Quantification de séquences spectrales de longueurs variables pour le codage de la parole à très bas débit

Geneviève Baudoin [(1)], Jan Cernocký [(1,2)], Gérard Chollet [(3)]

[(1)] ESIEE, Dpt Signaux-télécommunications, BP 99, Noisy Le Grand, 93162 CEDEX, France, baudoing@esiee.fr
[(2)] FEI VUT Brno, Purkynova 118, 61200 Brno, Czech Republic, cernocky@urel.fee.vutbr.cz
[(3)] ENST, Dpt Signal, 46 rue Barrault, 75013 Paris, France, chollet@sig.enst.fr

RÉSUMÉ

Ce papier traite du codage des paramètres spectraux pour le codage de parole à très bas débit. Nous présentons une nouvelle interprétation de recherches précédemment publiées par Chou-Lockabaugh et Cernocky-Baudoin-Chollet sur la quantification de séquences spectrales de longueurs variables, sous les noms respectifs de « Variable to Variable length Vector Quantization » (VVVQ) et de quantification par multigrammes (MGQ). Nous avons, d'autre part étudié l'influence de la limitation du retard introduit par la méthode et proposé une technique pour optimiser les performances en présence d'un retard maximum imposé. Nous avons ainsi trouvé qu'un retard de 400 ms est généralement suffisant. Enfin, nous proposons l'introduction de longues séquences dans le dictionnaire par interpolation linéaire des séquences courtes.

ABSTRACT

This paper deals with the coding of spectral envelope parameters for very low bit rate speech coding. We propose a new interpretation of already published research from Chou-Lockabaugh and Cernocky-Baudoin-Chollet on the quantization of variable length sequences of spectral vectors, named respectively Variable to Variable length Vector Quantization (VVVQ) and Multigrams Quantization (MGQ). We have also studied the influence of the limitation of the delay introduced by the method and proposed a technique for optimizing the performances when a maximal delay is imposed. It was found that, with this technique, a delay of 400 ms is generally sufficient. Finally, we propose the introduction of long sequences in the segmental codebook by linear interpolation of shorter ones.

1 Introduction

Pour coder la parole à des débits inférieurs à 500 bps, il est nécessaire de prendre en compte les dépendances intertrames en utilisant des techniques de quantification segmentale pour le codage des paramètres spectraux.

Chou and Loockabaugh [2] ont proposé une méthode de quantification de séquences spectrales de longueurs variables sous le nom de VVVQ (Variable to variable Vector Quantization). Elle donne des résultats satisfaisants (compréhensibilité) en monolocuteur pour des débits spectraux aussi bas que 50 bps. Mais sa complexité et le retard introduit sont importants.

Une méthode de quantification de même type a été proposée indépendamment par Cernocký, Baudoin et Chollet [4,6] sous le nom de quantification de séquences spectrales par des multigrammes (on la notera MGQ).

Ce papier traite les thèmes suivants : nouvelle interprétation et comparaison des 2 approches, étude du retard nécessaire et proposition d'une technique pour optimiser les performances en présence d'un retard maximum imposé, introduction dans le dictionnaire de séquences spectrales longues par interpolation linéaire de séquences plus courtes.

2. Description et comparaison des méthodes VVVQ et des multigrammes

2.1 La méthode VVVQ

Cette méthode segmente et quantifie une suite temporelle de vecteurs spectraux à l'aide d'une quantification vectorielle à dimension variable en utilisant un dictionnaire de séquences spectrales (suite de vecteurs spectraux) de longueurs variables de 1 à n vecteurs. Les séquences du dictionnaires sont codées par un code entropique et sont donc représentées par un nombre variable de bits dépendant de leur probabilité. Ainsi, à la fois la longueur des séquences du dictionnaire et le nombre de bits pour les coder sont variables, d'où le nom « Variable to variable Vector Quantization ». Le dictionnaire est obtenu, sur une base de données d'apprentissage, en minimisant la distorsion spectrale moyenne pour un débit spectral moyen limité. Une technique de multiplicateur de Lagrange est appliquée et le critère à optimiser s'écrit :

$$\min_{S_i \in S} \left(d_{S_i} + \lambda \, r_{S_i} \right) \qquad (1)$$

Où S est l'ensemble de toute les segmentations possibles de la base en segments de longueur inférieure à n, S_i est l'une d'elles, d_{Si} est la distorsion correspondante, r_{Si} le débit associé et λ le multiplicateur de Lagrange. Plus précisément :

$$d_{S_i} + \lambda \, r_{S_i} = \sum_{U_j \in S_i} \left(d_j + \lambda \, n_j \right) \qquad (2)$$

Où U_j est le $j^{ème}$ segment de S_i, n_j le nombre de bits de codage de U_j et d_j la distorsion sur ce segment (somme des distorsions sur tous les vecteurs du segment). Dans tout le papier, on suppose que : $n_j = -\log_2(proba(M_j))$ (3)

Où M_j est la séquence du dictionnaire associée à U_j. Le dictionnaire est initialisé avec Z séquences de vecteurs et leurs probabilités. Puis, on utilise un algorithme EM (Expectation Maximization) [1] itératif pour calculer le dictionnaire. A la $q^{ème}$ itération, le dictionnaire C_q contient Z séquences $M_{q,j}$ avec leurs probabilités $p(M_{q,j})$, le nouveau dictionnaire C_{q+1} est calculé en 2 étapes :

- étape 1 : Segmentation de la base de données en N segments en optimisant le critère (1) avec l'algorithme de

Viterbi. A chaque $M_{q,j}$ correspond une classe de $N_{q,j}$ séquences de la base de données codées par $M_{q,j}$.

- étape 2 : Mise à jour du dictionnaire. $M_{q+1,j}$ est le centroïde de la classe de $M_{q,j}$ et $p(M_{q+1,j})$ est estimé par $N_{q,j}$ / N.

2.2 La Méthode de quantification par multigrammes MGQ

Comme pour la VVVQ, le principe consiste à segmenter et quantifier les séquences de vecteurs spectraux à l'aide d'un dictionnaire de segments de longueurs variables, appelés multigrammes par analogie avec les modèles de langage.

Dans une première approche, les vecteurs spectraux étaient quantifiés vectoriellement et les multigrammes M_k étaient des séquences de 1 à n indices de quantifications. Le dictionnaire était obtenu en maximisant la vraisemblance conjointe de l'observation d'apprentissage (séquence d'indices de QV) et de la segmentation optimale S_{opt} [3]. Les segments étant supposés indépendants, le critère était de maximiser : $L(observation, S_{opt}) = \max_{S_i \in S} \prod_{M_k \in S_i} p(M_k)$ (4)

Le dictionnaire était initialisé avec les séquences présentes dans la base d'apprentissage et leurs nombres d'occurrences. Puis il était optimisé (probabilités seulement) à l'aide de l'algorithme EM, un codage entropique étant appliqué aux multigrammes. Les résultats furent insuffisants pour les tailles de QV supérieures à 128, à cause de la trop grande variabilité des séquences d'indices.

Aussi, une deuxième approche a-t-elle été développée. Les vecteurs spectraux ne sont plus transformés en symboles par QV. Un multigramme M_k est une séquence de 1 à n vecteurs spectraux et non plus d'indices. Une chaîne de vecteurs spectraux est segmentée en segments U_k qui sont quantifiés par les multigrammes M_k de façon à maximiser le critère L' :

$$L'(observation, S_{opt}) = \max_{S_i \in S} \prod_k p'(M_k) \quad (5)$$

Où $p'(M_k)$ est la probabilité pénalisée de M_k, définie comme le produit de la probabilité de M_k avec un facteur de pénalité Q dépendant de la distance d_k entre le segment observé U_k et le multigramme qui le code M_k.

$$p'(M_k) = p(M_k)Q(d_k) \qquad d_k = d(U_k, M_k) \quad (6)$$

$$Q(d) = \begin{cases} 1 - \dfrac{d}{d_{\max}} & \text{pour} \quad d \le d_{\max} \\ 0 & \text{pour} \quad d > d_{\max} \end{cases} \quad (7)$$

Où d_{max} est une constante arbitraire. Le nombre de multigrammes de chaque longueur dans le dictionnaire initial est limité à priori. Le dictionnaire segmental est initialisé, puis il est calculé itérativement avec l'algorithme EM en optimisant le critère (2.5). A chaque itération, les 2 étapes de l'algo-rithme EM s'effectuent comme pour la méthode VVVQ.

2.3 Nouvelle interprétation et comparaison des 2 méthodes

Bien que développées indépendamment, les 2 techniques sont très ressemblantes. La VVVQ est mieux formulée mathématiquement et est localement optimale en distorsion pour un débit et une structure de dictionnaire donnés.

L'approche MGQ apporte un éclairage différent. Elle va d'abord être reformulée puis dans le cadre de cette nouvelle interprétation les 2 approches vont être comparées.

Pour reformuler la méthode MGQ, on considère qu'une séquence de vecteurs spectraux est générée par une source qui émet des multigrammes (MG) de longueur variable indépendants entre eux. On considère de plus que les vecteurs spectraux (de dimension p) de ces MG ont une densité de probabilité gaussienne de matrice de covariance $\sigma^2 I$, I étant la matrice identité de dimension pxp. Les paramètres θ (multigrammes et probabilités) de la source sont identifiés en maximisant la vraisemblance conjointe de la segmentation optimale S_{opt} et de l'observation:

$$\max_\theta L(obs, S_{opt}) \Leftrightarrow \max_\theta L(S_{opt}) L(obs / S_{opt}) \quad (8)$$

$$L(S) = \prod_k p(M_k) \quad \text{et} \quad L(obs/S) = \prod_k p(U_k / M_k) \quad (9)$$

Où U_k est un segment de longueur l_k de la base d'apprentissage et M_k le multigramme par lequel U_k est quantifié dans la segmentation S. Selon le modèle gaussien proposé et en appliquant un logarithme, le critère est équivalent à :

$$\max \sum_k \left(\log(p(M_k)) - \sum_{j=1}^{l_k} \sum_{m=1}^{p} \frac{(c_{k,j,m} - m_{k,j,m})^2}{2\sigma^2} \right) \quad (10)$$

$$\Leftrightarrow \min \sum_k \left(\left(\sum_{j=1}^{l_k} d(\mathbf{c}_{k,j}, \mathbf{m}_{k,j}) \right) - 2\sigma^2 \log(p(M_k)) \right) \quad (11)$$

$c_{k,j,m}$ et $m_{k,j,m}$ sont les $m^{èmes}$ coefficients du $j^{ème}$ vecteur du segment U_k et du multigramme M_k. $d(\mathbf{c}_{k,j}, \mathbf{m}_{k,j})$ est la distance quadratique entre les $j^{ème}$ vecteurs de U_k et M_k.

On reconnaît dans l'équation (2.11) le critère de la VVVQ avec $\lambda = 2\log(2)\sigma^2$ et une distance quadratique sur les vecteurs spectraux.

D'autre part , il est possible d'interpréter le critère arbitraire de la méthode MGQ en remarquant que pour $d << d_{max}$:

$$\log(p) + \log(1 - \frac{d}{d_{\max}}) \approx \log(p) - \frac{d}{d_{\max}} \quad (12)$$

Avec $d_{max} = 2\log(2)\sigma^2$. Nous avons utilisé ici une densité de probabilité triangulaire, qui est proche d'une gaussienne quand d est petit devant d_{max}.

Une autre interprétation peut être obtenue en considérant que la source émet des multigrammes constants et indépendants auxquels s'ajoute un bruit blanc gaussien centré de σ^2.

Une différence supplémentaire entre les approches VVVQ et MGQ réside dans la mesure de distorsion spectrale utilisée. Chou & al ont travaillé avec une distance d'Itakura modifiée alors que nous avons utilisé une distance quadratique sur les coefficients cepstraux. Avec la distance d'Itakura modifiée, les interprétations précédantes doivent être appliquées au résiduel de prédiction linéaire supposé blanc et gaussien.

3 Limitation du retard

En théorie, la méthode introduit un retard égal à la durée totale du signal. Quand on limite ce retard à une valeur de

k_{max} trames, les performances se dégradent. La technique classique pour limiter le retard, consiste à utiliser un buffer de k_{max} trames, à imposer des points de segmentation aux extrémités du buffer et à vider le buffer tous les k_{max} trames .

Nous avons développé un nouvel algorithme. A chaque entrée d'un vecteur spectral dans le buffer, nous examinons si les n meilleures segmentations possibles du buffer depuis son origine jusqu'au dernier vecteur reçu coïncident jusqu'à une certaine position. Si un tel point existe, le buffer est vidé jusqu'à ce point. Tant que le buffer ne sature pas, la limitation du retard à k_{max} trames ne dégradent pas les performances.

Nous avons étudié les caractéristiques statistiques du remplissage du buffer pour différentes valeurs de n et de λ. Les résultats sont donnés dans la section 6.

4 Construction de longs multigrammes par interpolation

Allonger la longueur maximale possible des multigrammes entraîne l'augmentation rapide du nombre de vecteurs spectraux à estimer. Ainsi, pour 64 multigrammes par longueur, y a-t-il 8704 vecteurs à estimer pour n=16 et 35088 pour n=32. Aussi, dans le but d'augmenter la longueur maximale n des segments, sans avoir à accroître la taille de la base de données d'apprentissage, avons-nous construit un dictionnaire contenant des multigrammes de longueur 1 à n à partir d'un dictionnaire de longueur maximale n/2, en étirant par interpolation linéaire les multigrammes de longueur n/2 pour obtenir les longs multigrammes de longueur n/2+1 à n, prenant ainsi en compte le fait que les mêmes séquences acoustiques peuvent être prononcées à différentes vitesses.

Lors de l'apprentissage, les multigrammes de longueur n/2 sont actualisés à partir des segments de longueur n/2 qui leur ont été associés et à partir des segments de longueurs n/2+1 à n associés à ces multigrammes étirés. Dans ce dernier cas, la mise à jour se fait par contraction linéaire des longs segments. Les probabilités sont actualisées normalement pour toutes les longueurs. On sauve à chaque itération les multigrammes de taille 1 à n/2 et toutes les probabilités.

Les résultats obtenus (courbes distorsion-débit), pour n=16 avec étirement des multigrammes de longueur 8, sont supérieurs à ceux obtenus avec un dictionnaire non étiré de longueur maximale 12 et contenant le même nombre de vecteurs spectraux à estimer (fig 2). Mais la complexité est augmentée.

5 Résultats expérimentaux

définitions de la distorsion et du débit:

La distorsion spectrale est calculée en dB à partir d'une distance euclidienne entre les coefficients cepstraux originaux et quantifiés.

Le débit binaire est défini comme le nombre moyen de bits pour le codage d'un vecteur spectral. C'est un nombre moyen de bits par trame. Le débit binaire moyen par trame R, correspondant à dictionnaire de multigrammes avec un codage entropique, est le rapport de l'entropie H du dictionnaire à la longueur moyenne $\bar{l}$ des multigrammes :

$$R = \frac{H}{\bar{l}} = -\frac{\sum_{i=1}^{Z} p(M_i)\log_2 p(M_i)}{\sum_{i=1}^{Z} p(M_i) l(M_i)} \qquad (13)$$

Où $l(M_i)$ et $p(M_i)$ sont la longueur et la probabilité de M_i, et Z est le nombre de multigrammes dans le dictionnaire.

Base de données :

Nous avons utilisé un seul locuteur de la base de données PolyVar en français suisse. Elle contient des appels téléphoniques enregistrés sur une période de 6 mois, constitués de phrases lues, de mots épelés, de nombres, de quelques mots de contrôle et de parole spontanée. Le signal est numérisé à 8 kHz selon une loi A 8 bits. Les vecteurs spectraux sont formés de 10 coefficients LPCC calculés avec préemphase sur des fenêtres de Hamming de 20ms avec un recouvrement de 10ms. Le premier coefficient cepstral (lié à l'énergie) n'est pas utilisé. Le corpus a été divisé en 213270 vecteurs d'apprentissage et 122903 vecteurs de test.

Initialisation du dictionnaire:

Différentes initialisations du dictionnaire ont été comparées :

- Initialisation avec les multigrammes quantifiés les plus fréquents : après quantification vectorielle de la base d'apprentissage, on utilise pour chaque longueur l de multigramme les séquences quantifiées de longueur l les plus fréquentes.
- Initialisation par quantification matricielle (toutes les séquences du dictionnaire ont la même longueur) : pour chaque longueur l de multigramme, le dictionnaire de multigrammes est initialisé avec un dictionnaire de quantification matricielle à séquences de longueur l [2].
- Initialisation aléatoire naturelle : le dictionnaire de multigrammes est initialisé avec des séquences naturelles de vecteurs spectraux choisies au hasard [5].

Après quelques itérations de l'algorithme EM, les 3 initialisations ont donné des résultats similaires, nous avons donc utilisé la troisième dans les expériences.

Configurations de test :

Les résultats suivants ont été obtenus avec différentes topologies pour les dictionnaires de Multigrammes ou de quantification matricielle :

- MG16 Quantification par multigrammes avec n=16 et 64 multigrammes par longueur, 0≤λ≤1. Il y a 8704 vecteurs cepstraux dans le dictionnaire.
- MQ8704 Quantification Matricielle (λ=0) avec différents dictionnaires chacun correspondant a une longueur unique de séquence l entre 2 et 20 vecteurs, tous les dictionnaires contenant 8704 vecteurs cepstraux. Ainsi, pour l=8, y a-t-il 1088 séquences de 8 vecteurs dans le dictionnaire et pour l=16 544 séquences de 16 vecteurs.
- MQ1, MQ2, MQ4, 3 Quantifications matricielles avec codage entropique des séquences et des dictionnaires de 8704 séquences de longueur respectives 1, 2, 4, contenant 8704, 17408 and 34816 vecteurs cepstraux, 0≤λ≤1.
- MG8, MG12, MG8_étiré, quantifications par multigrammes avec respectivement n = 8, 12, 16, et 113, 64, 113 multigrammes par longueur, 0≤λ≤1. Les longs multigrammes (l = 9 à 16) de MG8_étiré sont obtenus en étirant les séquences de longueur 8. Il y a 4992 vecteurs à estimer pour MG12 et 4068 pour MG8 et MG8_étiré.

Limitation du retard, remplissage du buffer

La figure 1 représente, pour $0 \leq \lambda \leq 0.5$, les fonctions de répartition du remplissage du buffer (nombre de trames dans le buffer), pour MG16, lorsque l'on utilise le nouvel algorithme pour la limitation du retard.

En cas de limitation du retard à k_{max} trames, les performances sont nettement améliorées avec le nouvel algorithme par rapport à la vidange du buffer toutes les k_{max} trames. Par exemple, pour MG16, quand $0<\lambda<0.1$, un buffer de $k_{max}=40$ trames (retard maximal de 400 ms) donne des résultats équivalents à ceux obtenus avec un retard illimité. Pour $0.1<\lambda<0.2$, un buffer de 70 trames est suffisant.

Comparaison de la Quantification par multigrammes (VVVQ, MGQ) et de la Quantification Matricielle (MQ)

Chou & al ont comparé la VVVQ et l'ECMQ (Entropy Constrained MQ) pour une même complexité. Mais ils ne purent pas estimer les grands dictionnaires de MQ pour des longueurs de séquences supérieures à 4 à cause de la taille limitée de la base de données utilisée. Aussi avons nous de plus comparé les 2 approches pour un même nombre de vecteurs spectraux dans les dictionnaires (configuration MG16 vs MQ8704). La figure 3 donne les courbes distorsion-débit obtenues sur la base de test avec les configurations MG16, MQ8704, MQ1, MQ2, MQ4.

Pour les petits débits spectraux (moins de 2 bits/trame, 200 bits/s), la quantification par multigrammes est supérieure à la matricielle. Mais, quand la comparaison est faite pour un même nombre de vecteurs cepstraux dans les dictionnaires de MGQ ou de MQ, le gain en performance est assez faible pour une augmentation significative de la complexité.

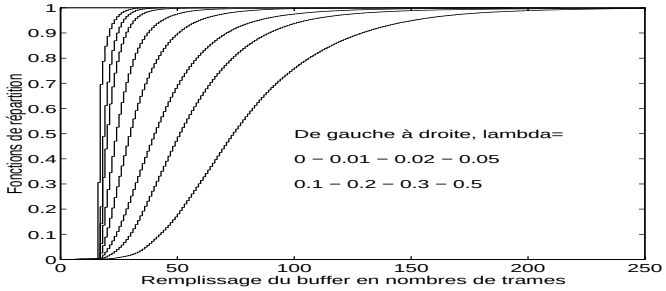

fig 1 : Fonctions de répartition du remplissage du buffer pour $0 \leq \lambda \leq 0.5$, n=16, MG16

6 Conclusion

Dans ce papier nous avons apporté une nouvelle interprétation de la VVVQ (Variable to Variable length Quantization) et une présentation unifiée de la VVVQ et de la quantification par multigrammes. La quantification vectorielle à longueur variable ou quantification par multigrammes n'a été testée que sur un seul locuteur, mais la nouvelle interprétation devrait permettre d'utiliser les techniques d'adaptation de la reconnaissance vocale.

La quantification par multigramme donne de meilleures performances en terme de distorsion-débit, que la quantification matricielle mais au prix d'un accroissement de complexité et de retard.

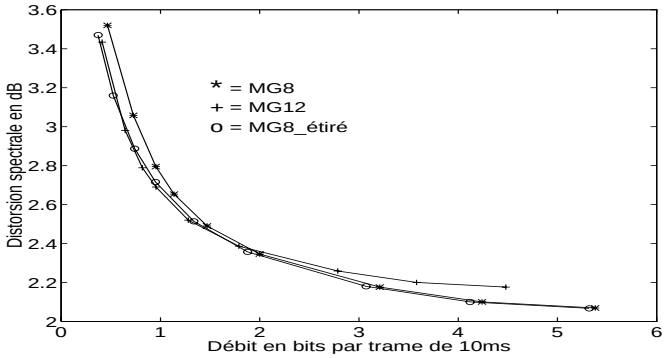

fig 2: Courbes distorsion-débit avec et sans étirement

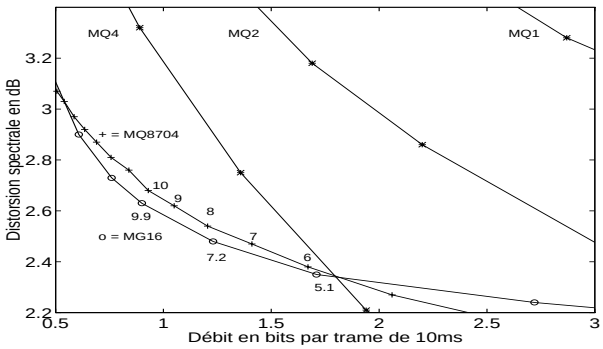

fig 3: Courbes distorsion-débit, les nombres le long des courbes MQ8704 et MG16 représentent respectivement les longueurs des séquences ou les longueurs moyennes.

Un défaut de la méthode est de ne pas explicitement prendre en compte la variabilité du rythme d'élocution. Un essai d'étirement linéaire des multigrammes a permis d'améliorer légèrement les performances mais insuffisamment. Aussi allons nous appliquer la quantification par multigramme aux cibles spectrales d'une décomposition temporelle.

7 Références

[1] Dempster P., Laird N. M. and Rubin D. B. *"Maximum Likelihood from Incomplete Data with the EM Algorithm"*, J. Roy. Stat., 39(1), pp.1-38, 1977.

[2] Chou A. and Lookabaugh T., *"Variable dimension vector quantization of linear predictive coefficients of speech"*, Proc. IEEE ICASSP 94, pp. I.505-508, Adelaide Australia, June 1994.

[3] Deligne S. and Bimbot F., *"Language Modelling by Variable Length Sequences: Theoretical Formulation and Evaluation of Multigrams"*, Proc. IEEE ICASSP 95, pp. 169-172, Detroit USA, 1995.

[4] Cernocký J., Baudoin G. and Chollet G., *"Speech Spectrum representation and coding using multigrams with distance"*, Proc. IEEE ICASSP 97, pp. II.1343-1346, Munchen Germany, April 1997.

[5] Roucos S., Schwartz R. and Makhoul J., *"Segment quantization for very low bit rate speech coding*", Proc. IEEE ICASSP 82, pp. 1565-1568, Paris, Apr 1982.

[6] Cernocký J. and Baudoin G., *"Représentation du spectre de parole par les multigrammes"*, Proc. XXI-es Journées d'Etude sur la Parole, pp.239-242, Avignon, Jun 1996.