UN RESEAU DE NEURONES POUR LA CLASSIFICATION ET LA RECONNAISSANCE DE LA PAROLE

Lot TCHEEKO

Ecole Nationale Supérieure Polytechnique

BP 8390 Yaoundé (CAMEROUN)

FAX (237) 23-18-41

Abstract : We apply neural networks to test and classify the digital representation of voice signal. The concept of eingenvectors is used to examine the correlation of five phonemes. The results of eingenvector projection for the correlation factors are presented. During training phase, the learning rate is gradually decreased in discrete steps until the network converges to a minimum error solution. Experiments results of the analysis using neural networks indicate that about 94.29 percent of the phonemes was recognized and well classifyed.

Keywords : Neural Networks, Eingenvectors, Phoneme

INTRODUCTION

L'utilisation des réseaux de Neurones (RN) pour la reconnaissance de la parole a fait l'objet de plusieurs travaux.[4, 3, 2] L'originalité de nos recherches provient :

- de l'application des algorithmes d'apprentissage par RN artificiels, sur des représentations numériques issues de l'analyse de tranches de signaux de parole, elles-même obtenues en utilisant une carte d'échantillonnage numérique PC Lab card PcL - 812 d'une part,

- de l'intégration des objets de Turbo Vision aux techniques de la POO qu'offre C++, pour implémenter des algorithmes très performants d'autre part.

Nous avons fait un échantillonnage des signaux de cinq phonèmes correspondant aux cinq voyelles de l'alphabet (a,e,i,o,u) à l'aide de la carte PCL-812 à la fréquence de 10 khz. Nous avons généré à partir des échantillons de ces voyelles, des facteurs de corrélation qui sont des vecteurs de dimension 12 contenant la plus part du contenu phonémique du signal, lesquels sont ensuite présentés aux RN. Mais auparavant, nous avons fait une analyse en composantes principales (ACP) de ces facteurs de corrélation afin de mieux juger les résultats obtenus ultérieurement. Plusieurs configurations de RN ont été simulées afin d'apprécier leurs efficacités respectives.

II. CLASSIFICATION ET FACTEURS DE CORRELATION

Des hypothèses simplifiées sur la structure du signal sont faites pour réduire la complexité des algorithmes nécessaires à la classification souhaitée.[1]. Dans le réseau multicouches utilisé pour la classification, le nombre de neurones sur la couche de sortie correspond au nombre de classe attendue[4].

En appliquant la tranformée de Fourier rapide (FFT) à notre signal, puis en appliquant la FFT inverse à la distribution énergétique en fréquence, nous obtenons par cette méthode 16 facteurs de corrélation par segment de 128 échantillons traités.

Nous avons préferé une autre méthode consistant en la détection des pics, ie qu'au lieu d'utiliser des segments constants de 120 échantillons, nous recherchons dans notre signal, les pics correspondant aux débuts et fins de répétition des tranches de signaux. Ceci dans le but d'affiner les résultats, compte tenu du fait que la périodicité du signal (12 ms) constitue une valeur moyenne de la période de répétition qui peut varier beaucoup plus, particulièrement chez les enfants et les femmes. La paramétrisation par facteurs de corrélation présente un réel gain en terme de place mémoire. En effet, nous passons de 120 échantillons codés chacun sur 10 bits (soit 150 octets) à 12 facteurs de corrélation chacun sur 4 octets (soit 48 octets). De plus, ces facteurs contiennent presque tout le contenu phonèmique du signal, condition nécessaire pour pouvoir différencier suffisamment les phonèmes étudiés.

III. ALGORITHMES NEURO-MIMETIQUES IMPLEMENTES

Dans nos travaux nous avons implémenté quatre algorithmes d'apprentissage correspondant à 4 types de réseaux : multi-couches, adaline, LVQ et cartes topologiques.

Dans notre implémentation, la couche de sortie est totalement connectée à celle cachée, et permet la codification des sorties des couches cachées compte tenu du fait que cette dernière est libre de construire sa représentation des données. Pour ce faire, elle dispose d'un ensemble de neurones dont la fonction de transition d'état est composée d'une somme pondérée suivie de la fonction identité. Sa règle d'apprentissage est régi par la loi de Widrow-Hoff.

IV. RESULTATS OBTENUS

La simulation a été réalisé en C++ avec une interface Homme-machine issue de l'utilisation des objets Turbo-Vision, sur un Z-386 à 16 Mhz.

Après acquisition des données, celles-ci sont traitées afin de produire les facteurs de corrélation associés, puis les signaux des cinq phonèmes (/a/, /e/, /i/, /o/, /u/) sont analysés par une ACP, puis projetés sur le plan principal deux à deux afin de retenir le maximum de variance. Cette projection nous permet d'avoir une idée qualitative de la séparation des nuages de points associés. Les résultats relèvent :

- une bonne séparation entre les couples de phonèmes (/a/,/e/), (/a/,/i/), (/a/,/o/), (/a/,/u), (/e/,/o/), et (/e/,/u/). Ce qui devrait permettre de déboucher sur une bonne identification entre eux par les RN.

- une faible séparation des couples (/i/,/o/) et (/i/,/u/), ce qui laisse entrevoir un problème de séparation de ces phonèmes

- une très mauvaise séparation du couple (/o/,/u/). Compte tenu du pourcentage de variante totale très élevé (99,4 %) contenu dans le plan principal, il faut entrevoir un réel problème de séparation de ces phonèmes.

Procédure expérimentale pour le RN

Au cours de l'apprentissage, nous présentons aux différentes configurations de RN deux ensembles d'apprentissage contenant des vecteurs dont les douze premières composantes sont les coefficients de corrélation des signaux étudiés. La dernière composante est une étiquette permettant d'identifier leur classe d'appartenance. Au cours de l'évaluation, nous soumettons aux différentes configurations de RN l'ensemble 2 à partir desquels nous générons des pourcentages de reconnaissance par phonème que nous regroupons dans une matrice de confusion.

Tableau : Récapitulatif des meilleurs % de reconnaissance par phonème et par type d'architecture.

ARCHITECTURE	MLP	LVQ	CARTE TOPO.
a	99.60	97.60	95.60
e	99.66	96.93	65.87
i	99.35	97.71	66.67
o	88.00	48.31	0.00
u	84.88	0.00	100.00
% MOYEN	94.29	68.11	65.63

CONCLUSIONS.

Les résultats obtenus, particulièrement avec le réseau multi-couche sont très intéressants, et nous appellent à d'avantage de travail afin d'améliorer ses performances par des modifications augmentant notablement la vitesse d'apprentissage. Nous pensons utiliser à cet effet soit des méthodes complexes qui modifient les coefficients synaptiques par calcul préalable du meilleur taux d'apprentissage ; soit un algorithme plus complexe qui ajoute des neurones cachées en cours d'apprentissage. La méthode par paramétrisation par facteurs de corrélation donne des résultats globalement satisfaisants pour les cinq phonèmes étudiés. Après avoir augmenté la vitesse d'apprentissage et validé la paramétrisation, nous envisageons d'etendre le système réalisé à la localisation dans une phrase numérisée, de la position des phonèmes étudiés.

BIBLIOGRAPHIE

[1]. K.S.FU, P.J. MIN AND T.J.LI, 1970 : "Feature selection in pattern recognition" IEEE Trans. Syst. Sci. Cybern., vol. SSC-6, pp 27-33 Jan. 1970

[2]. R. Paul Gorman & Terrence J. Sejnowski , 1988 : "Learned Classification of sonar targets using a massivelly parallel Networks" IEEE Transaction on acoustics, speech, and signal processing, vol 36, N° 7 july 1988

[3]. T. Kohonen, 1988 : "An introduction to Neural computing", Neural Networks, vol 1, pp 3 - 16, 1988

[4]. S. K. Pal, Sushmita Mitra, 1992 : "Multilayer Perceptron, Fuzzy sets, and classification" IEEE Transaction on Neural Networks, vol 3, N° 5, PP 683-696 Sept 92.

UN RESEAU DE NEURONES POUR LA CLASSIFICATION ET LA RECONNAISSANCE DE LA PAROLE

Lot TCHEEKO

Ecole Nationale Supérieure Polytechnique

BP 8390 Yaoundé (CAMEROUN)

FAX (237) 23-18-41

Abstract : We apply neural networks to test and classify the digital representation of voice signal. The concept of eingenvectors is used to examine the correlation of five phonemes. The results of eingenvector projection for the correlation factors are presented. During training phase, the learning rate is gradually decreased in discrete steps until the network converges to a minimum error solution. Experiments results of the analysis using neural networks indicate that about 94.29 percent of the phonemes was recognized and well classifyed.

Keywords : Neural Networks, Eingenvectors, Phoneme

INTRODUCTION

L'utilisation des réseaux de Neurones (RN) pour la reconnaissance de la parole a fait l'objet de plusieurs travaux.[4, 3, 2] L'originalité de nos recherches provient :

- de l'application des algorithmes d'apprentissage par RN artificiels, sur des représentations numériques issues de l'analyse de tranches de signaux de parole, elles-même obtenues en utilisant une carte d'échantillonnage numérique PC Lab card PcL - 812 d'une part,

- de l'intégration des objets de Turbo Vision aux techniques de la POO qu'offre C++, pour implémenter des algorithmes très performants d'autre part.

Nous avons fait un échantillonnage des signaux de cinq phonèmes correspondant aux cinq voyelles de l'alphabet (a,e,i,o,u) à l'aide de la carte PCL-812 à la fréquence de 10 khz. Nous avons généré à partir des échantillons de ces voyelles, des facteurs de corrélation qui sont des vecteurs de dimension 12 contenant la plus part du contenu phonémique du signal, lesquels sont ensuite présentés aux RN. Mais auparavant, nous avons fait une analyse en composantes principales (ACP) de ces facteurs de corrélation afin de mieux juger les résultats obtenus ultérieurement. Plusieurs configurations de RN ont été simulées afin d'apprécier leurs efficacités respectives.

II. CLASSIFICATION ET FACTEURS DE CORRELATION

Des hypothèses simplifiées sur la structure du signal sont faites pour réduire la complexité des algorithmes nécessaires à la classification souhaitée.[1]. Dans le réseau multicouches utilisé pour la classification, le nombre de neurones sur la couche de sortie correspond au nombre de classe attendue[4].

En appliquant la tranformée de Fourier rapide (FFT) à notre signal, puis en appliquant la FFT inverse à la distribution énergétique en fréquence, nous obtenons par cette méthode 16 facteurs de corrélation par segment de 128 échantillons traités.

Nous avons préferé une autre méthode consistant en la détection des pics, ie qu'au lieu d'utiliser des segments constants de 120 échantillons, nous recherchons dans notre signal, les pics correspondant aux débuts et fins de répétition des tranches de signaux. Ceci dans le but d'affiner les résultats, compte tenu du fait que la périodicité du signal (12 ms) constitue une valeur moyenne de la période de répétition qui peut varier beaucoup plus, particulièrement chez les enfants et les femmes. La paramétrisation par facteurs de corrélation présente un réel gain en terme de place mémoire. En effet, nous passons de 120 échantillons codés chacun sur 10 bits (soit 150 octets) à 12 facteurs de corrélation chacun sur 4 octets (soit 48 octets). De plus, ces facteurs contiennent presque tout le contenu phonèmique du signal, condition nécessaire pour pouvoir différencier suffisamment les phonèmes étudiés.

III. ALGORITHMES NEURO-MIMETIQUES IMPLEMENTES

Dans nos travaux nous avons implémenté quatre algorithmes d'apprentissage correspondant à 4 types de réseaux : multi-couches, adaline, LVQ et cartes topologiques.

Dans notre implémentation, la couche de sortie est totalement connectée à celle cachée, et permet la codification des sorties des couches cachées compte tenu du fait que cette dernière est libre de construire sa représentation des données. Pour ce faire, elle dispose d'un ensemble de neurones dont la fonction de transition d'état est composée d'une somme pondérée suivie de la fonction identité. Sa règle d'apprentissage est régi par la loi de Widrow-Hoff.

IV. RESULTATS OBTENUS

La simulation a été réalisé en C++ avec une interface Homme-machine issue de l'utilisation des objets Turbo-Vision, sur un Z-386 à 16 Mhz.

Après acquisition des données, celles-ci sont traitées afin de produire les facteurs de corrélation associés, puis les signaux des cinq phonèmes (/a/, /e/, /i/, /o/, /u/) sont analysés par une ACP, puis projetés sur le plan principal deux à deux afin de retenir le maximum de variance. Cette projection nous permet d'avoir une idée qualitative de la séparation des nuages de points associés. Les résultats relèvent :

- une bonne séparation entre les couples de phonèmes (/a/,/e/), (/a/,/i/), (/a/,/o/), (/a/,/u), (/e/,/o/), et (/e/,/u/). Ce qui devrait permettre de déboucher sur une bonne identification entre eux par les RN.

- une faible séparation des couples (/i/,/o/) et (/i/,/u/), ce qui laisse entrevoir un problème de séparation de ces phonèmes

- une très mauvaise séparation du couple (/o/,/u/). Compte tenu du pourcentage de variante totale très élevé (99,4 %) contenu dans le plan principal, il faut entrevoir un réel problème de séparation de ces phonèmes.

Procédure expérimentale pour le RN

Au cours de l'apprentissage, nous présentons aux différentes configurations de RN deux ensembles d'apprentissage contenant des vecteurs dont les douze premières composantes sont les coefficients de corrélation des signaux étudiés. La dernière composante est une étiquette permettant d'identifier leur classe d'appartenance. Au cours de l'évaluation, nous soumettons aux différentes configurations de RN l'ensemble 2 à partir desquels nous générons des pourcentages de reconnaissance par phonème que nous regroupons dans une matrice de confusion.

Tableau : Récapitulatif des meilleurs % de reconnaissance par phonème et par type d'architecture.

ARCHITECTURE	MLP	LVQ	CARTE TOPO.
a	99.60	97.60	95.60
e	99.66	96.93	65.87
i	99.35	97.71	66.67
o	88.00	48.31	0.00
u	84.88	0.00	100.00
% MOYEN	94.29	68.11	65.63

CONCLUSIONS.

Les résultats obtenus, particulièrement avec le réseau multi-couche sont très intéressants, et nous appellent à d'avantage de travail afin d'améliorer ses performances par des modifications augmentant notablement la vitesse d'apprentissage. Nous pensons utiliser à cet effet soit des méthodes complexes qui modifient les coefficients synaptiques par calcul préalable du meilleur taux d'apprentissage ; soit un algorithme plus complexe qui ajoute des neurones cachées en cours d'apprentissage. La méthode par paramétrisation par facteurs de corrélation donne des résultats globalement satisfaisants pour les cinq phonèmes étudiés. Après avoir augmenté la vitesse d'apprentissage et validé la paramétrisation, nous envisageons d'etendre le système réalisé à la localisation dans une phrase numérisée, de la position des phonèmes étudiés.

BIBLIOGRAPHIE

[1]. K.S.FU, P.J. MIN AND T.J.LI, 1970 : "Feature selection in pattern recognition" IEEE Trans. Syst. Sci. Cybern., vol. SSC-6, pp 27-33 Jan. 1970

[2]. R. Paul Gorman & Terrence J. Sejnowski , 1988 : "Learned Classification of sonar targets using a massivelly parallel Networks" IEEE Transaction on acoustics, speech, and signal processing, vol 36, N° 7 july 1988

[3]. T. Kohonen, 1988 : "An introduction to Neural computing", Neural Networks, vol 1, pp 3 - 16, 1988

[4]. S. K. Pal, Sushmita Mitra, 1992 : "Multilayer Perceptron, Fuzzy sets, and classification" IEEE Transaction on Neural Networks, vol 3, N° 5, PP 683-696 Sept 92.