TREIZIÈME COLLOQUE GRETSI - JUAN-LES-PINS DU 16 AU 20 SEPTEMBRE 1991

APPLICATION DE LA QUANTIFICATION VECTORIELLE AU CODAGE BAS DEBIT D'UN SIGNAL DE PAROLE[1]

S. DEKETELAERE, H. LEICH, R. BOITE

Faculté Polytechnique de Mons, Service de Théorie des Circuits et Traitement du Signal, Bd Dolez 31, 7000 MONS

RESUME

Un nouveau schéma de codage des paramètres de prédiction linéaire (LPC), basé sur la quantification vectorielle des paires de raies spectrales (LSP), a été mis au point. Il permet une transmission des paramètres LSP, sans dégradation audible, à un débit binaire de 660 bits/s, voire de 330 bits/s avec une dégradation limitée.
Ces résultats sont obtenus par la définition d'une nouvelle distance perceptuelle adaptée aux LSP et d'une nouvelle procédure de création du dictionnaire dérivée de la procédure "K-MEANS".

ABSTRACT

A new coding scheme of Linear Predictive Coefficients (LPC), based on Line Spectrum Pairs (LSP) Vector Quantization is presented. It transmit LSP parameters, without audible degradation, at a data-rate of 660 bps. It can also transmit these datas at 330 bps with a small degradation.
These results are obtained with a new perceptual distance definition adapted to LSP and a new dictionnaries design procedure derived from "K-MEANS" algorithm.

1 INTRODUCTION

Le choix d'un codeur pour la transmission de la parole résulte toujours d'un compromis entre plusieurs paramètres antagonistes que sont la qualité de la parole décodée, le débit intermédiaire transmis sur la ligne et la complexité de la mise en oeuvre. Lorsque la classe de codeur que l'on désire réaliser se situe dans les codeurs à bas débit, la modélisation autorégressive d'un signal de parole est bien adaptée. Elle fournit une bonne qualité de codage même à très bas débit et la complexité de mise en oeuvre en temps réel peut être maîtrisée par les possibilités actuelles des processeurs spécialisés de traitement de signal.

La mise au point d'un codeur bas débit de type vocodeur à prédiction linéaire soulève 2 problèmes : - Codage de l'excitation du filtre de synthèse - Codage des paramètres du filtre de synthèse de la modélisation autorégressive (Paramètres LPC). Le premier problème est actuellement très étudié. Des solutions efficaces existent (CELP, Harmonique, RELP, MBE,...). Il est cependant hors de notre propos.

Le second problème est traditionnellement résolu par la quantification scalaire des paramètres LPC sous une de leur forme particulière (parcors, cepstre, paires de raies spectrales (LSP)). Cependant notre préférence va aux paramètres LSP grâce à leur plus grande robustesse à la quantification et à l'interpolation.

Afin de diminuer la part du débit binaire affecté à la transmission des paramètres LPC, La quantification vectorielle (QV) de ceux-ci a été envisagée. Grâce à elle, appliquée aux paramètres LSP, un débit binaire de transmission de 660 bits/s (tranche de 30 ms codée sur 20 bits), soit une diminution proche de 50% par rapport à la quantification scalaire, est possible sans dégradation audible.

2 MODELISATION LSP

La modélisation autorégressive d'un signal de parole se fait par la sortie d'un filtre tout-pôles :

$$F(Z) = \frac{1}{A(Z)} = \frac{1}{1+\sum_{i=1}^{p} a_i Z^{-i}} \qquad (1)$$

où $[a_1, a_2, \ldots, a_p]^T$ sont les coefficients du filtre et p l'ordre de celui-ci. Si on définit des filtres autorégressifs, d'ordre j croissant, de la forme (1), les polynômes $A_j(Z)$ créés satisfont aux récurrences suivantes:

$$A_j(Z) = A_{j-1}(Z) + k_j Z^{-j} A_{j-1}(Z^{-1}), \quad j=1,\ldots,p$$

$$A_0(Z) = 1 \qquad (2)$$

$$[k_1, k_2, \ldots, k_p]^T : \textit{PARCORS}$$

On peut étendre l'ordre du filtre à (p+1) sans introduire aucune nouvelle information en choisissant le (p+1)ème coefficient PARCOR à la valeur +1 ou -1. l'équation (2) devient alors respectivement:

[1] Recherches subsidiées par la société S.A.I.T. ELECTRONICS, ch. de Ruisbroek 66, 1190 - BRUXELLES

$$P(Z) = A_{p+1} = A_p(Z) + Z^{-(p+1)} A_p(Z^{-1}) \quad (k_{p+1} = 1)$$
$$Q(Z) = A_{p+1} = A_p(Z) - Z^{-(p+1)} A_p(Z^{-1}) \quad (k_{p+1} = -1) \qquad (3)$$

Cette transformation est avantageuse car les polynômes P(Z) et Q(Z) sont respectivement symétrique et antisymétrique. Ils possèdent notamment les propriétés utiles suivantes :
- tous les zéros de P(Z) et Q(Z) sont sur le cercle unité ,
- les zéros de P(Z) et Q(Z) sont alternés sur ce même cercle unité. La première propriété permet d'exprimer ces zéros par $e^{j\omega_i}$, qui regroupés en un vecteur $\omega = [\omega_1, \omega_2, \ldots, \omega_p]^T$, caractérisent complètement le filtre autorégressif. C'est ce vecteur ω que l'on appelle paires de raies spectrales.

2.1 SIGNIFICATION PHYSIQUE

Le grand intérêt du recours aux paramètres LSP pour la représentation du filtre de synthèse de la modélisation autorégressive est la signification physique de ces derniers. En effet, il est maintenant bien connu [6] que la position des formants dans le spectre du signal de parole est liée à un rapprochement de deux paramètres LSP consécutifs dans le vecteur ω autour de la fréquence du formant considéré. Il est donc possible d'identifier grossièrement les zones auditivement importantes dans le spectre du signal de façon très aisée. Nous ferons usage de cette propriété pour la définition de la distance décrite au paragraphe suivant.

3 QUANTIFICATION DE LA DISTORSION

Lorsque la quantification vectorielle est utilisée pour la réalisation d'une chaîne de codage de la parole, le choix de la distance capable de différencier 2 configurations du conduit vocal provoquant 2 impressions auditives différentes est primordiale. Le critère objectif auquel on a habituellement recours est la distorsion spectrale définie par :

$$D(dB) = \frac{10}{\pi} \int_0^{f_{éch}} \log \frac{S_1(f)}{S_2(f)} df \qquad (4)$$

où D est la distorsion spectrale en décibels, $f_{éch}$ la fréquence d'échantillonnage et $S_1(f)$ et $S_2(f)$ les 2 enveloppes spectrales à comparer. Cette distorsion fournit une bonne corrélation avec l'impression auditive et sa forme mathématique permet des développements élégants dans le cas de la modélisation autorégressive (Distance Maximum de vraisemblance [1][7]). Néamoins, l'introduction d'une pondération spectrale dans la définition de la distorsion s'avère utile, notamment pour considérer l'effet psychoacoustique du masquage fréquentiel [4]. Ainsi, au critère (4), se substitue par exemple :

$$D(dB) = \frac{10}{\pi} \int_0^{f_{éch}} W(f) \log \frac{S_1(f)}{S_2(f)} df \qquad (5)$$

où W(f) est une fonction de pondération spectrale.

3.1 APPLICATION AUX LSP

Les paramètres LSP possèdent également une propriété de sélectivité [6]. Celle-ci se traduit par la localisation de la distorsion de l'enveloppe spectrale sur une faible partie de l'axe fréquentiel proche de la valeur du paramètre LSP perturbé. Grâce à cette propriété, les contributions des diverses raies spectrales à la distorsion sont en première approximation indépendantes.
La forme mathématique de la distance ne peut pas être choisie de manière quelconque. Elle doit permettre la recherche aisée du dictionnaire des centroïdes. La seule envisageable est du type euclidienne pondérée. Cependant, la pondération peut être fonction du vecteur à quantifier. Sa forme est :

$$d(\bar{X}, \bar{Y}) = \sum_{i=1}^{D} W_i(\bar{X}) (X_i - Y_i)^2 \qquad (6)$$

où X_i et Y_i sont les $i^{ème}$ composantes des vecteurs à comparer et $W_i(\bar{X})$ la pondération associée à la $i^{ème}$ composante.
Les pondérations sont suffisamment générales pour prendre en compte les effets psychoacoustiques variables dans le temps comme l'effet de masque fréquentiel puisque elles restent fonction du vecteur à coder. Nous les avons choisies égales au carré de l'inverse de la distance à la raie la plus proche:

$$W_i = \max\left(\frac{1}{(X_i - X_{i-1})^2}, \frac{1}{(X_{i+1} - X_i)^2}\right), \quad i = 1,..,D-1$$
$$W_0 = \frac{1}{(X_1 - X_0)^2} \qquad (8)$$
$$W_D = \frac{1}{(X_D - X_{D-1})^2}$$

La distance ainsi définie favorise une bonne quantification des zones formantiques importantes auditivement et accorde moins d'importance aux paramètres LSP dont la valeur se trouve, sur l'axe des fréquences, dans une zone correpondant à une vallée du spectre. Elle approche conceptuellement la distorsion définie par (5). Elle s'interprète physiquement comme la somme des rapports relatifs, mis au carré, de l'écart avec le LSP le plus proche, paramètre dont l'importance perceptuelle a été reconnue [5].

4 QUANTIFICATION VECTORIELLE

La quantification vectorielle (QV) code le vecteur représentatif du filtre autorégressif, $\bar{X} = [x_1, x_2, \ldots, x_n]^T$ dans la suite, comme un ensemble contrairement au codage scalaire qui quantifie chaque composante séparément l'une de l'autre. Exploitant ainsi les dépendances statistiques de la distribution de densité à n dimensions p(x) du signal, la QV est en principe supérieure à tous les autres schémas de codage [7][8]. La QV est entièrement définie par un ensemble de vecteurs $C = [\bar{C}_1, \bar{C}_2, \ldots, \bar{C}_L]$, où $\bar{C}_i$ est le $i^{ème}$ vecteur du "dictionnaire" C, et une partition disjointe R_i de l'espace à n dimensions, R^n, respectant la relation:

$$\bigcup_{i=1}^{L} R_i = R^n \qquad (8)$$

Supposons également que Q est l'opérateur de quantification qui assigne à un vecteur $\bar{X}$ le vecteur $\bar{C}_j$ du dictionnaire de manière telle que la partition R_j est donnée par l'ensemble:

$$\{ \bar{X} \mid Q(\bar{X}) = \bar{C}_j \} \qquad (9)$$

En supposant enfin que la distorsion introduite par le remplacement de $\bar{X}$ par $\bar{C}_j$ est donnée par $d(\bar{X}, \bar{C}_j)$, le dictionnaire est dit optimal s'il est tel que la distorsion globale moyenne $D = E(d(\bar{X}, \bar{C}_j))$ est minimale. Celle-ci s'écrit également :

$$D = \sum_{i=1}^{L} P_i D_i \qquad (10)$$

où P_i est la probabilité que $\bar{X}$ se trouve dans la partition de l'espace R_i et D_i est la distorsion moyenne calculée uniquement sur les vecteurs $\bar{X}$ tombant dans cette même partition.

La minimisation de D entraine [8] les conditions nécessaires mais non suffisantes :

$$1°)\quad Q(\overline{X}) = \overline{C}_i \text{ si } d(\overline{X},\overline{C}_i) \le d(\overline{X},\overline{C}_j),\ j\neq i,\ 1\le j\le L \qquad (11)$$

$$2°)\quad \text{minimisation de } D_i(\overline{C}_j) \rightarrow \overline{C}_j$$

Une troisième règle d'optimalité nettement moins connue, démontrée par A. GERSHO [9], et uniquement valable pour les grands dictionnaires (L -> ∞),énonce que les produits $P_i D_i$ doivent être constants pour toutes les classes si le dictionnaire est optimal. L'emploi de cette troisième règle permet d'introduire un contrôle indirect sur le seuil de la distorsion maximale de quantification comme la suite va nous le montrer.

4.1 ALGORITHME K-MEANS MODIFIE

L'utilisation des deux premières règles d'optimalités définit une procédure de création du dictionnaire optimal. C'est la procédure "K-moyenne" (K-MEANS) bien connue en reconnaissance de formes [7][8]. De cette procédure est déduit l'algorithme LBG [10] classiquement utilisé pour la recherche du dictionnaire optimal en traitement de la parole. La procédure K-MEANS se décompose en 3 phases distinctes :
- La phase d'initialisation qui consiste à choisir un dictionnaire à L centroïdes "quelconques"
- une phase de classification qui assigne tous les vecteurs $\overline{X}$ de la base de données d'entrainement avec un élément du dictionnaire C par la règle 1°)
- une phase d'optimisation du dictionnaire qui consiste à calculer de nouveaux centroïdes par la règle 2°).

Les deux dernières phases sont réitérées jusqu'à la stabilisation du dictionnaire. Cet algorithme cherche, légitimement, à minimiser la distorsion en moyenne pour tous les vecteurs d'entrainement mais ne permet aucun contrôle sur la répartition de celle-ci entre les divers vecteurs. Une nouvelle phase a été introduite à cet algorithme afin de mieux contrôler cette répartition. Elle se situe dans l'algorithme, après la phase d'optimisation. Son exécution se fait à chaque itération. Cette phase de "reclassement" se décompose de la façon suivante:

- calculer les produits $\alpha_i = P_i D_i$ pour toutes les classes
- classer les α_i par ordre croissant
- supprimer du dictionnaire les N premiers centroïdes résultant du classement des α_i et d'affecter les places libérées du dictionnaire à N nouveaux centroïdes construits comme étant égaux aux N derniers centroïdes du classement des α_i à une petite perturbation près.

Ce reclassement tend à uniformiser les α_i pour tous les centroïdes (3ème règle d'optimalité). Son évolution dans l'algorithme peut s'interpréter intuitivement. Lorsque α_i est grand, P_i ou D_i ou les deux sont grands. C'est-à-dire que soit ce centroïde est souvent choisi, soit la distorsion moyenne des vecteurs $\overline{X}$ associés au centroïde $\overline{C}_i$ est grande. Dans ce cas, l'ajout d'un nouveau centroïde "proche" de $\overline{C}_i$, permet après une ou deux itérations de stabilisation, de diminuer P_i et D_i, et donc α_i, puisque les vecteurs $\overline{X}$ sont maintenant associés à deux centroïdes au lieu d'un seul. Au contraire, lorsque α_i est petit, c'est que le centroïde est rarement choisi, ou bien que la distorsion moyenne D_i associée à ce dernier est faible. La suppression de ce centroïde entraîne la réaffectation des vecteurs $\overline{X}$ qui lui étaient associés et donc une augmentation de la distorsion de quantification pour ces vecteurs, mais puisque D_i était déjà faible, un autre centroïde est certainement proche et la réaffectation peut se réaliser sans une grande augmentation de la distorsion. Si P_i est faible, la suppression du centroïde n'est pas très gênante en moyenne puisque ce dernier est rarement choisi.
Le mécanisme de cette procédure introduit de façon indirecte le seuil de la distorsion maximale admissible. En effet, elle consiste à accepter une moins bonne quantification des vecteurs "trop bien" quantifiés pour diminuer la distorsion des vecteurs particulièrement mal codés.
Le choix du nombre de reclassement à chaque itération, N, est réalisé de manière empirique. Il est adopté le plus grand possible, mais de façon à ne pas faire remonter la distorsion globale D. Celle-ci peut maintenant augmenter car le déplacement "brutal" des centroïdes ne respectent plus les deux premières règles d'optimalités, suffisantes pour assurer la convergence. Cependant, si la valeur de N est choisie faible par rapport au nombre de centroïdes dans le dictionnaire (1 à 2 %), nous n'avons jamais constaté de remontée de la distorsion globale D. Si néanmoins ce phénomène devait se produire, il suffirait de supprimer cette phase de reclassement pour les itérations suivantes.

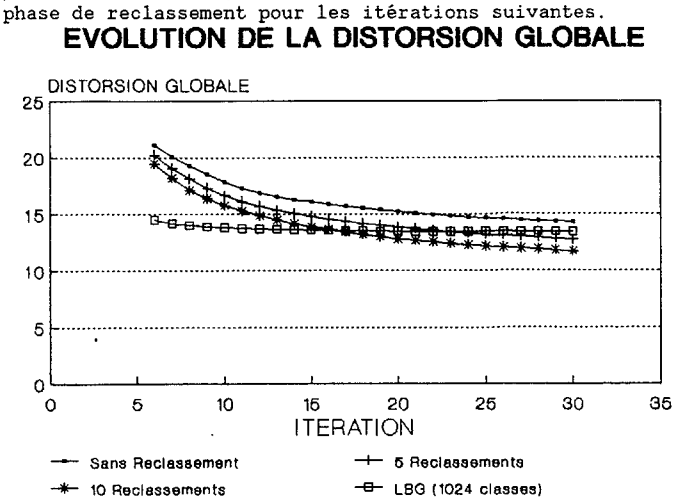

Figure 1 : Evolution de la distorsion globale

Les distorsions moyennes obtenues grâce à l'addition de cette phase sont consignés sur la figure 1. Ces résultats ont été obtenus avec la distance euclidienne pondérée définie par (6) et (7) et pour la création d'un dictionnaire à 1024 classes dont les vecteurs sont les 4 premiers paramètres LSP. On constate que la distorsion moyenne s'améliore légèrement avec l'augmentation du nombre de reclassement. Cependant, là n'est pas l'intérêt majeur de la nouvelle phase de reclassement. La figure 2 présente des résultats plus intéressants. Elle exprime la répartition des distorsions pondérées par leur fréquence d'apparition, α_i. On y constate que les vecteurs fort distordus par la quantification diminue très fortement lorsque le reclassement est inclus dans la recherche ; un effet de seuil maximal de la distorsion apparait.

4.2 STRUCTURE DU DICTIONNAIRE

La structure du dictionnaire que nous avons utilisée est du type Code produit [8]. Elle code les paramètres LSP en les regroupant dans 2 sous-vecteurs indépendants. Cette structure demande l'utilisation d'une distance séparable. La distance d'Itakura ne convient donc pas. Cependant, la distance euclidienne pondérée (6) convient parfaitement.

5 RESULTATS

La QV a été exploitée pour la réalisation de 2 schémas de

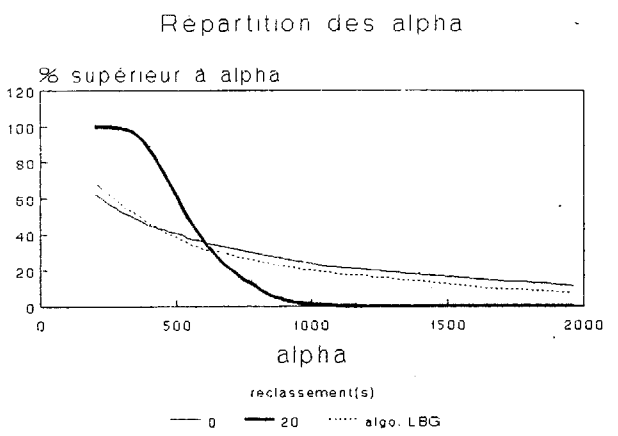

Figure 2 Répartition des distorsions pondérées par leur fréquence d'apparition

codage des paramètres LSP :
- schéma 20 bits : Codage sans dégradation audible nécessitant un débit binaire de transmission de 660 bits/s (tranche de 30 ms codée sur 20 bits).
- schéma 10 bits : Codage avec une légère dégradation mais n'utilisant que 330 bits/s (tranche de 30 ms codée sur 10 bits) pour la transmission de ces mêmes paramètres.
La base de données utilisée pour la création des dictionnaires est composée de phrases phonétiquement équilibrées d'environ 60 secondes chacune, prononcée par 13 locuteurs masculins et 7 locuteurs féminins. De cette base de données, environ 60000 vecteurs spectraux caractéristiques (pas de silence) sont retirés.
Dans les deux cas, l'algorithme K-Means modifié et la distance euclidienne pondérée spectralement ont servis à l'exploitation du dictionnaire. Dans le schéma 20 bits, une structure de type code produit est utilisée. Afin de favoriser une meilleure quantification des basses fréquence plus importante acoustiquement, le premier dictionnaire code les 4 premiers LSP et le deuxième les 6 suivants. La dimension des 2 dictionnaires est de 1024 centroïdes (2 X 10 bits). Dans le schéma 10 bits, un dictionnaire optimal à 1024 centroïdes (10 bits) de vecteurs LSP d'ordre 10 a été exploité. La figure 3 résume les résultats obtenus. Nous y

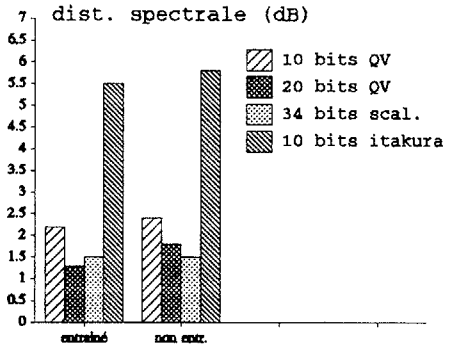

Figure 3 : Distorsion spectrale moyenne des différents schémas de quantification des paramètres LPC

avons introduit à titre de comparaison les résultats d'un codage scalaire sur 34 bits normalisé par le DOD dans le codeur 4800 bits/s CELP [11], ainsi que pour le dictionnaire à 10 bits, les résultats obtenus pour un dictionnaire à 1024 centroïdes utilisant la distance d'Itakura. Pour le schéma 20 bits, on constate que la quantification est meilleure pour les locuteurs entrainés que pour le schéma scalaire 34 bits. Il est cependant très légèrement moins bon pour les locuteurs non entrainés. Pour quantifier auditivement les erreurs de quantification des paramètres LSP, une écoute avec une excitation du filtre autorégressif "idéale" a été réalisée ; idéale voulant dire résultant du filtrage inverse avec les paramètres du filtre non quantifiés. Les différences entre le schéma scalaire et le schéma 20 bits sont inaudibles, que le locuteur soit entrainé ou non. Pour le schéma 10 bits, la dégradation devient légèrement audible. Elle reste cependant très nettement inférieure à celle du schéma utilisant un dictionnaire optimal et une distance d'Itakura (solution classique).

6 CONCLUSIONS

Nous avons montré qu'il est possible de quantifier les paramètres du filtre autorégressif caractérisant le conduit vocal avec un débit binaire de 660 bits/s, sans dégradation audible, et même sur 330 bits/s avec une dégradation limitée par l'utilisation de la quantification vectorielle des paramètres LSP. Ceci représente un gain de 50%, voire de 75% par rapport à un codage scalaire sur le débit nécessaire à la transmission des paramètres LSP.

BIBLIOGRAPHIE

[1] R. BOITE et M. KUNT
Traitement de la parole
Presse Polytechnique Romane, Edition, Lausanne 87

[2] F. ITAKURA
Line Spectrum Representation of Linear Predictive Coefficients of Speech Signals.
Journal of American Acoustic Society, 57, 535(a),1975

[3] F.K. SOONG, B.-W. JUANG
Line Spectrum Pair (LSP) and Speech data compression
IEEE ICASSP 84, CH 1945-5, pp 1.10.1-1.10.4

[4] CALLIOPE (Collectif d'auteurs)
La Parole et son Traitement Automatique
Chap. 2.2 : Psychoacoustique
Collection CNET-ENST, Masson, Paris 89

[5] H.J. COETZEE, T.P. BARNWELL III
A LSP based Speech quality measure
IEEE ICASSP 89, CH 2673-2, pp 596-599

[6] N. SUGAMURA, F. ITAKURA
Speech Analysis and synthesis methods developped at ECL in NTT - from LPC to LSP -
Speech Communication, num 5, 1986, pp 199-215

[7] R.M. GRAY
Vector Quantization
IEEE ASSP Magazine, vol 1, num 2, april 1984, pp 4-29

[8] J. MAKHOUL, S. ROUCOUS, H. GISH
Vector Quantization in Speech Coding
Proceed. IEEE, november 1985, pp 1551-1588

[9] A. GERSHO
Asymptotically Optimal Block Quantization
IEEE Trans. on Information Theory, Vol IT-25, num 4, July 1979

[10] Y. LINDE, A. BUZO and R.M. GRAY
An Algorithm for Vector Quantizer Design
I.E.E.E. Trans. on Communications, Vol COM-28,num 1, Jan 80

[11] J.P. CAMPBELL, T.E. TREMAIN, V.C. WELCH
The DOD 4,8 Kps standard (proposed federal standard 1016)
U.S. Department of Defense, Fort Meade, 1990