Un système hybride pour l'identification de traits phonétiques complexes de la langue arabe

Sid-Ahmed Selouani[*], *Jean Caelen*[**]

[*]Université de Moncton, 218, Bvd. J-D. Gauthier, E8S 1P6 Shippagan, Canada
Tél.: +(1) 506 336-3625 - Fax: +(1) 506 336-3477 - Mél : selouani@umcs.ca
[**]Laboratoire CLIPS, 385, rue de la Bibliothèque, 38041, Grenoble, France
Tél.: +(33) 4 76514634 - Fax : +(33) 4 76446675 - Mél: jean.caelen@imag.fr

ABSTRACT

This paper presents a new hybrid approach which aims to overcome the drawbacks of automatic speech recognition systems when faced with complex Arabic phonetic features such as emphasis, gemination and relevant vowel lengthening. The approach consists of using hearing/perception-based cues and dividing the global task of recognition into simple and well-defined sub-tasks. The sub-tasks are assigned to a set of Time-Delay Neural Networks using an autoregressive version of the backpropagation algorithm (AR-TDNNs). When they are incorporated in a hybrid structure, ARTDNNs act as post-processors of a HMM-based system. The reported results showed that for either static or dynamic acoustic features, the hybrid system performs significantly better than its corresponding baseline system.

1. INTRODUCTION

Les Systèmes de Reconnaissance Automatique de la Parole (SRAP), avec la diversité des techniques qui les sous-tendent, réagissent inégalement par rapport à la multitude de situations auxquelles ils sont confrontés en milieu réel [3]. Il est intéressant d'observer que dans toute cette panoplie de techniques, il n'existe pas de système adapté à toutes les situations de l'élocution. Il est bien connu par exemple, que le système à base de modèles de Markov cachés (*HMMs: Hidden Markov Models*), très performant quand il s'agit d'opérer un alignement temporel, présente quelques insuffisances sur le plan discriminatoire et de rejet. À l'opposé, les systèmes connexionnistes qui présentent d'excellentes prédispositions à la classification et à la modélisation acoustique, souffrent de leur incapacité à effectuer une normalisation temporelle [1]. Ainsi, la combinaison de systèmes de reconnaissance dans une perspective d'amélioration de performances constitue une approche attrayante, dont l'idée sous-jacente est d'exploiter les avantages de chaque type de système (HMM et réseaux de neurones par exemple). Dans ce papier, notre objectif est de montrer qu'il est possible de rendre efficace une hybridation de systèmes de reconnaissance, si certaines conditions inhérentes à la nature des tâches assignées à chaque système sont respectées. Le principe général de la méthode est expliqué en section 2. La section 3 décrit brièvement les réseaux d'experts connexionnistes récurrents et à délai utilisés dans la structure hybride. En section 4, les résultats obtenus par un système combinant les modèles de Markov cachées et les réseaux d'experts connexionnistes décris en section 3, sont présentés et commentés. En section 5, nous discuterons de l'incorporation de paramètres acoustiques dynamiques et de leur apport dans l'identification des traits phonétiques complexes de la langue arabe.

2. PRINCIPE DE L'APPROCHE HYBRIDE

Le principe général de l'approche hybride proposée consiste à effectuer, par un système de reconnaissance de base (SRA1), une identification phonémique complète mais qui ne produit pas d'étiquetage précis lorsqu'il s'agit de traits phonétiques complexes. Un système (SRA2) agira par la suite comme un module de post-traitement dont le but est d'affiner l'identification effectuée par SRA1, en désambiguïsant les résultats d'étiquetage des traits phonétiques complexes. Si nous prenons l'exemple de l'identification des voyelles, SRA1 doit produire en sortie une même étiquette pour la voyelle longue et sa correspondante brève. Ainsi les voyelles /a/ et /a:/ seront regroupées en une seule classe. SRA1 assignera l'étiquette /VA/ à chaque fois que l'une des deux voyelles est rencontrée. SRA2, spécialement entraîné pour discriminer la voyelle brève de sa correspondante longue, prendra la décision finale quant à l'étiquette définitive à apposer aux deux voyelles (v1 pour /a/ et V1 pour /a:/). Dans le cas de l'emphase, si nous prenons l'exemple de la consonne simple /t/ et de son opposée emphatique /t̲/, il n'est exigé du système SRA1 que de classer les deux consonnes dans une classe unique notée /T/. C'est le système d'appoint SRA2 qui se chargera d'effectuer une catégorisation des éléments appartenant à la classe /T/, en les séparant en deux classes : /t/ et /t̲/. Il en est de même pour les autres oppositions complexes. Une illustration de cette hybridation est donnée en Figure 1.

3. RÉSEAUX AR-TDNN

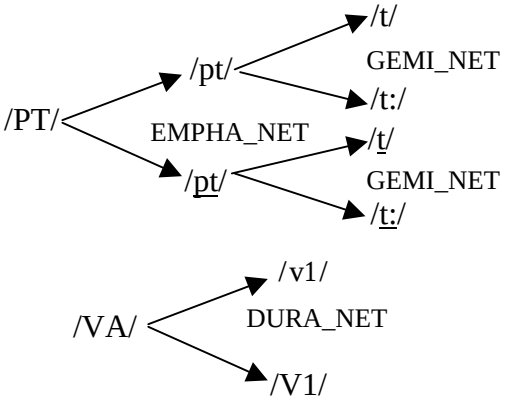
Un système de reconnaissance de la langue arabe doit être capable de distinguer un allongement du à une variation de débit par exemple, de celui pertinent, qui est du à une gémination ou à un allongement de la voyelle.

1ère étape :
Utiliser un modèle d'audition en analyse acoustique : énergies de 24 filtres sont utilisées par SRA1 et 8 indices acoustiques par SRA2.

2ème étape :
Etiqueter en zones stables et en zones de transitions :
/ t a t a / ⟶ sil / sil / sil / #P/ p1/ p1/ PV/ v1/ v1/ v1/ VP/ p1/ v1/ v1/ V#/ sil/ sil/

#P, PV, VP, V#, : transitions pause-plosive, plosive-voyelle, voyelle-plosive, voyelle-pause, respectivement.

3ème étape :
Discriminer sur la séquence produite par l'intermédiaire des experts de SRA2 :
sil / sil / sil / -/PT/ PT/ -/ VA/ VA/ VA/ -/ PT/ VA/ VA/ -/ sil/ sil/



4ème étape :
Filtrer les solutions :
/sil/ p1/ p1/ v1/ v1/ v1/ v1/ p1/ p1/ v1/ v1/ sil/

Figure 1. Principe de l'approche hybride

Cette perception de la durée pertinente doit être préalablement apprise par le système. Celui-ci doit simultanément mémoriser les contextes phonétiques et effectuer une intégration temporelle des fenêtres représentant la séquence à identifier. Russel et Bartley [4] ont montré que l'utilisation de l'algorithme auto-régressif de la rétropropagation de l'erreur (AR backpropagation) permet à un réseau de neurones d'acquérir cette capacité de mémorisation nécessaire dans le cas de l'identification d'événements temporellement instables. Nous pensons que si nous intégrons dans ce type de réseaux récurrents, une composante de décalage, tel que préconisé dans les réseaux à délais (TDNN : *Time Delay Neural Networks*) [7], nous augmentons la capacité du système à percevoir les traits phonétiques complexes même dans un contexte de forte coarticulation. Ainsi l'apprentissage de ce type de réseau ne va pas consister à activer la sortie du réseau lorsque le pattern à identifier se présentera à l'entrée, mais à reconnaître toute la séquence qui contiendra le phonème. Comme le montre la figure 2, la sortie sera activée graduellement dès le passage du contexte droit du pattern à identifier. Pour plus de détails sur les AR-TDNN, consulter [5]. Le système SRA2 est composé de réseaux multi-experts AR-TDNN.

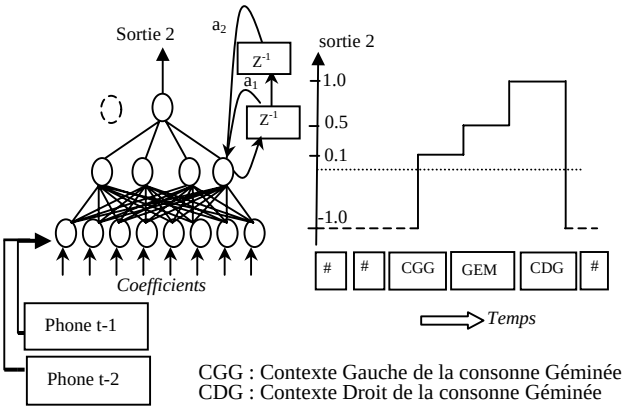


Figure 2. Réseau AR-TDNN

Des sous-tâches de classification sont assignées individuellement à trois experts : GEMI_NET, EMPHA_NET et DURA_NET dont les objectifs consistent respectivement à améliorer les performances de systèmes de reconnaissance de base dans le cas de la détection de la gémination, de l'emphase et de l'allongement pertinent de la durée dans la langue arabe. SRA2 utilise en entrée, huit indices acoustiques dérivés du modèle d'audition de Caelen [2].

4. SYSTÈME HMM/AR-TDNN

À l'heure actuelle, les SRAP basés sur les modèles de Markov cachés sont les plus performants. Ils présentent un avantage déterminant par rapport aux autres systèmes, celui d'utiliser le même formalisme probabiliste pour traiter toute la chaîne de reconnaissance jusqu'au modèle de langage. Les connaissances de tous les niveaux d'analyse : phonétiques, lexicales et syntaxiques peuvent y être intégrées. Ces modèles supposent que le phénomène à caractériser est un processus aléatoire inobservable qui se manifeste par des émissions elles-mêmes aléatoires. Notre système de base utilise des HMMs selon le modèle de Bakis, continu à trois états avec une densité mono-gaussienne par état. Les paramètres acoustiques d'entrée sont constitués par les énergies des 24 canaux d'un modèle auditif et de l'énergie de l'oreille moyenne [2]. Le corpus est constitué d'occurrences VCV de mots et de phrases prononcés par six locuteurs algériens (trois hommes et trois femmes). Globalement, nous retenons 3724 voyelles (dont 1348 longues), 1197 fricatives (182 géminées , 193 emphatiques), 1089 plosives (215 géminées, 273 emphatiques), 573 nasales et 413 liquides. Les semi-voyelles sont assimilées à leurs voyelles correspondantes. Une partie de ce corpus prélevée aléatoirement, n'ayant pas servie à l'apprentissage, est utilisée exclusivement en test. Les tables 1 et 2 donnent les matrices de confusion des HMMs de base et des HMMs de la configuration hybride. Ces derniers opèrent pour les voyelles une identification groupée des brèves et longues. Ainsi au lieu d'avoir six voyelles à classifier, les HMMs du système hybride ne discriminent que trois classes de voyelles : /VA/ qui regroupe les sous-classes /a/ et /a:/, /VU/

regroupe /u/ et /u:/ et enfin /VI/ qui est composée des sous-classes /i/ et /i:/. Les sous-classes sont discriminées par les réseaux AR-TDNN (DURA_NET) dont les scores pour chaque voyelle sont donnés par la table 3. La même démarche est adoptée dans le cas des emphatiques et géminées. Par exemple, la fricative simple /δ/, sa correspondante emphatique /δ̲/ et géminée /δ:/ sont regroupées et étiquetées au sein d'une même classe : /Fδ/. Les experts EMPHA_NET et GEMI_NET se chargeront d'effectuer la séparation emphatique/simple et géminée/simple avec les scores reportés en table 3. Notons que le taux de reconnaissance du système hybride HMM/AR-TDNN est déterminé en effectuant le produit du taux obtenu par les HMMs pour une classe donnée et du taux de discrimination des AR-TDNN dans la même classe. Le taux d'identification de la voyelle longue /a:/ par exemple, est calculé en faisant le produit du taux de reconnaissance par les HMMs de la classe /VA/ (94%) et du taux de discrimination des AR-TDNN pour la sous-classe /a:/ (96%), soit un taux final de 90%. Le regroupement en classes grossières, lorsqu'il s'agit d'oppositions complexes, a un impact positif sur le système HMM car les classes non touchées par ce regroupement sont mieux reconnues. En effet, en comparant les matrices de confusions (tables 1 et 2) nous remarquons une amélioration dans la détection des nasales et des fricatives de l'ordre de 13%. Cette amélioration est de l'ordre de 6% pour les plosives. Dans le cas des liquides, nous remarquons au contraire une baisse des performances de l'ordre de 9%. Globalement, nous pouvons dire que l'hybridation ciblée a un impact positif sur les performances moyennes du système. Ainsi, le fait de déléguer certaines tâches aux réseaux connexionnistes placés en post-traitement à des HMMs opère un nivellement des performances moyennes tout en les améliorant. Avec approximativement 90% de taux moyen d'identification correcte, le système hybride réalise une amélioration de près de 12% par rapport aux HMMs seuls. Cependant, nous relevons une légère défaillance de la configuration hybride dans la reconnaissance des voyelles longues /a:/ et /u:/. Contrairement à ce qui était attendu, les HMMs standards montrent une certaine efficacité à détecter les voyelles longues, mais voient leurs performances rapidement chuter lorsqu'ils s'agit de détecter leurs correspondantes brèves. Un déséquilibre des performances pouvant atteindre 15% a été observé (cas du /a/ et /a:/). Le même phénomène a été observé dans le cas des géminées. Tout ce passe comme si le prolongement du phonème favorise sa modélisation statistique, car le système HMM dispose de suffisamment de temps pour caractériser ce phonème.

5. PARAMÈTRES DYNAMIQUES

Nous avons introduit une information sur la dynamique temporelle du signal en utilisant des coefficients différentiels du premier et du second ordre. Ces coefficients estiment la pente de régression linéaire d'un paramètre à un instant donné.

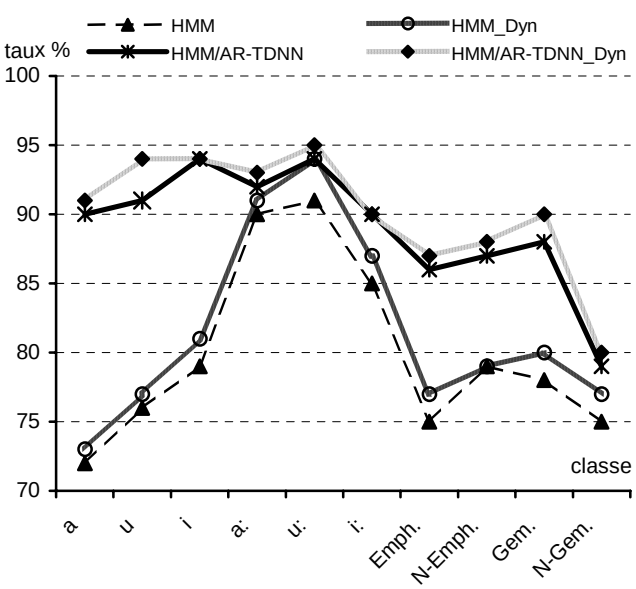


Figure 3. Taux de reconnaissance en % obtenus par le système de base et le système hybride en utilisant les paramètres statiques et dynamiques sur les six voyelles et les consonnes emphatiques, géminées (res. Emph. et Gem.) et leurs opposées (N-Emph. et N-Gem.).

Lorsqu'ils sont inclus dans une analyse acoustique en plus des paramètres initiaux, ils se révèlent efficaces dans l'amélioration des performances des différents systèmes de reconnaissance. Selon le même principe, nous avons inclus dans la configuration hybride HMM/AR-TDNN au niveau du système HMM pour chaque phone : huit indices acoustiques normalisés, huit dérivées premières de ces indices ainsi que leurs huit dérivées secondes. Afin de préserver une dimension raisonnable du vecteur d'analyse, nous avons préféré opérer sur les indices acoustiques plutôt que sur les sorties des 24 canaux du modèle auditif [6]. Comme le montre la figure 3, l'incorporation des paramètres différentiels améliore les performances globales de tous les systèmes.

6. CONCLUSION

La configuration hybride présentée et qui consiste à mettre en appoint à un système de base (HMMs) des experts connexionnistes (AR-TDNN) s'est révélée efficace. Une amélioration significative a été observée, notamment dans la reconnaissance de traits phonétiques complexes de la langue arabe (gémination, emphase et allongement pertinent des voyelles). Une configuration hybride sera d'autant plus efficace si la tâche globale de reconnaissance est divisée en sous-tâches simples et complémentaires qu'on assignerait à des experts adaptés et qui seraient entraînés exclusivement à la réalisation de ces sous-tâches. Les résultats ont également montré que l'introduction de paramètres différentiels auditifs procure une meilleure robustesse des systèmes de reconnaissance. Ce point fait actuellement l'objet d'une étude approfondie dans le contexte de la parole bruitée.

Table 1 Matrice de confusion et taux de reconnaissance du système HMM de base.

	/a:/	/u:/	/i:/	/a/	/u/	/i/	/δ/	/δ/	/δ:/	Nas.	Fric.	Liqu.	Pos.	Taux
/a:/	340	5	2	14	1	2	1	1	0	2	0	1	3	92%
/u:/	3	314	1	1	8	3	0	0	0	2	2	3	2	93%
/i:/	5	3	144	3	1	7	0	0	0	1	3	2	2	84%
/a/	32	26	16	367	15	12	2	3	2	9	2	9	2	73%
/u/	3	16	8	3	164	5	1	2	1	4	3	2	1	77%
/i/	4	2	18	8	10	171	1	0	0	1	1	0	1	78%
/δ/	1	2	2	0	0	1	82	12	10	0	3	1	0	72%
/δ/	3	0	0	1	0	1	5	80	7	2	4	3	1	75%
/δ:/	1	1	0	0	1	1	4	6	89	2	5	2	1	79%
Nas.	6	4	7	2	9	8	0	3	3	184	8	10	8	73%
Fric.	4	2	2	6	2	7	6	7	8	4	245	11	16	77%
Liqu.	2	3	2	4	5	2	0	1	1	2	3	129	3	82%
Plos.	3	4	3	5	1	1	0	3	2	12	9	14	221	80%

Table 2 Matrices de confusion du système HMM dans la configuration hybride .

	/VA/	/VU/	/VI/	/Fδ/	Nas.	Fric.	Liqu.	Pos.	Taux
/VA/	819	10	8	1	10	5	11	5	94%
/VU/	8	513	6	3	7	5	7	3	93%
/VI/	6	7	360	3	4	3	3	2	93%
/Fδ/	3	4	4	296	6	11	3	7	89%
Nas.	4	6	7	3	214	5	6	7	85%
Fric.	3	4	5	6	4	287	5	6	90%
Liqu.	7	8	4	5	10	4	114	5	73%
Plos.	3	2	3	5	7	8	10	240	86%

Table 3 Taux de discrimination des experts AR-TDNN dans la configuration hybride

/VA/		/VU/		/VI/		/Fδ/		/Fδ/	
/a/	/a:/	/u/	/u:/	/i/	/i:/	/δ/	/δ/	/δ/	/δ:/
97%	96%	98%	96%	98%	97%	94%	95%	95%	96%

BIBLIOGRAPHIE

[1] H. Bourlard and N. Morgan. *Connectionist speech recognition a hybrid approach*, Kluwer Academic Publishers, 1994.

[2] J. Caelen. Space/time data-information in A.R.I.A.L Project. *Speech Communications*, volume 4, pages 163-179, 1985.

[3] J.P. Haton. Modèles neuronaux et hybrides en reconnaissance de la parole: état des recherches. *Fondements et perspectives en traitement automatique de la parole*, H. Méloni ed., pages 139-154, 1995.

[4] Russel R. Leighton and C. Bartley. The autoregressive backpropagation algorithm. *International Joint Conference on neural networks*, volume2, pages 369-377, 1991.

[5] S.A. Selouani and J. Caelen. A hybrid Learning Vector Quantization/Time-Delay Neural Networks System for the recognition of Arabic speech. *IEEE-EURASIP workshop*, pages 709-713, Turkey, 1999.

[6] S.A. Selouani and J. Caelen. Recognition of Arabic phonetic features using neural networks and knowledge-based system: a comparative study". *International Journal on Artificial Intelligence Tools*, 8(1), pages 73-103, 1999.

[7] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano and K. Lang. Phoneme recognition using Time-Delay Neural Networks, *IEEE transactions on Audio Speech and Signal Processing*, 37(3), pages 328-339, 1989.