RECONNAISSANCE DE MOTS ISOLES PAR CARTES TOPOLOGIQUES AUTO-ORGANISATRICES

Abddelkader Amraoui, René Boite

Laboratoire de théorie des circuits et de traitement du signal
Faculté polytechnique de Mons Boulevard Dolez, 31 B7000 BELGIQUE
tel (32) 65.37.41.28 fax (32) 65.37.43.00

RESUME

Une classification efficiente des unités acoustiques, en particulier les phonèmes, est une étape cruciale en vue de la reconnaissance automatique de la parole indépendante du locuteur. Les réseaux de neurones artificiels sont connus pour leur grandes capacités de résolution des problèmes de classification. Un type particulier, en l'occurrence le réseau auto-organisé de Kohonen, est utilisé en même temps en tant que quantificateur vectoriel et aussi en tant que classificateur pour la reconnaissance des dix chiffres. L'approche originale consiste à modéliser chacun des mots du vocabulaire par un réseau qui lui est propre.

ABSTRACT

An efficient classification of acoustic units, particularely phonemes, is a crucial step towards speaker-independent automatic speech recognition. The artificial neural networks have a great capability of resolution of pattern classification. A particular type, the self-organized neural network of Kohonen, is used as vector quantizer and also as classifier to the aim of the ten french digits. An original approach which consits of modelling each word by its own network is presented.

1. INTRODUCTION

La difficulté dans les tâches de reconnaissance automatique de la parole indépendante du locuteur dicte des approches alternatives ou complémentaires par rapport à celles régulièrement utilisées et qui sont basées sur des techniques de représentation et de traitement de connaissances acoustiques , phonétiques et lexicales ou sur des modèles probabilistes de production de la parole. Ceci explique l'intérêt grandissant des chercheurs en traitement de la parole pour les réseaux de neurones artificiels (RNA) vu la facilité déconcertante avec laquelle le cerveau humain arrive à résoudre des tâches assez complexes de reconnaissance de formes (image, parole).

Les RNA cherchent donc à simuler "l'intelligence biologique". Ces modèles sont attractifs par leur capacité d'apprentissage et de généralisation ainsi que par leur possibilités de résolution de problèmes de classifications. En outre, ils offrent l'avantage de traitements distribués qui permet de réaliser des réponses en temps réel : on parle de méthodes connexionistes pour un traitement massivement parallèle.

La carte topologique auto-organisatrice (ou SOM pour Self Organizing Map) est l'un des RNA les plus connus [1]. C'est une technique développée par Kohonen au début des années 80.

Les SOMs ont été utilisées en reconnaissance automatique de la parole principalement en tant que quantificateur vectoriel, ou pour réaliser une répartition initiale afin de construire un classificateur de formes statiques. Les techniques d'apprentissage d'un tel type de RNA sont décrites à la section 2.

Dans cet article, une nouvelle approche en vue de la reconnaissance de mots isolés est décrite. Chacun des mots du vocabulaire est modélisé par une carte auto-organisatrice. Le nombre de phonèmes composant chacun des mots étant relativement réduit, une reconnaissance basée sur une étape préalable de classification phonémique a été pu être accompli.

2. LE RESEAU DE NEURONES AUTO-ORGANISE

Une carte topologique auto-organisatrice ou "carte de Kohonen" est une classe particulière de réseaux de neurones artificiels basée sur un apprentissage compétitif (dans le sens où une compétition s'opère entre les neurones et la plus proche suivant un certain critère est activée). Elle consiste généralement en une grille plane constituée de L cellules ou neurones artificiels [1] [4].

Chaque cellule i est caractérisée par un vecteur de poids (ou de référence) w_i de même dimension N que celle de

l'espace des vecteurs d'entrée et par ses coordonnées dans la grille.

Une forme d'entrée x(t) est présentée à toute les cellules; celle dont le vecteur de poids est le plus proche de x(t) est activée. Le phénomène d'auto-organisation est obtenu par utilisation de relations de voisinage spatial entre les cellules durant la phase d'apprentissage. Non seulement la cellule la plus proche, c, mais aussi son voisinage N_c, sont adaptées, en contraste avec les algorithmes classiques. Un exemple de voisinage est illustré à la figure 1.

La procédure d'entraînement commence par une initialisation des vecteurs de poids à des valeurs aléatoires. Une contrainte évidente est que les vecteurs initiaux soient distincts. Les vecteurs d'entrée x(t) sont présentés successivement au réseau où une adaptation des vecteurs de poids est opérée. Le processus est arrêté après un certain nombre d'itérations (typiquement de 300 à 500 fois le nombre de cellules dans le réseau) ou lorsqu'un critère d'arrêt est satisfait.

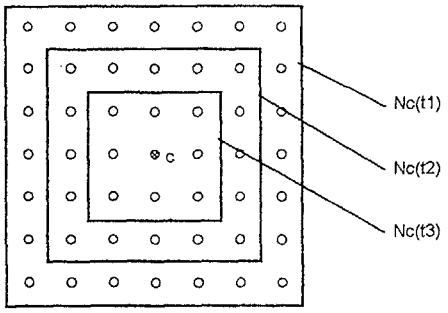

Fig 1 : Exemple de voisinages Nc pour une disposition rectangulaire des cellules, et à différents instants t1,t2,t3.

Pour une bonne évolution des vecteurs de poids, l'expérience montre qu'il est avantageux de prendre N_c très large au début du processus d'apprentissage, ce qui correspond à une résolution grossière et induit d'abord une disposition globale des vecteurs de poids. On réduit le voisinage N_c pour affiner la résolution de la carte; l'ordre topologique acquis précédemment n'est toutefois plus modifié. Il est même possible de terminer l'adaptation avec N_c=c c'est à dire adapter uniquement la cellule élue [1], [5].

L'algorithme d'adaptation d'une carte de Kohonen est le suivant:

1) La cellule élue c est la plus proche de la forme d'entrée x(t) suivant un critère de distance euclidienne:

$$\| x(t) - w_c(t) \| = \min_i \{ \| x(t) - w_i(t) \| \} \quad 1 \le i \le L \qquad (1)$$

2) Les cellules dans le voisinage topologique $N_c(t)$ de la cellule c sont adaptées à l'actuelle forme d'entrée en déplaçant leurs vecteurs de poids $w_i(t)$ vers x(t) suivant la formule suivante:

$$w_i(t+1) = w_i(t) + \alpha(t)[x(t) - w_i(t)] \quad pour\ i \in N_c(t) \qquad (2)$$

$$w_i(t+1) = w_i(t) \qquad ailleurs$$

L'apprentissage est réalisée en deux phases: la première correspondant à une adaptation de la cellule élue et celles situées dans son voisinage. La deuxième phase, beaucoup plus longue, où seule la cellule élue est adaptée.

La figure 2 illustre l'étape d'adaptation du vecteur de poids de la cellule élue et de son voisinage à l'entrée présentée.

A l'issue de ce processus, les cellules du réseau se sont accordées sur les caractéristiques des formes d'entrée présentées:

a) les vecteurs de poids se sont organisés de manière à ce que des cellules proches topologiquement soient sensibles à des signaux physiquement similaires: des neurones voisins sont donc associés à des stimuli voisins. ***Il faut noter que cette répartition particulière des vecteurs de poids s'est opérée bien que la procédure d'apprentissage ait été non supervisée c'est à dire qu' aucune information d'appartenance du vecteur d'entrée à une classe particulière n'a été fournie au réseau : on parle d'auto-organisation.***

b) les valeurs asymptotiques des vecteurs de poids w_i vont définir une quantification vectorielle de l'espace d'entrée. Leur distribution va approcher la fonction de densité de probabilité p(x) des données d'entrée.

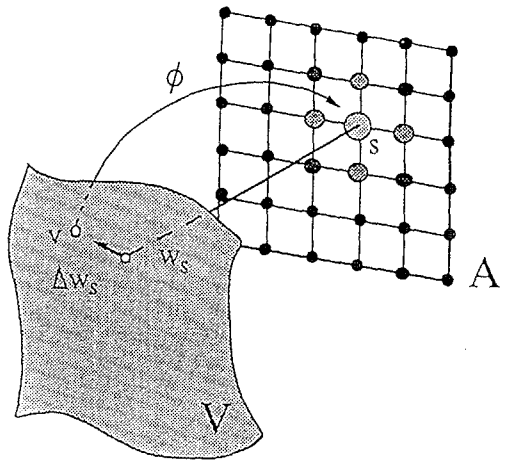

Fig 2 : Le vecteur d'entrée v sélectionne la cellule élue s. Toutes les cellules dans le voisinage de s déplacent leurs vecteurs de poids vers l'entrée v. L'amplitude du déplacement décroît quand la distance à s croît. Dans la figure, cette amplitude est indiquée par différentes tailles et niveaux de gris.

On peut utiliser le réseau après la phase d'entraînement comme classificateur de formes en donnant des étiquettes aux cellules. A cette fin, on compare systématiquement chaque vecteur de référence w_i à un ensemble de vecteurs à classification connue. On affecte une étiquette à chaque cellule suivant un critère de vote majoritaire pour toutes les classes. En général, plusieurs cellules représentent chacune des classes. Les cellules sont ainsi groupés en sous-ensembles, chacun correspondant à une classe particulière des formes d'entrée.

Kohonen, dans sa formulation des algorithmes **LVQ** (**L**earning **V**ector **Q**uantization) cherche à minimiser les erreurs de classification en ajustant de manière itérative les

positions des vecteurs de références. On trouvera dans la référence [1] une description détaillée de l'algorithme LVQ.

3. CLASSIFICATION PHONEMIQUE

Dans la référence [5], un système est présenté, le but duquel est de transcrire de la parole finnoise en une séquence de phonèmes. Le pourcentage de phonèmes reconnus correctement varie autour de 90 %. Cependant, ceci a été réalisé sous certaines conditions favorables qui peuvent être résumées comme suit:

1- les expériences ont été menées dans un contexte mono-locuteur.

1- La langue finnoise ne contient que 21 phonèmes (les plosives ont été traités comme une seule classe, ce qui ramène le nombre de classes phonémiques à discriminer à 19).

3- Les parties stables des phonèmes sont relativement importantes.

Des travaux de reconnaissance de phonèmes, indépendante du locuteur, par utilisation de différentes approches peuvent être cités.

Leung et Zue [6] ont utilisés un RNA du type perceptron multi-couches pour la reconnaissance des 16 voyelles anglaises. Un taux de reconnaissance de 54 % a été obtenu. Dans cette expérience, une segmentation manuelle a été nécessaire en phases d'entraînement et de test. Les performances de reconnaissance augmentaient de façon monotone de 30 % à 54 % quand le nombre des échantillons d'entraînement passait de 80 à 8000.

Nakagawa [7] appliqua une classification de formes statistique et la programmation dynamique pour la reconnaissance de phonème indépendante du locuteur. Il mentionna 51 % et 56 % avec les 2 approches pour la classification de 7 voyelles, et 71 % et 74 % pour 9 fricatives.

K. F. Lee et al [8] obtint un taux de reconnaissance de phonèmes égal à 58.8 % par utilisation de modèles de Markov cachés (HMM) .

Un système de reconnaissance de phonèmes comme étape principale dans le but d'une reconnaissance automatique de la parole, en langue française indépendante du locuteur, achèvera difficilement des performances à même de permettre de réaliser un tel but. Il est alors nécessaire d'intégrer des niveaux de connaissances lexicales supérieures. L'approche présentée dans cet article permet de ne traîter simultanément qu'un nombre réduit de phonèmes.

4. LE SYSTEME DE RECONNAISSANCE

4.1 Base de données et pré-traitement acoustique:

Le système est testé sur un vocabulaire des 10 chiffres de la langue française (zéro, un, deux, trois, quatre, cinq, six, sept, huit, neuf). La base de données est constituée du vocabulaire prononcé par 15 locuteurs mâles (10 pour l'entraînement et 5 pour les tests). Ainsi, la base complète est constituée de 450 expressions orales.

Le signal de parole est échantillonné à 10 Khz, pré accentué avec un filtre dont la fonction de transfert est donnée par $1\text{-}0.95\ z^{-1}$. Ensuite, une fenêtre de Hamming de largeur 30 ms est appliqué toutes les 10 ms. 12 coefficients de prédiction linéaire sont calculés toutes les 10 ms. Enfin, un ensemble de 14 coefficients cepstraux sont tirés à partir des coefficients LPC. Ainsi, le vecteur caractéristique a une dimension égale à 14.

4.2 Modélisation des mots:

En phase d'entraînement, tous les mots sont traités séparément. Le tableau 1 résume les dimensions des cartes et la transcription phonémique pour chaque mot. Il est à noter ici que les plosives (k,t,d) ont été divisés en deux classes phonémiques, l'une représentant la partie de retenue et l'autre l'explosion. L'étiquette silence a aussi été incluse.

	Transcription phonémique	Dimension carte X*X
ZERO	(Z E R O (	7*7
UN	(1 (	4*4
DEUX	(! D 2 (	6*6
TROIS	(- T R W A (	6*6
QUATRE	(- K A - T R (	8*8
CINQ	(S - K (	6*6
SIX	(S I S (	5*5
SEPT	(S 7 - T (	6*6
HUIT	(8 - T (	6*6
NEUF	(N 9 F (	6*6

(: étiquette silence
! : étiquette retenue occlusive voisée
- : étiquette retenue occlusive non voisée

Tableau 1

10 cartes distinctes sont créées par application des étapes suivantes:

1- Une procédure d'apprentissage non supervisée par utilisation des 10 versions du mot. La dimension de la carte (X,X) dépend du nombre de phonèmes différents constituant le mot. Les paramètres des règles d'adaptation (2) sont:

nombre d'itération en phase I = $20\ X^2$
nombre d'itération en phase II = $9 * 20\ X^2$
$\alpha_I(0) = 0.8$
$\alpha_{II}(0) = 0.08$

2- Un étiquetage de chaque cellule de la carte par une technique de vote majoritaire. L'ensemble d'entraînement contenant les 10 versions du mot a été segmenté manuellement et par suite utilisé durant cette étape.

3- Un arrangement direct des positions des vecteurs de références par utilisation du même sous-ensemble de parole segmentée. L'algorithme **LVQ** version 2 a été utilisé avec un nombre d'itération égal à 1000 itérations et le facteur d'adaptation initial égal à 0.05.

A la fin de ce processus, on obtient les 10 cartes qui agissent en même temps en tant que quantificateur vectoriel et aussi en tant que classificateur de phonèmes.

4.3 les tests de reconnaissance:

Le diagramme-bloc de l'étage de reconnaissance est présenté à la figure 3. Le mot inconnu est présenté simultanément aux 10 cartes. Différentes séquences de quasi-phonèmes (SQ) sont générées. Les distorsions moyennes (DM) du mot quantifié par chacune des cartes sont aussi calculées. Grâce au nombre réduit de phonèmes (de l'ordre de 4 par carte), il est suffisant que les SQ soient converties en transcriptions phonémiques (TP) simplement en concaténant les étiquettes identiques qui sont adjacentes et éliminer celles qui n'apparaissent qu'une fois en tant que sons transitoires **[2]**.

Chaque TP est comparée à la transcription phonémique exacte du mot respectif en mesurant la distance d'édition entre les deux chaînes **[3]**. On définit trois opérations élémentaires, dont les combinaisons permettent de transformer une chaîne en une autre. Ces opérations sont : la substitution, l'insertion et la destruction. On associe à chacune un coût individuel. On définit la distance d'édition comme le coût de la suite de transformations élémentaires la moins coûteuse pour transformer une chaîne en une autre.

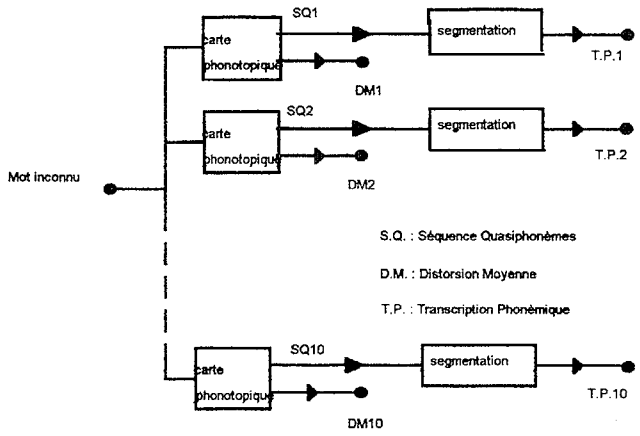

Fig 3 : Diagramme-bloc du système de reconnaissance

Les deux paramètres complémentaires de la distorsion moyenne et la distance d'édition permettent le choix du meilleur candidat comme étant le mot prononcé.

Le taux de reconnaissance obtenu avec la base de tests est égal à 90% c.à d. 15 mots ont été mal reconnus. Tous les 300 mots de la base d'entraînement ont été bien reconnus.

5. Conclusion et perspectives:

Nous avons proposé une approche où les réseaux de neurones artificiels auto-organisés sont utilisés comme quantificateur vectoriel vu qu'une distorsion moyenne du mot quantifié est calculée et aussi en tant que classificateur de formes statiques représentant le vecteur caractéristique d'une tranche de parole de 10 msec en classes phonémiques.

Ces cartes sont intégrés dans un système de reconnaissance des 10 chiffres indépendant du locuteur où elles ont servis à modéliser chacun des mots du vocabulaire.

Les résultats obtenus sont encourageant bien que la base de test utilisée n'est pas assez représentative. Néanmoins, c'est la contrainte de devoir disposer d'une base d'entraînement segmentée et étiquetée (10 versions de chaque mot dans nos expériences) qui a limité la taille de la base d'entraînement et par conséquent une capacité limitée de généralisation.

Le travail actuel porte sur l'étude d'une méthode plus efficace de passage de la séquence de quasi-phonèmes à une transcription phonémique **[2]**.

Une modification du vecteur caractéristique afin de tenir compte du contexte par concaténation de tranches successives améliorerait les résultats mais avec une charge de calcul plus importante **[4]**.

REFERENCES

[1] Teuvo Kohonen. *The self-organizing map.* Proceedings of the IEEE, 78(9):1464-1480, 1990.

[2] M.Kokkonen, K. Torkkola. *Using self-organizing maps and multi-layered feed-forwardnets to obtain phonemic transcriptions of spoken utterances.* Speech communication, vol. 9, n° 5-6, nov 15,1990.

[3] R.A.Wagner, M.J.Fisher. *The string to string correction problem.* JACM, vol. 21, n° 1 pp. 168-173 , 1974.

[4] Jari Kangs. *Phoneme recognition using time-dependant versions of self-organizing maps.* Proceedings of the IEEE 1991 ICASSP, Toronto, canada.

[5] K. Torkkola, J. Kangas, M. Kokkonen, S. Kaski, T. Kohonen. *Status report of the Finnish phonetic typewriter project.* Proceedings of the ICANN, 1991: 771-776,Espoo,Finland.

[6] H.C.Leung and V.W.Zue, " Some phonetic recognition experiments using artificial neural nets," IEEE, ICASSP,avril 1988.

[7] S.Nakagawa : "Speaker-independent phoneme recognition in contiluous speech by a statistical method and a stochastic dynamic time warpin method,",tech. Rep. CMU-CS-86-102, comput.sci.Dep, Carnegie Mellon Univ, jan. 1986.

[8] K.F.Lee and H.W.Hon ,"Speaker-independent phone recognition using hidden Markov models",IEEE, trans. on ASSP, novembre 1989.v