinte par la norme G722 du CCITT est suffisante mais le débit de 64 kbit/s est trop élevé. Récemment, le codage par transformée a montré d'excellentes performances pour la transmission des signaux sonores de haute-fidélité (Brandenburg & al., 1991). Ces bons résultats ont suggéré l'application de cette technique au codage de la parole à bande élargie. Il est alors intéressant de prendre en compte la nature voisée/non-voisée (V/NV) du signal de parole au cours de l'élaboration de l'algorithme. Dans cette communication, nous présentons un codeur à débit fixe (32 kbit/s) et un codeur à débit variable ayant un débit moyen de 24 kbit/s. Les résultats obtenus lors de l'évaluation subjective ont montré que la qualité de ces deux codeurs atteint celle de la norme G722 (64 kbit/s).

ABSTRACT

Wideband speech coding (0 to 7 kHz) finds a broad range of applications in modern telecommunication systems such as video-phone and teleconferencing. The quality delivered by the G722 CCITT standard is sufficient, but the bit-rate (64 kbit/s) is too high. Recently, transform coding has made great performances in high-fidelity transmission of audio signals (Brandenburg & al., 1991). This technique has prompted us to use it for wideband speech coding. In order to obtain a high quality of coded speech, the voiced/unvoiced (V/UV) nature of speech has been taken into account at each stage of the algorithm development. In this paper, a fixed bit-rate coder at 32 kbit/s and a variable bit-rate one with a mean bit-rate of 24 kbit/s are presented. Evaluation results about the subjective quality of both coders have shown that they perform as well as the standard G722 (64 kbit/s).

1. INTRODUCTION

Le codage de la parole à bande élargie (0 à 7 kHz) suscite un grand intérêt puisque le côté naturel et la richesse du signal de parole sont accrus par rapport à la bande téléphonique classique (0,3 à 3,4 kHz). La norme G722, un codeur en sous-bandes-MICDA, ne permet pas une transmission de haute qualité à un débit inférieur à 56 kbit/s. Deux techniques de codage peuvent atteindre ce but avec un débit inférieur ou égal à 32 kbit/s : le codage par transformée et le codage CELP. Dans cette communication, nous nous intéressons au codage par transformée.

Cette communication comporte deux parties essentielles. En section 2 un codeur à débit fixe de 32 kbit/s est détaillé. La section 3 traite du problème du débit variable. Les deux codeurs ont le même schéma de principe (fig. 1) qui est décrit ci-dessous. Il s'agit d'un codeur par transformée dont les paramètres ont été optimisés pour le codage du signal de parole.

Les signaux, échantillonnés à 16 kHz, sont traités par trames de 512 points. Chaque trame subit une transformation temps/fréquence à l'aide de la Transformée en Cosinus Discrète Modifiée (TCDM) (Princen & al., 1986). Une courbe de masquage est calculée à partir des coefficients de transformée (Dia & al., 1993). Les raies spectrales dont l'énergie se trouve en-dessous de cette courbe sont supprimées. Ceci permet l'élimination de 55 % des coefficients pour les sons voisés et 42 % pour les non voisés (moyennes obtenues sur un corpus de parole multi-locuteurs de 270 s). Pour coder les coefficients non masqués, une information auxiliaire comprenant la position des raies masquées et l'enveloppe spectrale est codée et transmise. Elle permet de calculer la même allocation des bits au codeur et au décodeur. Le débit nécessaire à la transmission de cette information auxiliaire a été réduit grâce à la séparation voisée/non-voisée et à un codage entropique. Pour masquer le bruit de quantification, les bits disponibles sont alloués selon un critère perceptuel. Les coefficients non masqués sont ensuite codés à l'aide de quantificateurs optimaux. Quatre jeux de quantificateurs sont disponibles en fonction de la nature voisée/non-voisée du signal et de l'efficacité du masquage.

2. CODEUR A DEBIT FIXE (32 kbit/s)

Dans cette partie, les différentes optimisations spécifiques au signal de parole sont présentées. Les parties de l'algorithme omises sont décrites dans Dia & al.(1992, 1993).

2.1 Codage de l'enveloppe spectrale

L'enveloppe spectrale est définie comme l'ensemble

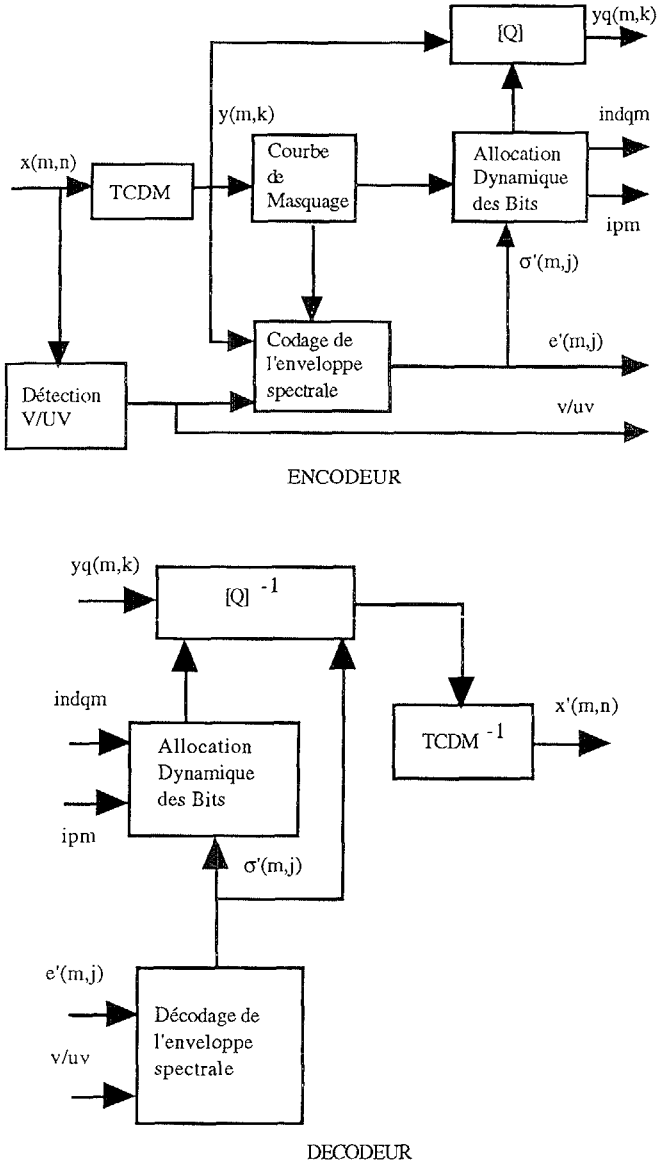

Fig. 1. Schéma de base du codeur. *indqm* indique si le masquage est effectif dans la bande j. *ipm* indique les positions des coefficients masqués dans la bande j.

$\{\sigma(m,j),\ j = 0,..., N_{sb}-1\}$ des écart-types $\sigma(m,j)$ des coefficients de transformée appartenant à la bande j. Elle permet de calculer la même allocation des bits à l'encodeur et au décodeur. Elle sert aussi de facteur de normalisation à la quantification des coefficients. Le spectre est découpé en N_{sb} (= 32) bandes de largeurs inégales. L'enveloppe spectrale est transmise grâce à un codage différentiel qui permet d'exploiter la redondance inter-trames (prédiction temporelle) ou intra-trame (prédiction fréquentielle).

L'erreur de prédiction temporelle est donnée par :

$$e(m,j) = \log(\sigma(m,j)) - a_1 \,.\, \log(\sigma'(m-1,j)) \qquad (1)$$

pour $j = 0,..., N_{sd} - 1$.

L'erreur de prédiction fréquentielle est donnée par :

$$e(m,j) = \log(\sigma(m,j)) - a'_1 \,.\, \log(\sigma'(m,j-1)) \qquad (2)$$

pour $j = 1,..., N_{sd} - 1$. $\sigma'(m,j)$ est la version quantifiée de $\sigma(m,j)$. a_1 et a'_1 les coefficients de prédiction sont proches de 1. En codage prédictif fréquentiel, $\sigma'(m,0)$ est quantifiée séparément sur 7 bits.

Les deux codeurs prédictifs sont appliqués en parallèle à l'enveloppe spectrale. Celui réclamant le moins de débit est retenu.

Une analyse statistique des distributions des erreurs de prédiction montre qu'un quantificateur de dynamique 80 dB et de pas 5 dB suffit au codage de l'enveloppe spectrale. La densité de probabilité (ddp) des mots de code en sortie du quantificateur est fortement non uniforme. Ce qui incite à l'utilisation d'un codage entropique (Huffman). Les ddp des mots de code sont différentes d'une bande à l'autre et varient selon la nature V/NV de la trame. Il est donc judicieux d'utiliser plusieurs codes de Huffman. La combinaison des différents cas conduirait au stockage de $2*2*N_{sb}-1 = 127$ tables. Ce nombre peut être réduit en regroupant les ddp, et en calculant un seul code de Huffman pour chaque groupe. La classification a été effectuée à l'aide d'un algorithme des K-moyennes utilisant la distance de Kullback (Kullback, 1959). Un regroupement en 8 classes, pour chaque condition (V/NV, T/F) conduisant au total à 32 codes de Huffman, s'est avéré suffisant. La table 1 met en évidence l'intérêt de la prise en compte de la séparation V/NV. En moyenne 85 bits par trame, i.e 16.6 % du débit total, suffisent au codage de l'enveloppe spectrale. (voir tab. 2)

Prédiction	Temporelle	Fréquentielle
Sans séparation	2.62	2.76
Voisée/non-voisée	2.59	2.72
Voisée/non-voisée + 8 classes	2.54	2.61

Tab. 1. Nombre de bits moyen par bande pour le codage de l'enveloppe spectrale.

2.2 Allocation dynamique des bits

L'allocation des bits est basée sur la minimisation, pour chaque coefficient, du rapport de la variance du bruit de quantification à la courbe de masquage associée (Johnston, 1988). Cette minimisation s'effectue sous la contrainte du nombre de bits disponibles. Ainsi, le nombre de bits distribués pour chaque coefficient R(m,j) appartenant à la bande j est calculé comme suit (Mahieux & Petit, 1990) :

$$R(m,j) = 0.5 \log_2 (\sigma'^2(m,j) / S_1(m,j)) + \lambda(m) \qquad (3)$$

où $\sigma'(m,j)$ est la version quantifiée de l'enveloppe spectrale. $S_1(m,j)$ est la courbe de masquage calculée à partir de l'ensemble des $\sigma'(m,j)$. $\lambda(m)$ est un terme dépendant du nombre de bits à distribuer.

L'allocation des bits est effectuée en deux étapes. Les bits sont d'abord alloués selon la règle ci-dessus sans distinguer la nature

masquée/non-masquée des coefficients. Ensuite, les bits alloués aux coefficients masqués sont collectés afin d'indiquer leurs positions dans chaque bande. Le surplus de bits est distribué aux coefficients non-masqués selon la même règle ci-dessus. Si le nombre de bits collectés dans une bande n'est pas suffisant pour indiquer les positions des raies masquées, le masquage est annulé et ces coefficients seront codés.

Expérimentalement, nous avons remarqué que pour certaines trames voisées, un grondement est perceptible en basses fréquences. Ceci est dû à l'imprécision de la courbe de masquage calculée à partir de l'enveloppe spectrale (voir fig. 3.b et fig. 3.c). Pour éviter ce phénomène, un nombre de bits fixe (5 bits par coefficient) est alloué aux 12 premiers coefficients.

2.3 Quantification optimale

Dans chaque bande j, c'est le rapport du coefficient y(m,k) à la valeur de l'enveloppe $\sigma'(m,j)$ qui est quantifié. Des quantificateurs optimaux (Max, 1960) ont été calculés à partir des densités de probabilité expérimentales. L'examen de ces ddp montre des variations importantes selon la nature V/NV de la trame et selon le pourcentage de raies masquées dans chaque bande. Lorsque le masquage est effectif (nombre de raies masquées non nul dans la bande), l'écart-type de la variable aléatoire $x = y(m,k) / \sigma'(m,j)$ est moins élevé que dans le cas inverse (0,58 en fig. 2.a et 0,36 en fig. 2.b).

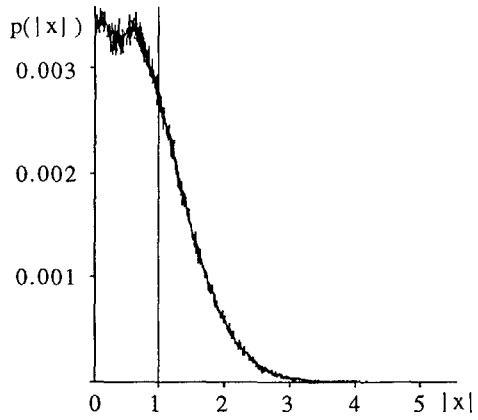

Fig. 2.a. Exemple de densité de probabilité pour y(m,k) recevant 2 bits. Cas voisé et masquage non effectif.

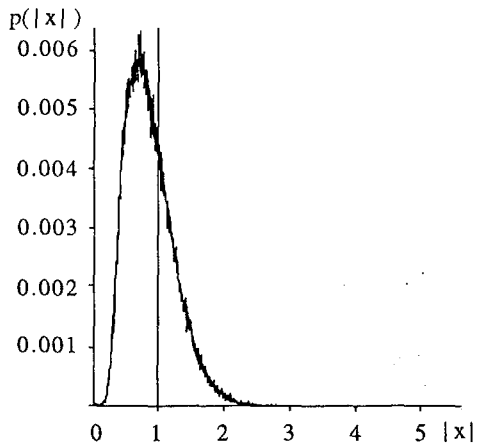

Fig. 2.b. Idem à 2.a mais cas voisé et masquage effectif.

On a donc distingué 4 cas pour le calcul des quantificateurs optimaux (V/NV, masquage-effectif/non-effectif). Ces quantificateurs ont été obtenus en résolvant les équations de Llyod-Max par un algorithme du gradient. Pour chaque nombre de bits (de 1 à 6), c'est la ddp expérimentale calculée à partir des coefficients recevant ce nombre de bits qui est considérée.

Un exemple illustrant les étapes principales de l'algorithme est décrit sur la figure 3.

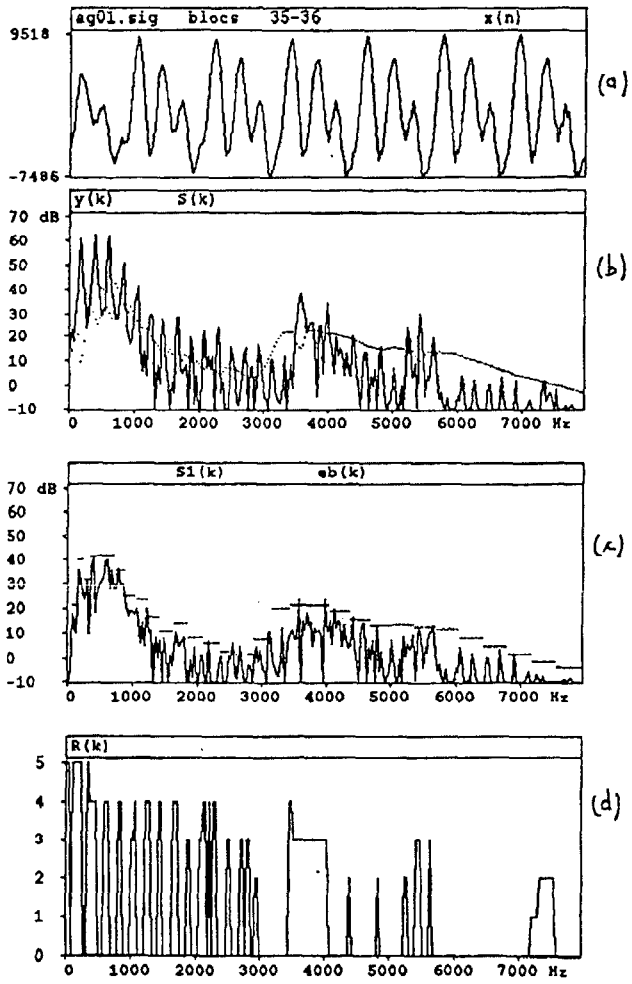

Fig. 3. Exemple des principales étapes de l'algorithme du codeur à débit fixe, pour une trame voisée. De haut en bas : (a) 512 échantillons du signal en entrée (32 ms), (b) coefficients de la TCDM et courbe de masquage calculée pour chaque coefficient (en pointillés), (c) courbe de masquage calculée à partir de l'enveloppe spectrale (en pointillés), et spectre du bruit de quantification, (d) nombre de bits alloués selon l'équation (3).

synchro-trame	V/NV	prédiction T/F	enveloppe spectrale	indications de masquage des bandes	positions des coefficients masqués	coefficients non-masqués
4	1	1	85	28	52	341

Tab. 2. Répartition (moyenne) du débit par trame (total : 512 bits).

3. CODEUR A DEBIT VARIABLE

L'algorithme du codeur à débit variable repose sur le même principe que celui du codeur à débit fixe. Mais la procédure d'allocation des bits n'inclut pas de contrainte de débit, c'est-à-dire le terme $\lambda(m)$ dans l'équation (3). Le débit moyen vaut alors 28 kbit/s. Cependant, la qualité est dégradée pour les trames voisées dont les coefficients en basses fréquences ne reçoivent pas assez de bits. Pour y remédier, une manière simple est de tenir compte de l'indice de tonalité calculé à partir de la mesure d'étalement spectral (Johnston, 1988). Les bits sont distribués selon la règle suivante :

$$R(m,j) = 0.5 \log_2 (\sigma'^2(m,j) / S_2(m,j)) \qquad (4)$$

avec $S_2(m,j) = \alpha(m) + (1-\alpha(m)) \, . \, S_1(m,j)$.

$\alpha(m)$ est l'indice de tonalité au bloc m. Il est transmis au décodeur à l'aide d'une quantification linéaire sur 6 bits. Le débit obtenu est alors de 34 kbit/s (écart-type de 6 kbit/s) et la qualité est jugée suffisante.

Ce débit moyen trop élevé peut être réduit en procédant à un codage différent lors des trames de silence. Nous supposons que dans une communication interactive normale, le silence occupe 40 % du temps total, un débit moyen de 24 kbit/s peut être atteint si, les trames de silence sont codées avec moins de 12 kbit/s.

Il est observé pendant ces trames de silence, que le signal à coder est assimilable à du bruit. Or l'oreille n'est pas sensible à l'information de phase pour ce type de son (Zwicker & Feldtkeller, 1981). Seules la répartition fréquentielle de l'énergie et la largeur de bande sont importantes. Le procédé retenu dans le codeur à débit variable consiste au décodeur, à moduler de manière aléatoire l'enveloppe spectrale dans chaque bande de fréquence. Les coefficients sont générés au décodeur en pondérant $\sigma'(m,j)$ par un générateur de nombres aléatoires à moyenne nulle ayant le même écart-type (0,88) que celui de la variable $x = y(m,k) / \sigma'(m,j)$. Il suffit de transmettre l'enveloppe spectrale, d'où une réduction importante du débit.

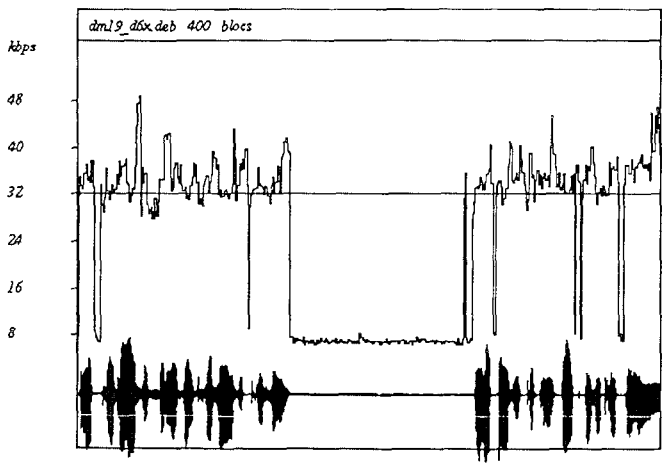

Fig. 4. Débit variable pour une paire de phrases de 3.84 s.

Pour une bonne qualité lors des transitions "séquence parlée" à silence, cette synthèse aléatoire des coefficients n'est pas appliquée aux basses fréquences. Les coefficients (< 1 kHz) sont codés par le codeur à débit variable.

Le débit moyen mesuré dans les trames de silence vaut 9,3 kbit/s. Le débit moyen obtenu pour chaque phrase est de 24 kbit/s. Un exemple illustrant l'évolution du débit pour une paire de phrases entrecoupées de silence est donné en figure 4.

4. EVALUATION

Des tests subjectifs formels ont été effectués pour comparer la performance des deux codeurs par rapport à celle du codeur de la norme G722. Ces tests de dégradation par paires consiste à présenter aux auditeurs des paires A-B, où A est le signal original et B le signal codé. B est noté par rapport à A sur l'échelle de dégradation (5 à 1) du CCIR. Le corpus utilisé est composé de 6 phrases (3 hommes et 3 femmes). Le codeur à débit fixe (32 kbit/s) a obtenu un score de 4,51 (sur 5) avec un écart-type de 0,63. Le score de 4,58 a été obtenu par le codeur à débit variable (écart-type de 0,59) et par le G722 à 64 kbit/s (écart-type de 0,63). Ces notes mettent en valeur les bonnes performances des deux codeurs par transformée.

5. CONCLUSION

Dans cette communication, deux codeurs ont été présentés : un codeur à débit fixe à 32 kbit/s et un autre à débit variable avec un débit moyen de 24 kbit/s. Ces deux codeurs ont les mêmes performances que la norme G722 à 64 kbit/s. Ces résultats montrent que le codage par transformée avec intégration des propriétés perceptuelles et optimisé en fonction des caractéristiques propres au signal de parole, est une technique adéquate pour la transmission de la parole à bande élargie (0 à 7 kHz) avec un débit moitié de celui de la norme G722.

REMERCIEMENTS

Cette étude a été partiellement financée par la convention CNET/ICP n° 90 7B 051/CMC.

REFERENCES

Brandenburg K., Herre H., Johnston J., Mahieux Y. & Schroeder E. (1991), Adaptive perceptual entropy coding of high quality music signals. Proceedings of the 90th AES convention, Paris, Preprint 3011, 1-11.

Dia H., Achab N. & Feng G. (1992), "Codage par transformée et segmentation automatique : vers un codeur à débit variable pour la parole à bande élargie (0 à 7 kHz)", 19es J.E.P, Bruxelles, 415-420.

Dia H., Feng G. & Mahieux Y. (1993), A 32 kbit/s wideband speech coder based on transform coding. Eurospeech' 93.

Johnston J.D. (1988), Transform coding of audio signals using perceptual noise criteria. IEEE Journal on selected areas in communications, vol. 6, n° 2, 314-323.

Kullback S. (1959), "Information theory and statistics". John Willey & sons Inc., New York.

Mahieux Y. & Petit J. (1990), Transform coding of audio signals at 64 kbits/s, Globecom'90, San Diego, 518-522.

Max J. (1960), "Quantizing for minimum distortion", IRE Trans. Inform. Theory, vol. IT-6, 7-12.

Princen J.P. & Bradley A. (1986), Analysis/synthesis filter bank design based on time domain aliasing cancellation, IEEE Trans. ASSP., vol. 34, n°5, 1153-1161.

Zwicker E. & Feldtkeller R. (1981), "Psychoacoustique, l'oreille réceptrice d'information", Ed. Masson.