Quantification adaptative des coefficients LSF pour le codage de la parole à bas débit

M. Djamah, M. Boudraa, B. Boudraa, M. Bouzid

Laboratoire Communication Parlée et Traitement de signal (LCPTS)
Faculté d'Electrique et d'informatique, USTHB, BP32, EL-ALIA, ALGERIE
Mél: mouloudjamah@yahoo.fr

ABSTRACT

This paper describes a variable-rate linear predictive coding (LPC) parameter quantization, in which the LSF coefficients quantization is done by using either the scalar quantization or the vector quantization using a adaptive codebook containing a previous LSF coefficients. The quantization sheme is integrated in to a basic CELP coder and a objective evaluation is done in order to evaluate the speech quality according the average bit-rate.

1. INTRODUCTION

Le codeur CELP [1, 2, 3] utilisé dans cette étude utilise une fréquence d'échantillonnage de 8 KHZ et une trame de 30 ms divisée en quatre sous-trames de 7.5 ms. Le signal synthétique est obtenu par le passage d'un signal d'excitation à travers un filtre de prédiction linéaire 1/A(z). L'excitation est calculée par sous-trame. Elle est le résultat de l'addition de deux excitations élémentaires, la première est un vecteur extrait du dictionnaire adaptatif (composé de 128 excitations passées) à l'indice i_a et pondéré par un gain g_a, la deuxième est un vecteur extrait du dictionnaire stochastique (composé de 512 excitations ternaires) à l'indice i_s et pondéré par un gain g_s. L'analyse du codeur CELP consiste à trouver les paramètres d'excitation (les indices et les gains) de manière à minimiser l'erreur perceptuelle entre le signal de parole originale et le signal synthétisé. Ce codeur correspond à un codeur à 4,8Kb/s. L'allocation des bits des différents paramètres à transmettre est donnée à la table 1.

Table 1 : Allocation des bits.

Paramètres /Trame	Bits/Trame	Débit (bits/s)
10 LSF	34	1 133.33
4 x (i_a, g_a)	4 x (7 + 5)	1 600
4 x (i_s, g_s)	4 x (9 + 5)	1 866.67

L'analyse par prédiction linéaire à court terme est réalisée une fois par trame de parole. Elle comporte l'estimation de 10 coefficients de prédiction a_i sur des trames de 30 ms (pondéré par la fenêtre de Hamming) et une expansion en largeur de bande de 15HZ. Cela consiste à remplacer les coefficients de prédiction a_i par les coefficients $a'_i = a_i \gamma^i$ (γ =0.994) [3]. Ces derniers définissent le filtre de prédiction linéaire 1/A(z).

$$A(z) = 1 + \sum_{k=1}^{P} a'_k z^{-k} \qquad (1)$$

Les coefficients a'_i sont convertis en coefficients LSF (Line Spectral Frequency) qui sont plus adaptés à la transmission.

2. INTERPOLATION DES COEFFICIENTS LSF

Les valeurs des coefficients LSF utilisés dans la détermination de l'excitation des sous-trames, sont obtenues par interpolation linéaire de deux ensembles de coefficients LSF (calculés pour deux trames d'analyse successives n et n+1) pour former un ensemble intermédiaire pour chacune des 4 sous-trames de la trame à coder (figure 1). L'interpolation linéaire utilise la pondération de la table 2 [4].

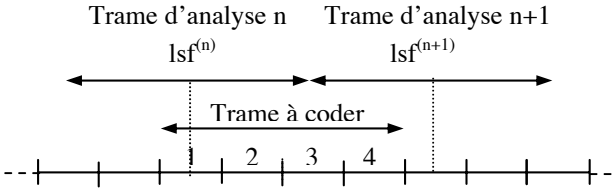

Figure 1 : Interpolation des LSF

Table 2 : Facteurs de pondération.

Numéro sous-trame	$lsf^{(n)}$	$lsf^{(n+1)}$
1	1	0
2	0.75	0.25
3	0.50	0.50
4	0.25	0.75

La trame d'analyse désigne la trame de parole (de 30ms) qui est pondérée par la fenêtre de Hamming et à partir de laquelle les coefficients LSF sont calculés. La trame à coder désigne la trame de parole à partir de laquelle le codage CELP est effectué. Notons que le centre de la trame d'analyse est aligné avec le centre de la première sous-trame de la trame à coder, ce qui implique que les coefficients LSF calculés à partir de cette trame d'analyse sont ceux de la première sous-trame, et les coefficients des autres sous-trames (numéro 2, 3 et 4) sont obtenus par interpolation entre les coefficients LSF calculés à partir de la trame d'analyse courante et de la trame d'analyse suivante [4]. La figure 2 montre les deux

premières trames à coder. Chaque trame est divisée en 4 sous-trames et chaque sous-trame est représentée par son propre ensemble de coefficients LSF calculé selon l'équation 2. Notons que dans cette figure le symbole "X" représente les coefficients LSF originaux et le symbole "0" représente les coefficients LSF interpolés.

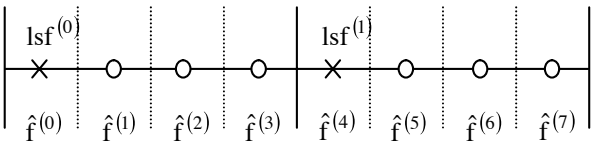

Figure 2 : Exemples de vecteurs LSF interpolés correspondants à huit sous-trames successives.

$$\hat{f}^{(i)} = \begin{cases} lsf^{\left(\frac{i}{4}\right)} & si\ (i \bmod 4) = 0 \\ \alpha_i\, lsf^{\left(\left[\frac{i}{4}\right]\right)} + (1-\alpha_i) lsf^{\left(\left[\frac{i}{4}\right]+1\right)} & ailleurs \end{cases} \quad (2)$$

$$\text{Avec :}\ \ \alpha_i = \frac{4-(i \bmod 4)}{4} \qquad i = 0,....,4N_F - 1 \quad (3)$$

Où N_F est le nombre de trames et 4 est le nombre de sous- trames par trames. [ ] : Désigne la partie entière.

3. MESURE DE DISTORSION

Pour évaluer la qualité de la parole synthétisée, la mesure la plus couramment utilisée dans le domaine temporel est le rapport signal-sur-bruit segmental SNRseg. Le signal est découpé en N_F segments de N échantillons chacun et on calcule une moyenne (s(n) est le signal de parole original et ŝ(n) le signal synthétisé) :

$$SNRseg = \frac{1}{N_F} \sum_{i=0}^{N_F-1} 10 \log_{10} \left(\frac{\sum_{j=0}^{N-1} s(Ni+j)^2}{\sum_{j=0}^{N-1} \left(s(Ni+j) - \hat{s}(Ni+j)\right)^2} \right) \quad (dB) \quad (4)$$

La distorsion spectrale est une autre mesure objective dans le domaine spectrale. La distorsion spectrale pour une trame i est définie par [4]:

$$SD_i = \sqrt{\frac{1}{F_s} \int_0^{F_s} \left[10 \log_{10} \left(\frac{S_i(f)}{\hat{S}_i(f)} \right) \right]^2 df} \ (dB) \quad (5)$$

Où F_s est la fréquence d'échantillonnage, $S_i(f)$ et $\hat{S}_i(f)$ sont les densités spectrales LPC (estimées à partir des coefficients de prédiction) de la trame numéro i et sont donnés par :

$$S_i(f) = \frac{1}{\left| A_i(e^{j2\Pi f / F_s}) \right|^2} \quad (6)$$

$$\hat{S}_i(f) = \frac{1}{\left| \hat{A}_i(e^{j2\Pi f / F_s}) \right|^2} \quad (7)$$

Où $A_i(z)$ et $\hat{A}_i(z)$ sont les polynômes correspondants respectivement aux coefficients de prédiction originaux (équation 1) et aux coefficients de prédictions quantifiés pour la $i^{ème}$ trame de parole. En passant à la notation numérique, l'équation (5) devient [4] :

$$SD_i = \sqrt{\frac{1}{n_1 - n_0} \sum_{n=n_0}^{n_1-1} \left[10 \log_{10} \left(\frac{S_i\left(e^{j2\Pi n / N}\right)}{\hat{S}_i\left(e^{j2\Pi n / N}\right)} \right) \right]} \ (dB) \quad (8)$$

Les densités spectrales sont évaluées par DFT (Transformée de Fourier Discrète) en utilisant l'algorithme FFT. n_0 et n_1 correspondent respectivement à 0 et 97, si la distorsion SD_i est calculée dans l'intervalle de fréquence allant de 0Hz à 3kHz (le signal est échantillonné à 8kHz et on utilise N=256 points pour le calcul de la DFT, la résolution entre deux points est donc de 8kHz/256 = 31,25 Hz).

La distance euclidienne pondérée est une mesure qui utilise les coefficients LSF. En effet les coefficients LSF ont une relation avec les formants et les valets de l'enveloppe spectrale calculée par l'équation (6). Si f et $\hat{f}$ sont deux vecteurs composés chacun de P coefficients LSF originaux et quantifiés respectivement, alors leurs distance euclidienne pondérée et définie par [4]:

$$d(f, \hat{f}) = \sum_{i=1}^{P} c_i w_i (f_i - \hat{f}_i)^2 \quad (9)$$

w_i est le poids assigné au $i^{ème}$ coefficient LSF, défini par Paliwal et Atal [5] par:

$$w_i = \left| S(f_i) \right|^r \quad (10)$$

Où $S(f_i)$ est la densité spectrale LPC estimée à la fréquence f_i et associée au vecteur à tester. r est une constante empirique qui contrôle le poids des différents coefficients LSF et est déterminée expérimentalement. Paliwal et Atal ont trouvé que r=0.15 est une valeur satisfaisante. Donc la pondération dépend de la valeur de la densité spectrale LPC à la fréquence du coefficient LSF considérée. L'oreille humaine est plus sensible aux basses fréquences qu'aux hautes fréquences. Pour exploiter cette caractéristique, Paliwal et Atal [5] ont introduit un coefficients de pondération constant défini par :

$$c_i = \begin{cases} 1.0, & 1 \le i \le 8 \\ 0.8, & i = 9 \\ 0.4, & i = 10 \end{cases} \quad (11)$$

4. QUANTIFICATION ADAPTATIF DES LSF

Dans notre codeur CELP, un ensemble de P coefficients LSF (où P=10 est l'ordre de la prédiction linéaire) sont calculés à chaque passage d'une trame de parole (de 30ms) à l'autre. Ce vecteur de 10 coefficients LSF est comparé en utilisant la distance euclidienne pondérée (équation 9) à tous les vecteurs d'un dictionnaire constitué de L vecteurs LSF. Si 'index' est l'indice du vecteur le plus ressemblant au vecteur testé, alors la distorsion spectrale SD_{index} (équation 8) correspondant à ce vecteur est calculée. Si SD_{index} est inférieur ou égale à un seuil T prédéfini alors une quantification vectorielle est effectuée c'est-à-dire que le vecteur issu du dictionnaire à l'indice 'index' remplacera le vecteur d'origine et l'indice 'index' est transmis au décodeur. Si SD_{index} est supérieur au seuil T une quantification scalaire, selon la norme FS1016 [3], est réalisée sur le vecteur LSF d'origine. Avant de passer à la trame de parole suivante, le dictionnaire LSF est actualisé en remplaçant les 4 plus anciens vecteurs par 4 vecteurs LSF calculés par interpolation entre le vecteur LSF quantifié de la trame courante et celle de la trame précédente (on parle alors de dictionnaire LSF adaptatif). Pour plus de détail, on donne ci-dessous l'algorithme correspondant au traitement de la trame numéro n+1 (voir figure 1). Notons que la trame d'analyse et la trame à coder sont deux trames différentes ce qui est montré à la figure 1.
À l'encodeur, on effectue les étapes suivantes:

1) Calculer les coefficients LSF correspondant à la trame d'analyse n+1: $lsf^{(n+1)} = [lsf_0^{(n+1)}, lsf_1^{(n+1)}, ..., lsf_P^{(n+1)}]$.

2) Calculer les distances $D(i) = d(lsf^{(n+1)}, \hat{f}_a^{(i)})$ i=0,..,L-1 (en utilisant l'équation 9) où les $\hat{f}_a^{(i)}$ sont les vecteurs du dictionnaire LSF adaptatif composé de L vecteurs.

3) Sélectionner le vecteur du dictionnaire adaptatif correspondant à la distorsion minimale c'est-à-dire le vecteur $\hat{f}_a^{(index)}$ avec $index = \arg\min(D(i))$ i=0,…,L-1.

4) Calculer la distorsion spectrale du vecteur sélectionné $SD_{index} = d(lsf^{(n+1)}, \hat{f}_a^{(index)})$ (équation 8).

5) Si $SD_{index} \leq T$ (T est un seuil prédéfini) alors faire $lsf_q^{(n+1)} = f_a^{(index)}$ puis calculer 4 vecteurs LSF par interpolation des vecteurs $lsf_q^{(n)}$ et $lsf_q^{(n+1)}$ (figure 1). $lsf_q^{(n)}$ est le vecteur LSF quantifié correspondant à la trame d'analyse numéro n. L'indice 'index' est transmis au décodeur. Un bit d'état, mis à zéro, est transmis au décodeur pour l'informer qu'une quantification vectorielle a été effectuée à l'encodeur. Aller à l'étape 7.

6) Si $SD_{index} > T$ alors faire une quantification scalaire du vecteur $lsf^{(n+1)}$ ce qui donne le vecteur $lsf_q^{(n+1)}$ puis calculer 4 vecteurs LSF par interpolation des vecteurs $lsf_q^{(n)}$ et $lsf_q^{(n+1)}$. Le vecteur $lsf_q^{(n+1)}$ est transmis au décodeur. Un bit d'état, mis à 1, est transmis au décodeur pour l'informer qu'une quantification scalaire a été effectuée à l'encodeur. Aller à l'étape 7.

7) Les 4 vecteurs calculés par interpolation à l'étape 5 ou à l'étape 6 sont utilisés pour le codage des 4 sous-trames de la trame de parole à coder. Le dictionnaire LSF adaptatif de l'encodeur est actualisé en décalant les vecteurs de 4 positions vers le bas et en insérant les 4 vecteurs interpolés aux indices 0, 1, 2 et 3.

Au décodeur, on effectue les étapes suivantes :

1) Si une quantification vectorielle a été effectué au niveau de l'encodeur (le bit d'état reçu est à zéro) alors un vecteur d'indice 'index' (reçu de l'encodeur) est sélectionné à partir du dictionnaire LSF adaptatif. Faire $lsf_q^{(n+1)} = f_a^{(index)}$ et comme pour l'encodeur, 4 vecteurs LSF sont calculés par interpolation des deux vecteurs $lsf_q^{(n)}$ et $lsf_q^{(n+1)}$. Aller à l'étape 3

2) Si une quantification scalaire a été effectué au niveau de l'encodeur (le bit d'état reçu est à 1) alors calculer 4 vecteurs LSF par interpolation des deux vecteurs $lsf_q^{(n)}$ et $\hat{f}_q^{(n+1)}$ (reçu de l'encodeur). Aller à l'étape 3.

3) Les 4 vecteurs calculés par interpolation à l'étape 1 ou à l'étape 2 sont utilisés pour le décodage d'une trame de parole. Le dictionnaire adaptatif du décodeur est actualisé en décalant les vecteurs de 4 positions vers le bas et en insérant les 4 vecteurs interpolés aux indices 0, 1, 2 et 3.

Notons que dans le travail présenté par Jozsef et al [6], le dictionnaire LSF est actualisé en calculant par prédiction linéaire (à chaque passage d'une trame à l'autre) 4 vecteurs de coefficients LSF à partir des trames de parole synthétisées passées. Ici l'approche est différente puisque les 4 vecteurs sont calculés par interpolation (les calculs sont plus rapides) à partir de deux vecteurs LSF quantifiés calculés à partir de la parole d'origine.

5. EVALUATION DES RÉSULTATS

L'algorithme proposé a été intégré dans un codeur CELP de base. Pour évaluer la qualité de la parole reconstruite, nous avons synthétisé environ 143s de parole phonétiquement équilibrée prononcé par 3 locuteurs masculins et 3 locuteurs féminins où chaque locuteur a prononcé 10 phrases [7]. Notons que nous avons effectué un rééchantillonnage à 8kHz puisque au départ le corpus de parole a été enregistré avec une fréquence de 10kHz. Nous avons utilisé le rapport signal-sur -bruit segmental SNRseg (équations 4) et la distorsion spectrale (équation 8) pour évaluer les performances de l'algorithme proposé en fonction du débit du codeur. Le nombre de bits utilisés pour le codage des 10 coefficients LSF est variable d'une trame à l'autre selon qu'on utilise une quantification scalaire (34 + 1 bits) ou une quantification vectorielle ($\log_2$ (L) + 1 bits avec L le nombre de vecteur du dictionnaire LSF et '1' pour le bit d'état). On parle alors de codeur CELP à débits variable. Le débit moyen du codeur dépend du seuil T utilisé et du nombre de

vecteur L du dictionnaire adaptatif. La figure 3 montre la distribution des indices du dictionnaire à transmettre lorsque le dictionnaire LSF adaptatif est composé de 64 vecteurs.

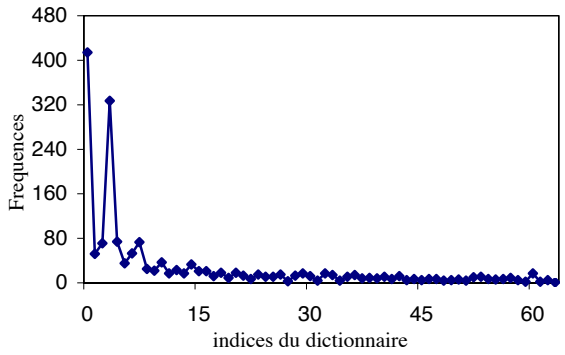

Figure 3 : Distribution des indices du dictionnaire LSF adaptatif composé de 64 vecteurs. 2294 trames de parole ont été traitées et 1740 indices ont été calculés. Le seuil T est pris égal à 4.5dB.

Dans la plus part des cas, les vecteurs LSF sélectionnés correspondent aux trames de parole les plus récentes (vecteurs aux indices bas) par rapport à la trame traitée. Ce qui implique qu'il n'est pas nécessaire de prendre un dictionnaire à très grande taille. Dans notre cas la taille du dictionnaire LSF adaptatif sera de 128 vecteurs. Les résultats obtenus pour le traitement d'un corpus de parole de 143s cité plus haut sont résumés dans la table 3 où on a fait varier le seuil T de 2 à 5 dB avec un pas de 0.5dB. On donne à la deuxième colonne de la table le débit moyen (en bits par trame) correspondant à chaque seuil T. Notons que l'augmentation du seuil T entraîne une utilisation plus fréquente du dictionnaire adaptatif ce qui diminuera le débit nécessaire à la transmission des coefficients LSF. Le seuil T=0dB correspond au cas où seul une quantification scalaire est utilisée pour les coefficients LSF. Ce cas servira de référence pour évaluer les performances du codeur CELP utilisant la quantification adaptatif des coefficients LSF.

Table 3 : Evaluation objective pour un dictionnaire LSF adaptatif de 128 vecteurs

			SD		
T [dB]	Débit LSF [bits/trame]	SNRseg [dB]	Moy [dB]	2-4 dB %	>4 dB %
0	34	8.72	1.39	8.4	0.0
2	27.4	8.77	1.47	7.6	0.0
2.5	23.5	8.70	1.62	25.1	0.0
3	20.3	8.69	1.84	39.7	0.0
3.5	17.4	8.61	2.09	52.6	0.0
4	14.9	8.55	2.41	64.5	0.0
4.5	13.4	8.48	2.67	60.8	10.3
5	12.3	8.45	2.94	57.3	19.5

La distorsion spectrale SD est utilisée pour évaluer la performance du quantificateur. La distorsion spectrale moyenne est calculée dans la quatrième colonne. Le nombre de trame (en pourcentage) ayant une distorsion spectrale dans l'intervalle 2-4dB (outlier type1) et supérieure à 4 dB (outlier type2) [4] est donnée respectivement à la colonne cinq et six. Notons que pour un seuil T = 2.5dB, le débit moyen pour la transmission des coefficients LSF est de 23.5 bits/trame au lieu de 34bits/trame, le rapport signal sur bruit SNRseg a diminué de 0.02dB, la distorsion spectrale moyenne a augmenté de 0.23dB et les 'outlier type1' ont augmenté de 16.7%.

6. CONCLUSION

Dans ce travail, nous avons introduit un quantificateur adaptatif des coefficients LSF dans un codeur CELP de base donnant lieu ainsi à un codeur CELP à débit variable. Une évaluation objective a été réalisée en fonction du débit moyen du codeur. Néanmoins une évaluation subjective, sur un corpus de parole plus important, de la qualité de la parole synthétisée est nécessaire. Nous envisageons d'expérimenter l'intégration de notre algorithme dans le codeur standard FS1016 [3]. Le dictionnaire LSF adaptatif pourrait être remplacé par un dictionnaire mixte où il y aurait une partie adaptative (telle que décrite dans cet article) et une partie fixe qui serait conçu en utilisant des algorithmes de quantification vectorielle tel que l'algorithme LBG.

BIBLIOGRAPHIE

[1] M.R Schroeder and B.S. Atal. Code-Excited Linear Prediction (CELP): High-quality speech at very low bit rates. *Proc. ICASSP*, pp. 937-940. March 1985.

[2] N.MOREAU. Codage prédictif du signal de parole à débit réduit: une présentation unifiée. *Annales des Télécom.*, vol. 46, n° 3-4, pp. 223-239, 1991.

[3] Fenichel R and Bodson D. Details to assist in implementation of Federal Standard 1016 CELP. *Technical Information Bulletin 92-1, National Communication system*, 1992.

[4] Tamanna Islam. Interpolation of Linear Prediction Coefficients for Speech Coding. *Master of Engineering, McGill University*, Canada,April 2000.

[5] K. K. Paliwal and B.S. Atal. Efficient vector quantization of LPC parameters at 24 bits/frame. *IEEE Trans, Speech and Audio Processing*, vol.1, pp.3-14, Jan. 1993.

[6] Jozsef Vass, Yunxin Zhao and Xinhua Zhuang. Adaptive Forward-Backward Quantizer for Low Bit Rate High-Quality Speech Coding. *IEEE Transactions on Speech and Audio Processing*, Vol.5, N0. 6, pp.552-557, Nov. 1997.

[7] M. Boudraa, B. Boudraa and B. Guerin. Mise en place de phrases arabes phonétiquement équilibrées. *XIX ème JEP*, Bruxelles, Mai 1992.