DIXIEME COLLOQUE SUR LE TRAITEMENT DU
SIGNAL ET SES APPLICATIONS

NICE du 20 au 24 MAI 1985

APPLICATION DE LA QUANTIFICATION VECTORIELLE A LA RECONNAISSANCE DE LA PAROLE*

R. BOITE, H. LEICH, G. ZANELLATO

Service d'Electronique et Techniques Numériques, Faculté Polytechnique de Mons
31, boulevard Dolez, B-7000 MONS (Belgique)

RESUME

On sait que l'analyse de la parole est effectuée sur des tranches de l'ordre de 10 ms : les paramètres associés à une tranche constituent un vecteur spectral. L'expérience montre que l'ensemble des vecteurs spectraux peut être partitionné en un nombre fini de classes M : les vecteurs $X_j^{(i)}$ appartenant à une même classe i sont représentés par un vecteur "moyen" C_i appelé "centroïde". Les vecteurs sont classés de façon à minimiser la distorsion totale, c'est-à-dire la somme étendue à tous les vecteurs de la distance entre $X_j^{(i)}$ et C_i.

La distance entre deux vecteurs spectraux est la distance d'ITAKURA liée aux paramètres issus de la prédiction linéaire (LPC) ou encore la distance cepstrale.

Une classification dite par "éclatements binaires" a été choisie; elle permet une quantification très rapide (mais sous-optimale) des mots à reconnaître; cet algorithme est brièvement décrit et on précise la valeur obtenue pour la distorsion totale en fonction du nombre de classes.

Quant à la qualité effective de la classification, elle est déterminée par les résultats obtenus pour la reconnaissance de la parole. On donne les résultats obtenus pour deux méthodes de reconnaissance : la comparaison dynamique des vecteurs spectraux et la méthode basée sur une modélisation de chaque mot par un automate probabiliste.

SUMMARY

The analysis of speech signals is made on successive frames of about 10 to 20 ms; each frame is thus characterized by a set of parameters denoted as "spectral vector". The experience shows that the whole set of the spectral vectors can be partitionned into a finite set of M classes; each class is represented by a vector called "centroïd". Such a partition corresponds to a "vector quantization"; it is designed to minimize the total distorsion.

The ITAKURA distorsion measure is used as a distance between two spectral vectors; it is based on the LPC parameters; the cepstral distance has also been used.

The "binary splitting" procedure has been chosen for the partitionning; such a procedure is efficient for vector quantization in real time, as explained in the paper. That procedure is brievly described; the mean distorsion is given as a function of the number of classes.

The real effectiveness of the quantization is of course related to the results obtained during speech recognition. Such results are given for two recognition algorithms : the Dynamic Time Warping and another method which is based on statistical concepts.

* Ce travail a été partiellement subventionné par le Ministère Belge des Affaires Economiques sous la convention IRSIA-IWONL n° 4280.

APPLICATION DE LA QUANTIFICATION VECTORIELLE A LA RECONNAISSANCE DE LA PAROLE

1. INTRODUCTION

L'objectif actuel de notre groupe de recherches en Reconnaissance de la Parole est la comparaison de deux algorithmes pour la reconnaissance de mots isolés appartenant à un vocabulaire limité à N mots (20 mots constituant le vocabulaire de commande d'une calculatrice et les 26 lettres de l'alphabet). Les algorithmes étudiés sont :

(a) l'algorithme "Dynamic Time Warping" (D.T.W.), bien documenté dans la littérature [1,2];

(b) un algorithme basé sur une méthode statistique qui modélise la production de chaque mot du vocabulaire par un Automate Probabiliste (A.P.) [3].

On sait que l'algorithme D.T.W. calcule une distance entre le mot à reconnaître et chaque mot du vocabulaire de référence. Chaque mot est constitué par une suite de "tranches" dont la durée est de l'ordre de 10 ms; une analyse LPC effectuée sur chaque tranche permet de définir un "vecteur spectral" qui la caractérise.

La distance de ITAKURA [1] a été choisie pour caractériser la "distance locale", c'est-à-dire la distance entre une tranche du mot inconnu et une tranche d'un mot de référence. Si I désigne le nombre de tranches du mot à tester et J celui du mot de référence, l'algorithme D.T.W. détermine la distance entre ces deux mots en cherchant un chemin optimal dans une grille constituée de I x J points.

Ceci implique le calcul de $\gamma.I.J$ distances locales; le coefficient $\gamma < 1$ tient compte d'une limitation (heuristique) de la zone de recherche du chemin optimal; la reconnaissance d'un mot exige donc, en principe, $N \gamma I_m J_m$ calculs de distance (I_m et J_m désignent des valeurs moyennes), si chaque mot fournit une seule référence. On sait aussi que, pour un système indépendant du locuteur, chaque mot est souvent représenté par un nombre de références qui peut varier de 5 à 10.

Afin de limiter le temps de calcul nécessaire à la reconnaissance, il paraît intéressant de "quantifier" l'espace des vecteurs spectraux. Une partition de cet espace en M classes a donc été opérée; chaque classe est représentée par un "centroïde".

Les distances entre chaque paire de centroïdes sont calculées une fois pour toutes et mémorisées. L'algorithme de quantification doit être efficient pour permettre un calcul en temps réel du centroïde associé à chaque tranche du mot à reconnaître; les distances locales sont alors simplement lues dans un tableau.

Dans l'algorithme A.P., un automate probabiliste est associé à chaque mot du vocabulaire. Les transitions entre les états de l'automate sont caractérisées par des lois de probabilité; d'autre part, en chaque état, les "sons" sont émis selon une certaine loi de probabilité.

L'algorithme de VITERBI [3] permet de déterminer la probabilité avec laquelle un modèle donné a pu produire un mot donné; il doit donc être appliqué à tous les modèles de la référence pour identifier celui qui présente la plus grande probabilité d'avoir produit le mot à reconnaître.

Il est clair qu'une quantification vectorielle permet de réduire d'une façon importante le temps de calcul exigé par l'algorithme A.P., pour l'entraînement des modèles et pour la reconnaissance proprement dite.

La section 2 est consacrée à une description sommaire de la procédure adoptée pour la quantification vectorielle en vue de son application à la reconnaissance.

La réduction du temps de calcul, ainsi que l'accroissement éventuel du taux d'erreurs pour les algorithmes D.T.W. et A.P. sont étudiés respectivement dans les sections 3 et 4.

Enfin, une discussion sommaire sur le choix de la distance a été développée dans la section 5.

2. ORGANISATION GENERALE DE LA CLASSIFICATION

Par "quantification vectorielle", on entend une quantification globale des vecteurs spectraux, par opposition à une quantification indépendante de chaque paramètre. On définit donc une partition de l'espace des paramètres en M classes; les vecteurs appartenant à une même classe sont représentés par un "vecteur moyen" appelé centroïde.

Il est clair que les "erreurs de quantification" dépendent du nombre de classes M; pour la transmission de la parole, des résultats intéressants sont obtenus par une quantification séparée des sons voisés et des sons non voisés avec un nombre total de classes de l'ordre de 1024 ou 2048.

Pour la reconnaissance de la parole, on peut estimer qu'un nombre de classes plus restreint (M = 64 ou 128) doit suffire; il en résulte une réduction sensible du volume de calcul quel que soit l'algorithme choisi.

L'ensemble des cinq versions du vocabulaire (230

mots prononcés par le même locuteur) constitue environ 130 secondes de signal. L'analyse LPC (algorithme de LEROUX-GUEGUEN [4]) a été effectuée dans les conditions suivantes :

- fréquence d'échantillonnage 8680 Hz;
- préaccentuation et fenêtre de HAMMING;
- blocs de 264 échantillons avec recouvrements de 88 échantillons;
- ordre de la prédiction p = 8.

On dispose de 12.748 vecteurs spectraux que l'on désire grouper selon $M = 2^m$ classes distinctes. Les vecteurs $X_j^{(i)}$ appartenant à une classe i seront représentés par un vecteur moyen C_i appelé centroïde; il faut minimiser la "distorsion totale"

$$D_{TOT} = \sum_{i=1}^{M} \left[\sum_{j} d(X_j^{(i)}, C_i) \right]$$

où d(.,.) est la distance entre deux vecteurs.

Les premiers travaux ont été basés sur la distance de ITAKURA [1] qui s'exprime par :

$$d_{ITA}(X_1,X_2) = \left[r_a^{(1)}(o)\, r_x^{(2)}(o) + 2 \sum_{k=1}^{p} r_a^{(1)}(k)\, r_x^{(2)}(k) \right] / \sigma_2^{(2)} - 1$$

où $r_a(k)$: autocorrélation des coefficients de prédiction de $X^{(1)}$

$r_x(k)$: autocorrélation de la tranche (2) dont l'analyse a fourni $X^{(2)}$

$\sigma_x^{(2)}$: gain du modèle AR associé à cette même tranche

L'algorithme de quantification est basé sur la méthode itérative de LOYDT [5]; si le nombre de classes M est fixé à priori :

(a) pour un ensemble de centroïdes C_i donné, D_{TOT} est minimum lorsque chaque vecteur X_j est affecté à la classe i dont le centroïde C_i est le plus proche;

(b) une classification "provisoire" étant ainsi effectuée, la distorsion totale dans chaque classe i peut être réduite si l'on détermine une meilleure position du centroïde C_i.

Ceci conduit très clairement à un algorithme itératif initialisé par un choix arbitraire des C_i ("points germes").

Le calcul des centroïdes C_i est opéré une fois pour toutes; les mots de la référence sont analysés et leurs vecteurs spectraux sont aussi quantifiés une fois pour toutes et mémorisés. On calcule également la distance entre chaque paire de centroïdes; les résultats sont mémorisés dans un tableau.

Le point délicat qui subsiste est donc la quantification, à opérer en "temps réel", des vecteurs spectraux produits par l'analyse LPC d'un mot à reconnaître. Chaque vecteur doit, en principe, être comparé à $M = 2^m$ centroïdes pour être affecté à une classe.

Dans le but de réduire le nombre de comparaisons, la classification initiale a, en réalité, été effectuée selon le principe des "éclatements binaires" successifs depuis M = 2 jusqu'à $M = 2^m$. Soit J le nombre de classes d'une partition : on optimise successivement des partitions qui comportent $J = 2,4,8,\ldots,2^m=M$ classes.

La procédure est initialisée avec deux centroïdes définis par les vecteurs de coefficients de corrélation partielle (PARCOR) :

$$K_1 = [\ +1/2,0,\ldots,0\ ]$$

$$K_2 = [\ -1/2,0,\ldots,0\ ]$$

ce qui correspond approximativement à une partition en sons voisés et en sons non voisés.

Pour "éclater" une classe, il suffit de perturber légèrement les coefficients PARCOR associés à leur centroïde ($K_i \leftarrow 0{,}99\ K_i$ et $1{,}01\ K_i$).

On garde en mémoire le tableau des centroïdes pour chaque stade de la quantification : l'affectation des vecteurs du mot à reconnaître à une des 2^m classes s'effectue donc par une suite de m comparaisons successives comportant chacune un simple choix binaire.

La quantification effectuée par une telle méthode est clairement sous-optimale; les M centroïdes obtenus par éclatement binaire peuvent cependant être utilisés pour initialiser une procédure classique : on diminue ainsi la distorsion totale, mais on perd la possibilité d'une quantification "rapide" des vecteurs.

La partition par éclatements binaires des 12.748 vecteurs en 64 classes demande 86 minutes de CPU (VAX 780) et 115 minutes pour 128 classes. Une quantification globale à partir de ces 64 centroïdes utilisés pour l'initialisation exige 3 h 15 min de CPU (plus de 5 h pour M = 128).

Un certain nombre de critères objectifs peuvent être retenus pour caractériser la qualité d'une quantification vectorielle. On ne donnera ici que l'évolution de la distorsion moyenne par vecteur D_{TOT}/N en fonction du nombre de classes M (figure 1).

APPLICATION DE LA QUANTIFICATION VECTORIELLE
A LA RECONNAISSANCE DE LA PAROLE

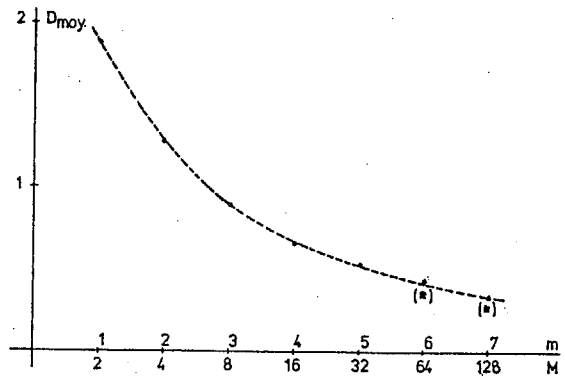

Figure 1 : Distorsion moyenne en fonction du nombre de classes.

3. APPLICATION A L'ALGORITHME D.T.W.

L'intérêt essentiel de la quantification vectorielle selon la procédure qui vient d'être décrite est bien entendu la réduction du volume de calcul au moment de la reconnaissance.

Comme cela a été expliqué plus haut, la reconnaissance d'un mot exige normalement le calcul de $C = N\gamma\ I_m\ J_m$ distances locales, soit environ $C(p + 1)$ opérations (multiplication et addition); elle exige aussi C comparaisons de trois nombres.

La quantification des I vecteurs spectraux d'un mot selon $M = 2^m$ classes exige m I comparaisons de vecteurs, soit environ $2\ m\ I(p + 1)$ opérations; les distances locales sont lues simplement dans un tableau; il y a encore C comparaisons de trois nombres.

Pour fixer les idées, supposons $N = 46$, $I_m = J_m = 65$, $m = 6$, $p = 8$, $\gamma = 0.25$. Le nombre d'opérations passe de $46.\frac{1}{4}.(65)^2.9 = 437.288$ à $12.65.9 = 7020$.

Ce gain en temps de calcul est cependant accompagné d'une augmentation du taux d'erreurs. Le tableau suivant récapitule les résultats obtenus (B : quantification par éclatements binaires, Q : quantification globale).

Conditions	Erreurs %	T_{CAL} (s/s)
Sans quantification	1.6	11.3
B(m = 6)	3.1	5.9
B(m = 7)	2.7	5.6
Q(m = 6)	2.3	6.4
Q(m = 7)	2.4	7.8

4. METHODE STATISTIQUE (A.P.)

En l'occurrence, la quantification vectorielle provoque une discrétisation des sons pouvant être émis par chaque état d'un modèle. L'émission est régie par une loi de probabilité discrète qui est déterminée à partir de séquences d'entraînement.

L'algorithme de VITERBI [3] permet de déterminer le "chemin optimal" dans le graphe des transitions : il fournit donc la probabilité maximale avec laquelle un mot à identifier peut avoir été produit par chaque modèle de référence.

Les avantages et les inconvénients de la quantification vectorielle sont donc identiques à ceux observés lors de l'étude de l'algorithme D.T.W. Les résultats préliminaires concernant un système multi-locuteur confirment ce fait; pour une quantification selon 128 classes, le taux d'erreurs est doublé, mais le temps de calcul est divisé par trois par rapport à un système fonctionnant sans quantification.

5. REMARQUE : DISTANCE CEPSTRALE

La procédure de quantification vectorielle est en principe indépendante du choix de la distance entre vecteurs spectraux.

Une investigation de la statistique de la distance de ITAKURA et de celle de la distance cepstrale a été effectuée pour les phonèmes de la langue française [6]; la comparaison du rapport de l'écart-type à la valeur moyenne observée montre que la distance cepstrale permet une meilleure discrimination entre sons voisins.

Il s'agit, en outre, d'une distance au sens strict, propriété que ne présente pas la "distance" d'ITAKURA. Elle a donc été utilisée dans le système de reconnaissance basé sur la modélisation par automates statistiques.

6. REFERENCES

[1] F. ITAKURA,
"Minimum prediction principle applied to speech recognition",
IEEE Trans., ASSP-23, February 1975, pp. 67-72.

[2] H. SAKOE & S. CHIBA,
"Dynamic programming algorithm optimization for spoken word recognition",
IEEE Trans., ASSP-26, February 1978, pp. 43-49.

[3] S.E. LEVINSON et al.,
"An introduction to the application of the theory of probabilistic functions of a MARKOV process to automatic speech recognition",
B.S.T.J., V62, n° 4, April 1983, pp. 1035-1105.

[4] J. LEROUX, C. GUEGUEN,
"A fixed point computation of partial correlation coefficients",
IEEE Trans., ASSP-25, June 1977, pp. 257-259.

[5] R.M. GRAY,
"Vector quantization",
IEEE, ASSP Magazine, April 1984, pp. 4-29.

[6] R. BOITE et al.,
"A comparison of two distance measures for speech recognition",
Proc. ECCTD'85 (Prague).