MODELISATION NEURO-PREDICTIVE POUR LA CLASSIFICATION PHONETIQUE DE LA LANGUE ARABE

M.DIDICHE[1] & A.DEBILOU [2]

Laboratoire de l'identification commande, contrôle et communication l'I3C
Université Med Khider Biskra ALGERIE
m.didiche@hotmail.fr

RÉSUMÉ

La modélisation neuro-prédictive pour la classification phonétique de la langue arabe est une branche de la reconnaissance automatique de la parole qui suppose l'application de deux processus fondamentaux : la para métrisation du signal de parole et le décodage phonétique. La para métrisation acoustique a pour but d'extraire l'information pertinente du signal acoustique afin de fournir une description aussi complète et représentative que possible. Sa mise en œuvre repose essentiellement sur des méthodes issues du traitement du signal. Le décodage phonétique consiste à classifier des formes acoustiques en vue de la reconnaissance. Différents décodeurs ont été utilisé (Modèles de Markov Cachés, Ondelettes, Réseaux de neurones……etc.) ayant chacun leur propre algorithme d'apprentissage. Dans la littérature, nous avons constaté que le processus de paramétrisation acoustique et le processus décodage (classification) utilisent des critères d'optimisation différents, alors qu'ils ont pour objectif commun la reconnaissance des unités phonétiques. Dans ce papier, Nous allons procéder à la mise en forme d'un signal de parole que l'on injectera dans un réseau de neurone MLP(Multi Layer Perceptron) et ensuite faire une comparaison entre les résultats obtenus par les MFCC()Mel Frequency Cepstral Coding et NPC(Neuronal Predictive Coding).Notre contribution sera sans doute l'implication da le langue arabe dans ce processus.

Mots Clés : extraction de caractéristique, reconnaissance de la parole, MFCC, NPC et MLF

1 INTRODUCTION

Dans l'objectif d'améliorer les résultats en reconnaissance de la parole, plusieurs méthode peuvent être adopté. L'une d'elle est l'extraction de caractéristiques, dont les vecteurs d'observation sont déterminés à l'aide de méthode temporelles comme le LPC (linear Predictive coding) ou MFCC (Mel Frequency Cepstral Coding). De nombreux travaux ont montré l'importance de l'étape d'extraction de caractéristique [4,1,8] qui est un élément de mise au point d'un système de reconnaissance.

Mais comme la parole n'est pas stationnaire, donc notre approche est une extension au domaine non linéaire du codage LPC par les réseaux de neurones multicouche. Les réseaux de neurones sont des outils puissants et commodes pour résoudre des problèmes complexes.

L'étape la plus importante lors de la construction d'un système de reconnaissance connexionniste est l'apprentissage, qui consiste a mettre a jours les valeurs des connexions de ce réseaux afin de réussir la tache qui lui est assigné. Sachant que cet apprentissage est basé sur l'algorithme de la rétro propagation du gradient. Dans cette article, nous nous intéressons aux traitements de l'information vocal et aux algorithmes pour l'extraction des LPC, MFCC, FFT, et enfin MLP.

2 TRAITEMENT DE L'INFORMATION VOCAL

2.1 PRETRAITEMENT

Le but des prétraitements acoustiques pour la reconnaissance de la parole est de réduire la quantité d'information du signal de parole et de faciliter la classification en mettant en évidence des invariants par rapport aux événements acoustiques. La procédure d'isolation consiste à déterminer les limites du mot en réalisant une segmentation parole/non parole du signal vocal car il est assez coûteux en temps et inutile d'analyser la totalité du signal. On utilisera les critères d'énergie et de passage par zéro. L'introduction d'une pré-emphase sur le signal de parole avant son analyse tient au souci de rehausser les amplitudes fréquentielles faibles par rapport aux hautes amplitudes afin de tenir compte de l'ensemble du signal. Le procédé le plus simple, c'est d'appliquer un filtre de préaccentuation donnée par fonction de transfert.

2.1.1 PREACCENTUATION :

La préaccentuation est une opération de filtrage d'un signal de parole s(n) avec un filtre dont la fonction de transfert H(Z) est donnée par :

$$\mathbf{H(Z) = 1 - \mu Z^{-1}}$$

La valeur la plus utilisée pour **μ** est **0,95**.

Si s(n) est le signal de la parole préaccentué alors : $S_p(n)=s(n) – 0,95\ S(n-1)$

Cette opération permet d'accentuer les hautes fréquences du signal.

2.1.2 FENETRAGE

Le signal de la parole est de nature non stationnaire, il est donc nécessaire, avant d'extraire les paramètres de la reconnaissance de la subdiviser en segments. Cette étape permet d'obtenir pour chaque segment de parole un signal quasi stationnaire.

Les discontinuités aux extrémités des segments peuvent être amoindries en multipliant chaque segment par une fenêtre de Hamming. La fenêtre de Hamming est donnée par l'équation suivante :

$$s_2(n) = s_1(n).(0,54+0,46.\cos(\frac{2\pi n}{N-1} - \pi))$$

Où N est le nombre d'échantillons du segment**.**

2.1.3 FFT :

Cette étape transforme le signal de parole en domaine fréquentiel [1] avec la formule :

$$X_n=\sum_{k=0}^{N-1} x_k e^{-\frac{2\pi jkn}{N}}, \quad n=0,1\ldots,N-1 \quad \text{où } j=\sqrt{-1}$$

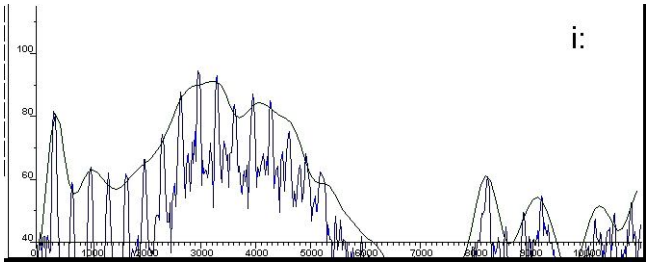

Figure 2: FFT (bleu riche en harmonique) et Cepstre (vert) pour le phonème i: d'une locutrice

2.2 CALCUL DES COEFFICIENTS

Dans cette étape, nous survolerons les deux méthodes les plus utilisées d'extraction des paramètres. La première est basée sur le principe de production de la parole et la deuxième sur le principe de perception de la parole.

2.2.1 CODAGE LINEAIRE PREDICTIF LPC

Le codage linéaire prédictif est une méthode d'extraction basée sur le principe qu'un échantillon du signal peut être estimé à l'aide d'une composition linéaire de p échantillons précédentes. Elle est fondée sur un modèle Autorégressif [1,8].

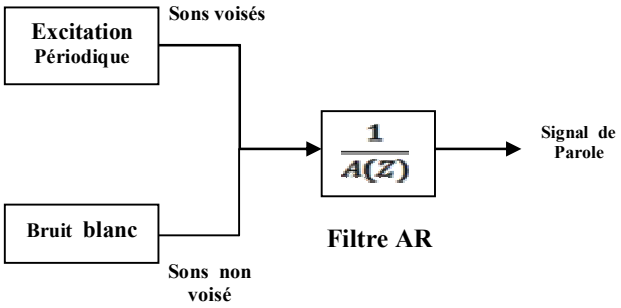

Figure 1: Modèle autorégressif de production de la parole

Comme le conduit vocale peut être assimilé à un filtre récursif, à pôle seulement défini par :

$$y(z) = \frac{G}{A(Z)} \times U(Z) \quad \text{Et} \quad A(z) = 1 - \sum_{i=1}^{p} a_i z^{-i}$$

Ou : G est le gain, G/A(Z) est la fonction de transfert du filtre, les paramètres a_i sont les coefficients de prédiction, Y(Z) est le signal de sortie du filtre et U(Z) représente le signal d'excitation, qui est formé d'impulsion périodique pour le son voisé.

L'équation donne :

$$G\ U(z) = Y(Z) - Y(Z)\sum_{i=1}^{p} a_i z^{-i}$$

Dans le domaine temporel, l'équation devient :

$$G\ U(n) = Y(n) - \sum_{i=1}^{p} a_i\, y^{(n-i)} = Y(n) - \hat{Y}(n)$$

Ou y est le signal estimé à partir de la composition linéaire des p échantillons passés.

Pour l'estimation des paramètres a_i, on doit minimiser l'erreur quadratique suivante :

$$E_n = \sum_{m=0}^{n-1+p} [Y_n(m) - \hat{Y}_n(m)]^2$$

$$\frac{\delta E_n}{\delta a_k} = 0, 1 \le k \le p$$

Pour résoudre ces équations, on peut utiliser la matrice d'autocorrélation qui est une matrice de Toeplitz. Elle permet de résoudre l'équation à l'aide de l'algorithme Levinson-Durbin comme suite :

- Initialisation **: $E^{(0)}$= r(0)**

Bo $k_i = \dfrac{r(i) - \sum_{j=1}^{i-1} a_j^{(i-1)} r(|i-j|)}{E^{i-1}}$

Avec : $1 \le i \le p$, $a_j^{(i)} = k_i$

$$a_j^{(i)} = a_j^{(i-1)} - k_i \alpha_{(i-j)}^{(i-1)} \quad E^{(i)} = (1 - k_i^2) E^{(i-1)}$$

- Coefficients : $a_j = a_i^p, 1 \le j \le p$

$$A_t(z) = \sum_{i=0}^{M} a_t(i) z^{-i}, a_t(0) = 1$$

2.2.2 MFCC

La methode MFCC est une méthode d'extraction des paramètres selon l'echelle MEL. En effet, la perception de la parole par le système auditif humain est fondée sur une echelle fréquentielle semblable a l'echelle MEL. Cette echelle est linèaire aux basses fréquences et logarithmique en hautes fréquences et elle est donnée selon l'equation suivante:

$$Mel(zf) = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$

B = 2595 et **C = 700**, **f** représente la fréquence

Après une préaccentuation et une subdivision en différents segments avec fenêtrage du signal de parole, on applique la méthode MFCC qui consiste à calculer la transformée de Fourier de chaque segment. Puis à utiliser des filtres triangulaire, espacés suivant l'échelle Mel, pour filtrer cette transformée est obtenir les énergies à partir du module au carré de la transformée de Fourier.

Finalement, on calcul la transformée discrète en cosinus (DCT) des logarithmes des énergies obtenues par les filtres triangulaires afin d'extraire les coefficients MFCC utilisés pour la reconnaissance. Ces coefficients sont donnés par l'équation suivante:

$$C_i = \sqrt{\frac{2}{k}} \sum_{j=1}^{k} (\ln(m_j)) \cos\left(\frac{\pi i}{K}(j-0.5)\right)$$

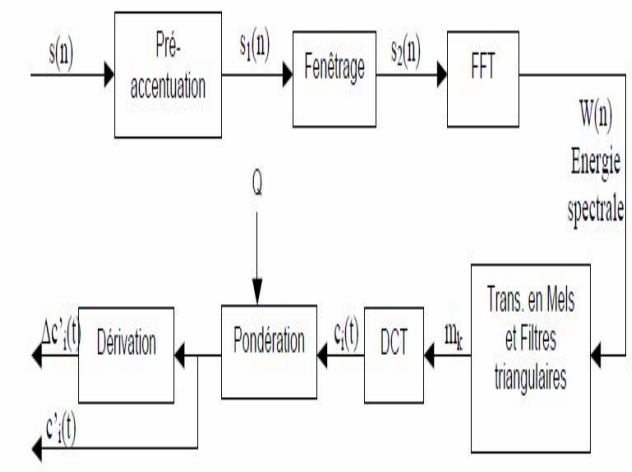

Figure 3: étapes de calcul des coefficients MFCC

3 METHODE UTILISEE

Le codage neuro-prédictif est une extension du codage LPC, donc une méthode de codage temporelle, le codeur NPC extrait les caractéristiques non linéaires d'un phonème. Il est basé sur un MLP à une couche cachée suivi d'une couche de sortie à 1 neurone appelé cellule de prédiction. L'étape d'apprentissage consiste à prédire un échantillon (extrait du signal acoustique d'un phonème) à partir des n échantillons précédents. Tous ces coefficients sont injecté dans un réseau de neurone MLP [6,7] qui doit déterminer l'erreur quadratique afin de réajuster les poids de la couche d'entré jusqu'à pouvoir avoir une erreur désirée acceptable, cette algorithme est basé sur la rétro propagation du gradient.

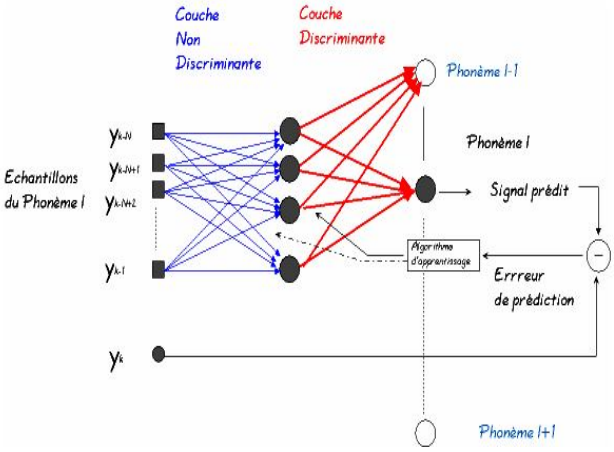

Figure 4 : Codeur NPC

La rétro propagation de l'erreur:

$$W_{ij}^{l}(k+1) = W_{ij}^{l}(k) - \Delta W_{ij}^{l}$$

4 RESULTAT ET DISCUTION

Afin de tester les performances de notre système, nous allons l'appliquer sur notre base de données. Pour cela, nous allons extraire une suite de vecteurs de coefficients (MFCC/NPC) de nos fichiers wav, et lancer le mécanisme (apprentissage/test).

4.1 Base de voyelle avec un MLP

Commentaire

Sur 1000 itérations avec 12 coefficient NPC, et 16 coefficient MFCC. Les résultats montrent une supériorité des MFCC que soit en taux d'apprentissage (89%MFCC, 85%NPC) ou en taux de reconnaissance (66%MFCC, 58%NPC). Le codage MFCC étant une méthode fréquentielle arrive mieux à reconnaître les voyelles que le codage NPC.les coefficients MFCC montre une meilleure résistance au bruit que les coefficients NPC.

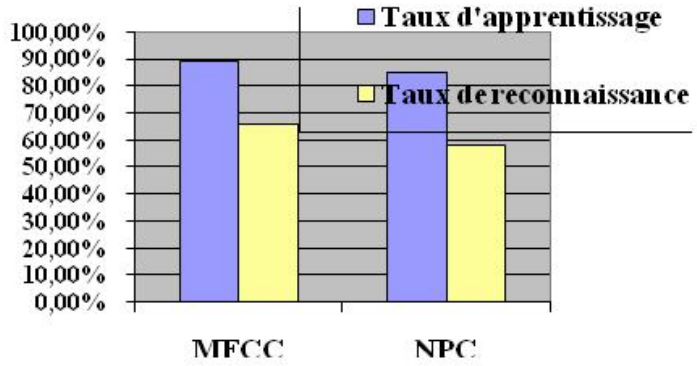

Figure 5: résultat avec MFCC et NPC

4.2 Base de mot avec un MLP

MFCC : Taux de reconnaissance sur les consonnes

Tableau 1: résultat par MFCC

ب	ض د	ط ت	ک	ق	ظ ذ	ع	ص س	م	ن	ل	ر
15%	46%	55%	33%	13%	85%	0%	98%	71%	97%	97%	67%

NPC : Taux de reconnaissance sur les consonnes

Tableau 2: résultat par NPC

ب	ض د	ط ت	ک	ق	ظ ذ	ع	ص س	م	ن	ل	ر
72%	62%	82%	62%	68%	94%	63%	98%	73%	63%	65%	69%

Commentaire

Une large avance en taux de reconnaissance NPC pour les plosives, ce qui veut dire qu'on arrive à mieux prendre en compte les phonèmes rapides dans le temps, contrairement au codage MFCC qui arrive plutôt à prendre en compte les consonnes de longue durée comme les nasales.

5 CONCLUSION

J'ai présenté la mise en œuvre de deux méthodes de codage destinées à la reconnaissance de parole et qui sont les paramètres qui se sont imposés lors de l'extraction des caractéristiques dans l'analyse de la parole. A savoir le NPC tiré de la phase de production et le MFCC tiré de la perception. Cette étude nous a montré l'importance du contexte en reconnaissance de la parole, donc là où des classifieurs ne prennent pas en compte le temps tel que le MLP et qui ont montré des faiblesses en taux de reconnaissance.

REFERENCE

[1] Maïtine Bergounioux «Mathématiques pour le traitement du signal :cours et exrcices corrigés » Dunod, paris 2010 p :270-279

[2] Z.Hamaiza et M.Bedda « Analyse et synthése de la parole Arabe » université annaba et biskra, 2011

[3] Abdelkader Benyattou et Hiba Khelil « Application du réseau de neurone gamma à la reconnaissance de la parole » USTO oran SETIT 2007

[4] O.Deroo « Modèles Dependant du context et Méthodes de Fusion de données Appliquées à la reconnaissance de la parole par Modèles Hybrides HMM/MPL, ''Faculté Polytechnique de Mons, 1998''

[5] Rimah Amani, Dorra Ben Ayed et Noureddine Ellouz « Application de la méthode Adaboost à la reconnaissace Automatique de la parole » département de genie electrique, ENIT tunis Tunisie 2011

[6] Patrice Wira « Réseaux de Neurones artificiels : architectures et applications », université de haute Alsace, laboratoire MIPS avril 2009, p : 32, 49-56,

[7] Claude Touzet "Introduction au connexionnisme: cours, exercices et travaux pratiques" juillet 1992, p: 65-67,112

[8] M.Bellanger « Traitement Numérique du signal : Théorie et pratique » edition Masson 1987, p : 363

[9] D.E.Kouloughli « Grammaire de l'arabe d'aujourd'hui 2001»

[10] Tahar Saidane, Mounir Zrigui et Mohamed Ben Ahmed «La transcription orthographique phonétique de la langue arabe 1999»

[11] AMROUCHE A., DEBIECHE M., TALEB-AHMED A., (2010). An efficient speech recognition system in adverse conditions using the nonparametric regression. Engineering application of artificiel intelligence, 23(1), pp 85-94.

[12] AMROUCHE A., TALEB-AHMED A., ROUVAEN. J-M., YAGOUB M. (2009) improvement of the speech recognition in noisy environments using a nonparametric regression. International journal of parallel , emergent and distributed system, vol 34, issue 1, pp .49-67 …