

Reconnaissance automatique des occlusives de l'arabe standard

R. BULOT et A. BÉTARI

G.I.A. Luminy-Marseille, Faculté des Sciences, 163 avenue de Luminy, 13288 Marseille cedex 9, France

abstract : standard arabic is distinctive from other Indo-European languages by the articulation of sounds in the back part of the vocal track, by the feature of gemination and by the complexity of certain consonants from a velarisation. The stop consonants of Arabic do not escape these particularities and form the object of our study within the frame of speech recognition. With the help of a mixed system using Prolog rules and neural networks conjointly, we locate and identify the occlusives of Arabic as well as the nasal consonants in an ascendant phase of Acoustic-Phonetic Decoding.

résumé : l'arabe standard se distingue des langues indo-européennes par l'articulation de sons dans la partie arrière du conduit vocal, par le trait de gémination et par la complication de certaines consonnes d'une vélarisation. Les consonnes interrompues de l'arabe n'échappent pas à ces particularités et font l'objet de notre étude dans le cadre de la reconnaissance de la parole. A l'aide d'un système mixte utilisant conjointement des règles Prolog et des réseaux de neurones, nous localisons et identifions les occlusives de l'arabe ainsi que les consonnes nasales dans une phase ascendante du Décodage Acoustico-Phonétique.

1. INTRODUCTION

Dans cet article, nous nous intéressons à la reconnaissance automatique des consonnes interrompues dont le point d'articulation peut être très arrière dans le conduit vocal et dont la réalisation peut être compliquée d'une emphatisation. Bien que les consonnes nasales (/m/, /n/) ne soient pas à proprement parler des occlusives, nous les traiterons comme telles pour des raisons stratégiques de reconnaissance.

2. DESCRIPTION ACOUSTIQUE

Les consonnes interrompues de l'arabe sont au nombre de 10 (/ʔ/, /k/, /t/, /ʃ/, /q/, /b/, /d/, /d̤/, /m/, /n/) et se distinguent des autres phonèmes par la brièveté du son turbulent intense (explosion) qui suit la phase d'interruption, et par la rapidité des transitions qui mènent à la voyelle suivante ou qui proviennent de la voyelle précédente. Les contraintes phonétiques de la langue font que ces consonnes ne peuvent exister qu'au voisinage immédiat d'une voyelle. Pour une étude plus détaillée nous invitons le lecteur à consulter les travaux [1][2][3][4][5][6].

3. LA RECONNAISSANCE DES OCCLUSIVES

3.1. SONIA, un environnement pour la parole

Notre système de Décodage Acoustico-Phonétique (DAP) a été développé dans un environnement Prolog II adapté au traitement de connaissances numériques et symboliques [7]. La souplesse et la précision des outils disponibles :

- multiple paramétrisation d'une même portion de signal,
- reconnaissance de formes (détection de schémas de colline, de vallée, ...),
- réseaux de neurones,
- définition de contraintes,

permettent de décrire de manière indéterministe des événements acoustiques et phonétiques ainsi que les contextes spécifiques dans lesquels ceux-ci sont pertinents. Ainsi, à partir d'un ensemble de règles Prolog, on localise et on identifie des unités phonétiques qui sont mémorisées dans un treillis de résultats. Un ensemble restreint de règles générales assure la supervision du processus de reconnaissance.

3.2. Stratégie

Nous disposons d'un ensemble de règles (une centaine) permettant de détecter, à l'aide de schémas de formes élémentaires, les segments susceptibles de correspondre à une occlusive. Ces règles sont volontairement peu sélectives (de nombreux segments sont retenus à tort) pour que les occlusives soient localisées dans les contextes les moins favorables. Cette description peu contraignante des sons a aussi pour but de préserver le caractère multi-locuteur de la segmentation. Les caractéristiques plus spécifiques au locuteur (par exemple, la répartition des énergies dans les spectres en fonction des sons) seront modélisées par un réseau de neurones. On dispose ainsi d'un système de reconnaissance adaptatif où les spécificités du locuteur sont apprises de manière automatique.

La stratégie consiste tout d'abord à rechercher des minima locaux d'énergie dans certaines bandes de fréquence qui peuvent révéler la présence d'occlusives [2] [7]. Les segments ainsi détectés servent de point d'ancrage pour rechercher par la suite un événement temporel qui caractérise la phase possible d'explosion. La segmentation obtenue est indéterministe et plusieurs interprétations peuvent être proposées sur une même portion de signal.

Cette pré-sélection des segments à caractère occlusif présente un taux d'oubli inférieur à 3% (quelques /l/ posent problème dans des contextes de voyelles fermées où les dépressions d'énergie sont à peine marquées, principalement avec la voyelle /u/). Lorsqu'elles sont détectées, les occlusives sont assez bien délimitées pour autant que l'on puisse parler de frontière de phonème. Par contre, de nombreuses consonnes vocaliques (/l/, /r/, /w/, /j/, etc.) sont également retenues par les règles et le réseau aura pour charge de les classer non occlusives.

3.3. Identification

Pour chaque segment qu'elles détectent, ces règles proposent un spectre "pertinent" (sur 24 canaux) pour chacune des deux phases probables d'occlusion et d'explosion (figure 1). Afin de prendre en compte le contexte acoustique à droite de la consonne⁽¹⁾ (le plus influent), nous avons retenu un troisième spectre choisi 40 ms. après l'explosion. Après un léger pré-traitement, les données sont fournies à un réseau qui retourne une liste ordonnée de candidats valués (seules sont retenues les occlusives dont le score est supérieur à 0.5). Les exemples d'apprentissage ont été directement fournis par nos règles de segmentation à partir d'un ensemble de phrases. Ainsi, le réseau effectue son apprentissage dans des conditions similaires à celles rencontrées en situation de reconnaissance.

¹ Le corpus d'apprentissage dont nous disposons est trop limité pour envisager une caractérisation en tenant compte simultanément du contexte gauche et du contexte droit (ce qui impliquerait plus d'un millier d'exemples par phonème).

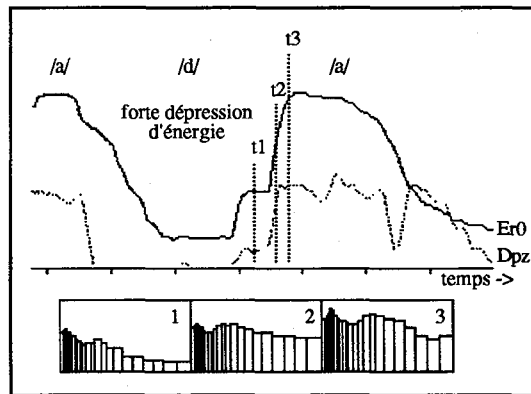


Figure 1 : Les règles de segmentation sélectionnent 3 spectres t_1 , t_2 , t_3 représentatifs respectivement de l'occlusion, de l'explosion et de la transition vocale.

3.4. Le réseau utilisé

L'environnement que nous utilisons a été enrichi avec de nouveaux outils permettant de définir et de gérer des réseaux neuronaux [8]. Il s'agit de réseaux multi-couches sans cycle dont la valeur des nœuds varie dans l'intervalle $[0, 1]$. Les valeurs des connexions sont ajustées par l'algorithme de rétro-propagation du gradient [9] [10]. Ces réseaux sont essentiellement utilisés pour la reconnaissance de formes et sont pilotés par des règles qui sélectionnent dans le signal les données d'entrée et récupèrent les valeurs de sortie.

Pour la classification des occlusives de l'arabe, nous avons retenu une structure avec une seule couche cachée ; les tests que nous avons effectués ont montré que 20 nœuds étaient suffisants pour une bonne classification et permettaient une bonne généralisation des phénomènes lors de l'apprentissage. Nous avons renoncé dans un premier temps à distinguer /m/ et /n/ dans la phase remontante de l'APD et ces deux consonnes sont représentées par la même cellule dans le réseau.

Dans l'architecture utilisée, le caractère occlusif du segment analysé est mis en évidence par la forme du premier spectre et le gain d'énergie entre le premier et le deuxième spectre. Le trait de voisement sera détecté essentiellement sur le premier spectre ainsi que sur le second (apparition ou non du premier formant). Enfin, le lieu d'articulation sera interprété en fonction des concentrations d'énergie dans le deuxième spectre (explosion) et les déplacements de ces concentrations entre le deuxième et le troisième spectre (transition). En particulier, ces déplacements d'énergie permettront de décider du caractère emphatique de la consonne analysée.

L'apprentissage a été effectué sur un corpus de phrases spécifiques où seules les occlusives ont été étiquetées (environ 300 exemples). Cet apprentissage nécessite environ 10 heures de calcul sur un SUN (SPARC 1 à 16 MIPS) pour que les valeurs des connexions entre les nœuds se stabilisent. Le taux de réussite pour la classification des exemples d'apprentissage est de l'ordre de 96%.

4. LES RESULTATS

Les taux de Reconnaissance (localisation + identification) sont de l'ordre de 87 % sur l'ensemble des occlusives. Si /y/ est bien détecté dans tous les contextes ainsi que /d/ en contexte /a/, l'emphatisation de /d/ est difficilement reconnue par notre système lorsque cette consonne est suivie d'une voyelle fermée (/ d i / et / d u /).

Tableau 1 : les résultats obtenus en reconnaissance sur plus de 400 occlusives énoncées en parole continue sont résumés dans la matrice de confusion ci-dessus. La colonne * comptabilise le nombre de fois où le bon phonème apparaît dans les deux premiers candidats. 95% des segments localisés à tort par les règles comme occlusifs sont rejetés par le réseau.

pho.	/ʔ/	/k/	/t/	/t̪/	/q/	/b/	/d/	/d̪/	/m,n	*
/ʔ/	82	0	0	0	7	0	2	0	0	84
/k/	3	68	8	4	12	0	0	0	0	77
/t/	2	0	85	5	0	0	0	0	0	92
/t̪/	10	0	10	80	0	0	0	0	0	90
/q/	0	0	4	5	91	0	0	0	0	95
/b/	0	0	0	0	0	81	6	2	3	90
/d/	0	0	0	0	0	3	87	0	2	90
/d̪/	0	0	0	5	5	0	60	30	0	70
/m,n/	0	0	0	0	0	0	4	0	90	90

5. CONCLUSION

Les résultats obtenus par notre réseau sont assez encourageant surtout lorsqu'on sait qu'une part importante des erreurs est imputable aux règles de segmentation. Par exemple les séquences /ta/ sont souvent regroupées dans un même segment (avec des valuations proches pour /t/ et /a/). Ces situations ne pourraient être récupérées que dans une analyse descendante en mettant en évidence une élévation du F1 au voisinage de /t/.

La classification des consonnes interrompues pourrait être améliorée en intégrant le contexte acoustique gauche dans les données analysées mais ceci nécessite l'utilisation de corpus d'apprentissage beaucoup plus importants si l'on veut disposer de tous les contextes possibles. Parallèlement, certaines erreurs pourraient être récupérées dans une phase d'analyse descendante où nous pourrions exploiter tous les indices que nous avons évoqués, notamment sur le caractère emphatique.

6. REFERENCES

- [1] Ghazali S. : "La coarticulation de l'emphase en arabe", *Arabica*, vol. 28 (1982) ; pp. 251-277.
- [2] Betari A. : "Caractérisation des phonèmes de l'arabe standard en vue d'une reconnaissance automatique de la parole", Thèse de l'université d'Aix-Marseille II (1993).
- [3] Bonnot J.F. : "Recherche expérimentale sur la nature des consonnes emphatiques de l'arabe classique". Travaux de l'Institut de Phonétique de Strasbourg, rap. n° 9 (1977).
- [4] Chiadmi K. : "Contribution à la synthèse par formants de l'arabe : étude de la pharyngalisation". Thèse 3ème cycle, Université de Rabat. Maroc (1986).
- [5] Dellatre P. : "Pharyngeal features in the consonants of Arabic, German, Spanish, French and American English". *Phonetica* (1971) pp. 261-268.
- [6] Klatt D. H., Stevens K.N. : "Pharyngeal consonants". Research laboratory of Electronics L.I.T. Quarterly Progress, Report n° 93, 1969 ; pp. 208-216.
- [7] Bulot R., Nocera P. : "Rule driven Neural Networks for Acoustic-Phonetic Decoding". ICSLP 90, Kobe - Japan, November (1990).
- [8] Nocera P. : "Utilisation de réseaux neuronaux et de connaissances explicites pour le décodage acoustico-phonétique". Thèse de l'université d'Avignon et des pays de Vaucluse (1992).
- [9] Rumelhart D. E., Clelland J. L. Mc. and The PDP Research Group : "Parallel Distributed processing : Explorations in the microstructure of cognition Vol 1 : Foundations" (1986).
- [10] Le Cun Y. : "Modèles connexionnistes de l'apprentissage" Thèse, Paris 1987.