QUANTIFICATION DE SÉQUENCES SPECTRALES DE LONGUEURS VARIABLES POUR LE CODAGE DE PAROLE À TRÈS BAS DÉBIT

Geneviève Baudoin [1], *Jan Černocký* [1,2]

[1]ESIEE, Département Signal, BP 99, Noisy Le Grand, 93162 CEDEX, baudoing@esiee.fr
[2]VUT Brno, Brno, République tchèque, cernocky@urel.fee.vutbr.cz

Rubrique: 2.3
Problème posé Le problème posé est celui de la quantification de l'enveloppe spectrale du signal pour le codage de parole à très bas débit (< 500 bits par seconde) avec conservation de la compréhensibilité.
Originalité du travail Le travail apporte une nouvelle interprétation de recherches précédemment publiées . On a de plus étudié l'influence de la limitation du retard introduit par la méthode et l'intérêt d'un nouveau type de dictionnaire.
Les résultats nouveaux La nouvelle interprétation apporte un sens au multiplicateur de Lagrange du critère d'optimisation et devrait permettre de nouveaux développemnts. Il a été trouvé qu'un retard de 200 ms est suffisant pour les débits concernés.

1. INTRODUCTION

Pour des applications de codage de parole à des débits inférieurs à 500 bps, Chou et Loockabaugh 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 100 bps. Elle comporte toutefois certaines limitations: retard théorique égal à la durée du signal, complexité, grande bases d'apprentissage. Et plusieurs questions restent ouvertes, comme l'adaptation au locuteur.

Une méthode de quantification de même type a été proposée indépendamment par Cernocky et Baudoin sous le nom de représentation de séquences spectrales par les multigrammes (on la notera MG). Ce papier traite les thèmes suivants : 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 de grandes longueurs par interpolation linéaire de séquences plus courtes.

2. DESCRIPTION ET COMPARAISON DES MÉTHODES VVVQ ET MG

2.1. Méthode VVVQ

Cette méthode quantifie l'enveloppe spectrale à l'aide d'un dictionnaire contenant des séquences spectrales de longueurs variables de 1 à n trames.

Par ailleurs, les séquences du dictionnaires sont codées à l'aide d'un code entropique et sont donc représentées par un nombre variable de bits lié à la probabilité des séquences.

Le dictionnaire est obtenu sur une base de données d'apprentissage en minimisant la distortion spectrale moyenne pour un débit limité. Une technique de multiplicateur de Lagrange est appliquée et le critère a optimiser s'écrit: $\min_{S_i \in S} d_{S_i} + \lambda r_{S_i}$, où S est l'ensemble de toutes les segmentations possibles de la base, S_i est l'une de ces segmentations, d_i la distortion correspondante, r_i le débit associé, et λ le multiplicateur de Lagrange. Plus précisément : $d_{S_i} + \lambda r_{S_i} = \sum_{s_j \in S_j} d_{i,j} + \lambda n_{i,j}$, où s_j est le $j^{j ème}$ segment de la segmentation S_i, $n_{i,j}$ le nombre de bits pour coder ce segment et $d_{i,j}$ la distortion sur ce segment(somme des distortions sur tous les vecteurs du segment).
$n_{i,j} = -\log(proba(M_{i,j})$, où $M_{i,j}$ est la séquence du dictionnaire associée à $s_{i,j}$.

Un algorithme itératif de type EM est utilisé pour calculer le dictionnaire après son initialisation. chaque itération comporte 2 étapes:
1. On segmente et on quantifie la base avec le dictionnaire correspondant à l'itération, de façon à optimiser le critère. On utilise un algorithme de Viterbi.
2. On actualise le dictionnaire en mettant à jour les probabilités et les séquences du dictionnaire en fonction des segments de la base qui leur ont été associés.

2.2. Méthode de représentation du spectre par des multigrammes MG

Comme pour la VVVQ, l'idée de base est de segmenter et quantifier les séquences de vecteurs spectraux à l'aide d'un dictionnaire de séquences caractéristiques de longueurs variables. Les séquences du dictionnaire sont appelées multigrammes.
Une première approche a consisté à quantifier vectoriel-

lement les vecteurs spectraux sur q valeurs puis à segmenter la chaîne des indices de QV . Les multigrammes sont ici des séquences d'indices. Le dictionnaire est calculé en optimisant sur une base d'apprentissage la vraissemblance conjointe L de la segmentation optimale et de l'observation, c'est à dire de la chaîne U des indices de la base : $\max_{S_i} L(U, S_i)$. Où $L(U, S_i)$ est la vraissemblance conjointe de la chaîne U et d'une segmentation S_i. Un algorithme EM a été utilisé pour apprendre le dictionnaire. Puis les multigrammes ont été codés par un codage entropique. La distortion introduite par la méthode est celle de la QV. On a comparé le débit obtenu avec celui de la QV avec codage entropique. Pour des tailles de QV petites cette approche permet de gagner 1 à 2 bits par trame. Mais pour des tailles de QV supérieures à 512 les résultats ne sont pas bons, la variabilité des séquences d'indices étant trop grande pour des séquences acoustiquement proches.

Aussi une $2^{ème}$ approche a-t-telle été développée. Les vecteurs spectraux ne sont plus transformés en symboles par QV. Un multigramme M_k est une suite de vecteurs spectraux, et non plus d'indices. Une chaîne de vecteurs spectraux est segmentée et quantifiée de manière à optimiser le critère $\max L(S) = \max_{S_i \in S} \prod_k p'(M_k)$ Où $p'(M_k)$ est la probabilité pénalisée du multigramme M_k définie comme le produit de la probabilité de M_k par un terme de pénalisation Q lié à la distance entre M_k et le segment U_k qu'il code.

$$Q[d] = \begin{cases} 1 - \frac{d}{d_{max}} & \text{pour} \quad d \leq d_{max} \\ 0 & \text{pour} \quad d > d_{max} \end{cases} \qquad (1)$$

Où d_{max} est une constante correspondant à la distance maximale pour laquelle p' peut être nulle.

2.3. Nouvelle interprétation et comparaison des 2 méthodes

Bien que développées indépendamment, les 2 techniques sont très ressemblantes. La méthode VVVQ a l'avantage d'être mieux formulée mathématiquement et d'optimiser la distortion pour une débit donné et une structure de dictionnaire donnée. L'approche MG apporte un éclairage différent. Elle va être reformulée plus proprement ici. Et dans le cadre de cette nouvelle interprétation, les 2 approches vont être comparées.

Pour reformuler la méthode MG, on considère qu'une chaîne de vecteurs spectraux est générée par une source qui émet des MG de longueur variable indépendants entre eux. On considère de plus que les vecteurs spectraux de ces MG ont une densité de probabilité gaussienne de matrice de covariance $\sigma^2 I$, où I est la matrice identité de dimension p (la taille des vecteurs). Les paramètres θ (probabilités et séquences) de la source sont déterminés de facon à maximiser la vraissembalance L de la segmentation optimale S_{opt} pour l'observation:

$$\max_{\theta} L(S_{opt}/obs) \quad \Leftrightarrow \quad \max L(S_{opt}) L(obs/S_{opt}) \qquad (2)$$

$$L(S) = \prod_k p(M_k) \qquad L(obs/S) = \prod_k p(U_k/M_k) \qquad (3)$$

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 de départ on a:

$$p(U_k/M_k) = \prod_{j=1}^{l_k} K \exp\left(\frac{-1}{2\sigma^2} \sum_1^p (c_{k,j,m} - m_{k,j,m})^2\right) \qquad (4)$$

avec K constante. Soit en passant en log:

$$\max \sum_k \sum_{j=1}^{l_k} \sum_{m=1}^{p} \frac{-1}{2\sigma^2} (c_{k,j,m} - m_{k,j,m})^2 + log(p(M_k)) \Leftrightarrow \qquad (5)$$

$$\min \sum_k \sum_{j=1}^{l_k} d(c_{k,j}, m_{k,j}) - 2\sigma^2 log(p(M_k)) \qquad (6)$$

Où $c_{k,j,m}$ et $m_{k,j,m}$ sont respectivement les coefficients m du vecteur j du segment U_k et du multigrame M_k, et où $d(c_{k,j}, m_{k,j})$ est une distance quadratique entre le vecteur j du segment et du multigramme associé.

On reconnait dans la dernière formule le critère de la VVVQ avec $\lambda = 2\sigma^2$ et une distance quadratique sur les vecteurs spectraux.

D'autre part, on peut interpréter le critère arbitraire de la méthode MG en considérant que, pour $d << dmax$:

$$log(p) + log(1 - \frac{d}{dmax}) \simeq log(p) - \frac{d}{dmax} \qquad (7)$$

3. LIMITATION DU RETARD

Le retard théorique introduit par la méthode est égal à la durée totale du signal.

quand on limite le retard à une valeur kmax trop petite, les performances se dégradent. La technique classique, pour limiter le retard consiste à utiliser un buffer de kmax trames,à imposer des points de segmentation à chaque extrémité et à vider le buffer tous les kmax trames.

Nous avons développé un nouvel algorithme pour limiter les dégradations de performance en présence d'un retard limité. A chaque nouveau vecteur spectral, on recherche la présence d'un point commun de segmentation dans le buffer. c'est à dire que l'on examine si les n meilleures segmentations possibles du buffer depuis son origine jusqu'au denier vecteur reçu ne sont pas formées des mêmes segments jusqu'à un point du buffer que l'on a appelé point commun de segmentation. 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 kmax ne dégrade pas les résultats.

Nous avons étudié les caractéristiques statistiques du remplissage du buffer, et ceci pour différentes valeurs de n et de λ. La figure 1 donne un exemple de fonctions de répartition du remplissage du buffer.

Pour kmax=20 (retard de 200 ms) les performances sont beaucoup améliorées avec cet algorithme en comparaison avec une vidange du buffer toutes les kmax trames.

en conclusion, un retard de 200 ms permet d'obtenir des résultats optimaux pour n=1 à n=16 et $0 < \lambda < 0.05$. Pour n=16 et $\lambda > 0.1$ un retard de 40 trames est nécessaire.

4. FORMATION DE MULTIGRAMMES LONGS PAR INTERPOLATION DE MULTIGRAMMES PLUS COURTS

Augmenter la longueur maximale possible des multigrammes du dictionnaaire, nécessite l'augmentation de la base d'apprentissage. Ainsi, pour n=16 et 64 multigrammes par longueur, y a-t-il 8704 vecteurs à entraîner. aussi dans le but d'augmenter la longueur maximale sans avoir à augmenter la base de données, avons nous constitué un dictionnaire avec des multigrammes de longueur 1 à n obtenu à partir d'un dictionnaire de longueur maximale n/2 dans lequel on étire par interpolation linéaire les multigrammes de taille n/2 pour obtenir ceux de taille n/2+1 à n. Lors de l'apprentissage, les MG de taille n/2 sont actualisés à partir des séquences de longueur n/2 qui leur ont été associées et des séquences de longueur n/2+1 à n associées à ces MG étirés. dans ce dernier cas la mise à jour se fait par compression linéaire des séquences longues. Les probabilités sont actualisées normalement pour toutes les longueurs. On sauve à chaque itération, les MG de taille 1 à n/2 et toutes les probabilités.

Les résultats obtenus (courbe distortion - débit) sont inférieurs à ceux obtenus avec un dictionnaire normal de longueur maximale n. ils sont très légérement supérieurs à ceux obtenus avec un dictionnaire de longueur maximale n/2, mais au prix d'une beaucoup plus grande complexité. Cette approche s'est donc révélée inefficace.

5. CONDITIONS ET RÉSULTATS EXPÉRIMENTAUX

les expériences ont été réalisées avec une base de données téléphonique monolocuteur en français.
Les vecteurs spectraux utilisés sont les LPCC (Linear Predictive Cepstral Coefficients) d'ordre p=10, calculés sur des trames de 20 ms toutes les 10ms. Le corpus d'apprentissage comprenait 213270 vecteurs, et celui de test 122903 vecteurs.

La figure 2 donne les courbes distortion-débit obtenues sur la base de test avec la méthode VVVQ pour les configurations:
MG16: n=16 et 64 MG par longueur
MG5: n=5 , 64 MG de longueur 1, et 128 MG de longueur 2 à 5
Des résultats de quantification matricielle MQ sont aussi donnés.

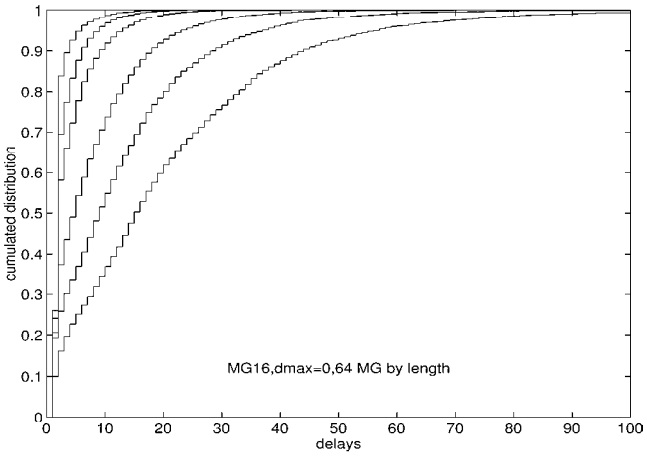

fig 1: fonctions de répartition du remplissage du buffer pour λ =0,0.01,0.02,0.05,0.1,0.2, kmax=512 n=16

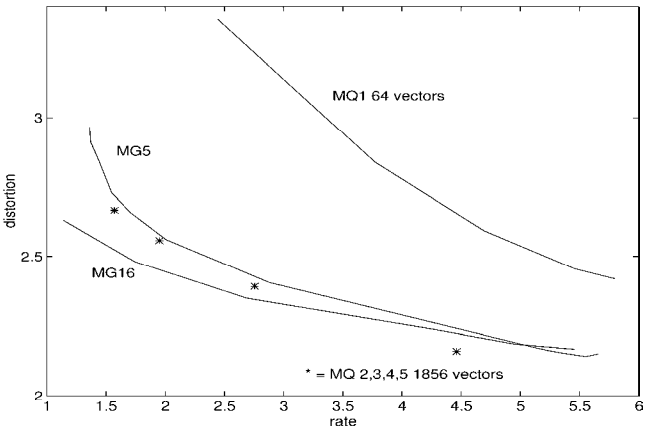

fig 2: Courbes distortion-débit pour les configs MG5, MG16 et quelques configs de MQ