Conception d'un Système pour La Reconnaissance De Mots Enchaînés Arabes

H.DAHMANI (1), M.BEDDA (2)
(1) Laboratoire D'Automatique Et Signaux, Université BADJI-Mokhtar, Annaba, Algérie
Enseignante à L'Université de M'sila
FAX : 0213 35 556410/550404
E.mail : habiba.dahmani@caramail.com
(2) Laboratoire D'Automatique Et Signaux ,Université BADJI-Mokhtar, Annaba, Algérie
E.mail : mouldi_bedda@yahoo.fr

Résumé

Pour la conception d'un système de reconnaissance des chiffres arabes enchaînés (des chaînes comportant deux chiffres et trois chiffres), nous avons utilisé l'algorithme de Bridle et Nqkagawa : la programmation dynamique en une seule passe (the one-pass dynamic programming) qu'est l'approche la plus utilisée actuellement dans les systèmes de reconnaissance par rapport à celle de Sakoe (the two level dynamic programming) et celle de Myers et « the level building dynamic programming) tout en essayant de résoudre le problème principal de la reconnaissance enchaînée ou continue ; l'effet de la coarticulation.

Mots clés

La reconnaissance enchaînée, the one-pass algorithm, la programmation dynamique

1. Introduction

La parole est l'un des moyens les plus naturels par lequel des personnes communiquent. Cependant, à ce jour, la commande de machines est en général effectuée par des gestes ou par le truchement de langages artificiels, tels les langages de programmation ou les commandes des automatismes. Pour des raisons de facilité d'interaction, l'homme a depuis longtemps été tenté de concevoir des machines dont les commandes seraient directement activées par la parole. Dans cette perspective, il s'agit de développer des systèmes capables de reconnaître la parole, de la comprendre et d'exécuter les actions résultant de la compréhension du message.
La reconnaissance de mots enchaînés est un des domaines de la reconnaissance de la parole les plus prometteurs et les plus passionnants. Ce constat est dû aux nombreux avantages de ce type de reconnaissance par rapport à la reconnaissance de mots isolés nécessitant pour leur bon fonctionnement qu'il y ait un silence entre les mots, les systèmes de reconnaissance de mots enchaînés autorisent une élocution continue. Le débit d'information avec de tels systèmes est par conséquent supérieur à celui que l'on obtient avec les systèmes de reconnaissance de mots isolés.
Ainsi les systèmes de reconnaissance de mots enchaînés se trouvent tout à fait adapté à la commande de machines complexes et par conséquent devraient intéresser un nombre de secteurs importants.

2. Détection de la parole

La détection de fin de mots après l'acquisition du signal est une opération indispensable pour la construction d'une base de données, identification et reconnaissance. Nous avons utilisé un algorithme de détection de fin de mots qui se base sur les deux mesures de l'énergie et le taux de passage par zéro (L.R. Rabiner, M.RSAMBUR, 1975).

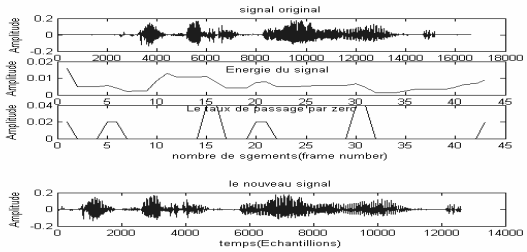


Figure 1 : La détection du début et fin de la chaîne 021

3. Modélisation Acoustique

La chaîne d'analyse suit les étapes suivantes : échantillonnage (fs=11025hz), un filtre de pré-accentuation de fonction de transfert $H(z)=1-0.95z^{-1}$. Puis une analyse à court terme est donc réalisée sur une fenêtre glissante de longueur N=256 échantillons (~23ms) sur laquelle le signal est supposé quasi-stationnaire, pour un pas de recouvrement entre deux trames successives égal à N/2(128). Chaque trame est présentée par 12 coefficients MFCC.
Nous avons utilisé la fenêtre de Hamming donnée par : $s_n = w_n . s_n$ (1)
avec : $w_n = 0,54 - 0,46.\cos(2\pi (n/ N –1))$ (2)
Nous avons utilisé la distance euclidienne qui donne a priori la même importance à chacun des coefficients.

4. Application de la PD à la reconnaissance enchaîné

Les variations de la vitesse d'élocution introduisent des distorsions non linéaires. Afin de compenser celles-ci, on procède par l'alignement temporel non linéaire qui consiste à construire un chemin de recalage entre deux formes acoustiques. La programmation dynamique qui consiste à calculer récursivement la distance accumulée minimale pour chaque point (i,j) suivant des contraintes locales et d'autres globale (Rabiner et al., 1978).
L'idée de base de la programmation dynamique est : qu'à un point (i,j) on continue juste avec le chemin de la plus petite distance des points suivants (i-1,j-1), (i-1,j) ou (i,j-1).
Si nous désignons par D(i,j)et d(i,j) les distances globale et locale respectivement , donc nous exprimons la formulation mathématique de la PD par la relation récurrente suivante :
$D(i,j)=\min[D(i-1,j),D(i-1,j-1),D(i,j-1)] +d(i,j)$ (3)
Avec la condition initiale : $D(1,1) =d(1,1)$.

Le problème de la reconnaissance de mots enchaînés est beaucoup plus difficile que celui de la reconnaissance de mots isolés pour plusieurs raisons telles que :

- Les distorsions temporelles plus accentuées ici du fait de la plus grande variabilité de la vitesse d'élocution.
- Le phénomène de la coarticulation qui a pour effet de déformer considérablement les zones initiales et terminales de mot.

4.1 Principe de la reconnaissance :

Soit V un ensemble de formes références constituant le vocabulaire de l'application. Soit P une phrase inconnue prononcée par un locuteur à partir du vocabulaire V. un système de reconnaissance de mots enchaînés se doit de déterminer tous les éléments de V se trouvant dans P, la vitesse d'élocution ainsi que le nombre de mots constituant la phrase à reconnaître étant des inconnues du problème.
Plusieurs variantes de programmation dynamique ont été proposées. Les principales sont (J.Dimartino,1984) (L.Rabiner,B.H. Juang, 1993) :

- **L'algorithme de programmation à deux niveaux** : (two-level dynamic programming), développé par Sakoe et Chiba (H.Sakoe, S.Chiba, 1978).
- **L'algorithme à construction de niveaux** : (Level Building) proposé par Myers (C.S.Myers et al., 1980).
- **L'algorithme de programmation dynamique en une passe : (one-pass dynamic time warping)** proposé par Bridle et Nakagawa (J.S Bridle et al., 1982) (S.Nakagawa, 1983).

Le tableau suivant justifie notre choix pour l'algorithme de Bridle et de Nakagawa.

L'algorithme	NDL
A deux niveaux	83200
A niveau de construction	10666
De PD en une passe	6400

Figure 2 : Comparaison entre les trois principaux algorithmes en fonction de leur nombre de distances locales effectuées (NDL) (J.Dimartino, 1984)

L'algorithme de PD en une passe est cependant communément adopté dans les systèmes de reconnaissance actuels. Nous allons le décrire brièvement t dans ce qui suit:

4.2 L'algorithme de Bridle et de Nakagawa : « the one pass » :

L'algorithme « one–pass » est l'algorithme le plus facile en implémentation défini comme extension de l'algorithme DTW. Au lieu d'aligner seulement une seule séquence maintenant on peut aligner toutes les références simultanément comme c'est montré par la figure 4 ci-dessous :

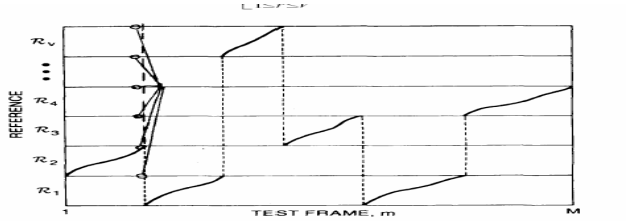


Figure2 : Exemple d'un chemin de recalage trouvé par le one-pass

Le chemin de recalage n'est pas connu durant le processus de la comparaison juste en atteignant la fin de la séquence on peut le trouver en utilisant le bactracking. On note bien que le but du « one-pass » est de trouver le chemin et pas le score de dissemblance.
L'alignement à l'intérieure de la référence k est similaire à celui fait pour les mots isolés, est donnée par la relation récursive déjà rencontrée avec la contrainte locale choisie (dans notre cas la contrainte locale du type III) (H.DAHMANI, 2003).

$$D(i,j,k) = d(i,j,k) + \min (D(i-2,j-1,k), D(i-1,j-1,k), D(i-1,j-2,k)) \quad (4)$$

Où les formes de référence sont indexées par k=1,….K, le temps des trames de formes de références indexée par j=1,…..,J(k) et la forme Test par i=1 ,….I

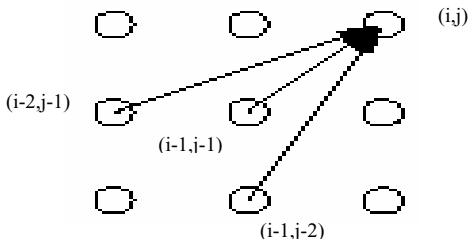
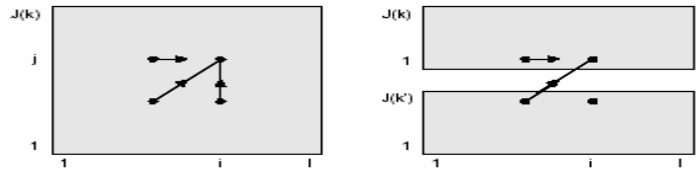


Figure 4: Contrainte locale type III

Pour la limite de la référence (c'est à dire à la trame j=1) une contrainte spéciale est utilisée comme c'est illustré par la figure 6.



La contrainte locale à l'intérieure d'une référence — La contrainte locale entre références

Figure 5 : Différentes contraintes utilisées à l'intérieure et entre références

5. L'apprentissage

Pour l'apprentissage des chiffres enchaînés on a construit deux bases de données :
La base **BAC2** (**B**ase d'**A**pprentissage de **C**haînes de **2** chiffres) **:** contenant les chaînes de chiffres {12, 21, 31, 34, 42} et la base **BAC3** (**B**ase d'**A**pprentissage de **C**haînes de **3** chiffres **):**contient les chaînes {021,024,025,026,027,029} avec la participation de 10 locuteurs (05 femmes et 05 hommes)de 03 occurrences pour chacun et une base des références isolées avec les mêmes locuteurs participant aux bases précédentes.

6. Résultats et Discussion.

Des chaînes comportant deux chiffres

Chaîne à reconnaître	**BAC2/MFCC**
12	93.33
21	96.00
31	63.04
34	76.92
42	96.15
Taux moyen	**85.08**

Figure 6 : Le taux de reconnaissance enchaînée avec la base d'apprentissage.

- **Des chaînes comportant trois chiffre**s

Les premiers résultats obtenus par la base **BAC3**, ont donné un taux de l'ordre de **42 %** de chaînes reconnues. Ce faible score peut s'expliquer par le fait que les mots en contexte subissent de fortes modifications dues aux effets de coarticulation d'un coté et d'un autre coté, les locuteurs participant à l'élaboration de la base n' étaient pas entraînés. Pour résoudre ces problèmes, nous devons segmenter les chaînes, c'est-à-dire déterminer le début et la fin de chacun des mots à l'intérieur d'une chaîne et sélectionner de bons locuteurs.

- **La segmentation**

Au début, on a essayé de prendre la totalité de la base de données des références isolées et la totalité des locuteurs, en appliquant les différentes contraintes (H.DAHMANI, 2003), en

considérant les chaînes qui ne sont pas entièrement reconnues (il suffisait que le nombre de chiffres constituant la chaîne soit reconnus et au moins un seul chiffre soit reconnu). Mais cette méthode ne nous a pas donné satisfaction puisqu'elle a conduit à :

- Des taux de segmentations faibles pour chaque locuteur (la vitesse d'élocution de chaque locuteur influe sur celle des autres locuteurs).
- Des mots de durées irréalistes.

Pour remédier à ces problèmes, nous avons pris en considération ce qui suit :

- Pour augmenter le taux de la segmentation, on a considéré chaque locuteur à part (ses références uniquement comme base d'identification).
- Utilisations de la contrainte type III symétrique ; avec celle-ci, les dilatations temporelles ne peuvent varier que dans l'intervalle [1/2 2].
- Considérer que les chaînes entièrement reconnues.
- **Une première segmentation**.

locuteurs	**Le nombre d'occurrences obtenues par segmentation / MFCC**									
chaîne	**F3**	**F4**	**F5**	**F6**	**F7**	**H1**	**H2**	**H3**	**H6**	**H9**
021	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**3**
024	**3**	**2**	**0**	**0**	**0**	**0**	**0**	**1**	**1**	**3**
025	**1**	**1**	**1**	**0**	**0**	**0**	**1**	**2**	**0**	**1**
026	**0**	**0**	**0**	**0**	**0**	**0**	**2**	**1**	**0**	**1**
027	**2**	**1**	**0**	**0**	**0**	**0**	**0**	**2**	**0**	**2**
029	**3**	**0**	**0**	**0**	**0**	**0**	**0**	**3**	**0**	**5**

Figure 7 : les chaînes totalement reconnues par locuteur

- **BAC3S1 : la** première base de données des références isolées en considérant les deux contextes droit et gauche résultant de la première segmentation.
- **BAC3S2** : la deuxième base des références isolées résultant de la deuxième segmentation.

Chaîne à reconnaître	**BAC3S1 / MFCC**
029	**85.56**
024	**83.95**
027	**78.89**
025	**71.26**
026	**56.32**
021	**55.55**
Le tau moyen	**71.92**

Figure 8 : Le taux de reconnaissance avec la base d'apprentissage BAC3S1

- **Deuxième segmentation** :

Chaîne à reconnaître	**BAC3S2 / MFCC**
029	**100**
027	**93.33**
025	**86.26**
024	**85.10**
021	**80.00**
026	**78.16**
Le taux moyen	**87.14**

Figure 9 : Le taux de reconnaissance la base BAC3S2

L'amélioration est significative : désormais le taux de reconnaissance s'élève à 87.14 % de chaînes entièrement reconnues

7. Conclusion :

On a expérimenté l'extension de la PD pour la reconnaissance des chiffres enchaînés (pour des chaînes de 02 et 03 chiffres) tout en essayant de résoudre le problème de la coarticulation entre chiffres par des segmentations successives. Les résultats obtenus étaient satisfaisants.
Si, de par l'extrême complexité du problème de la reconnaissance de la parole et la limitation des capacités des machines actuelles, que le volume du travail qu'on a voulu effectuer et les résultats que nous avons obtenus n'étaient pas toujours à la hauteur de nos espérances. Cependant, ils ont montré des caractéristiques intéressantes qui nous poussent à approfondir nos recherches dans plusieurs directions où il est nécessaire d'avoir une certaine phonétique pour la parole arabe et aussi d'avoir des bases de données standard pour pouvoir valider les résultats des études faites dans ce domaine.

References

Rabiner L.R., RSAMBUR M.(1975), An Algorithms for determining the endpoints of Isolated utterances, *the Bell System Technical Journal*, 1975.
MAKHOUL J.(1 975), linear prediction: *a tutorial review*, proc IEEE vol, pp 580 –561.
DAHMANI H.(2003),la Conception d'un Système de Reconnaissance de Mots Isolés et Enchaînés, *Mémoire de Magister*, Institut d'Electronique, Université d'Annaba.
Rabiner L.R., Rosenberg A.E., Levinson S.E.(1978),Considerations In Dynamic Time Warping Algorithm for Dicrete Word Recognition, *IEEE Trans*, ASSP, vol.ASSP-26, pp.575-582.
Yu Zeng. (2000), dynamic programming and the application in speech recognition system,
Sakoe H., Chiba S. (1978), Dynamic Programming Algorithms for spoken word recognition, IEEE Transactions on Acoustics, *Speech, and Signal Processing*, Vol.Assp-26, N°1.
Myers C.S., Rabiner L.R., Rosenberg A.E. (1980), performance tradeoffs in Dynamic Time Warping Algorithms for Isolated word recognition, IEEE Transactions on Acoustics, *Speech, and Signal Processing*, Vol.Assp-28, N°6.
Bridle J.S., Brown MD., Chamberlain R .M.(1982),An Algorithm for connected word recognition, *Proc 1982 IEEE International Conference on Acoustics*, Speech, and Signal Processing, Paris, France, pp.899-902.
Nakagawa S. (1983), A connected spoken word recognition method by dynamic Programming Pattern Matching Algorithm, *IEEE International conference on acoustics* speech and signal Processing, Boston, pp. 296-299.
Dimartino J. (1984), Contribution à la reconnaissance globale de la parole mots isolés et mots enchaînés, Thèse, Université de Nancy I.
Rabiner L., Juang B.H. (1993), *Fundamentals of speech recognition*, Englewood Cliffs, N.J.: Prentice Hall.