Utilisation des techniques d'intelligence artificielle pour le décodage phonétique de l'Arabe standard

M. Djoudi

CRIN — INRIA Lorraine
Campus Scientifique
B.P. 239
54506 Vandœuvre-lès-Nancy CEDEX
France

Mots clés : Arabe standard, Décodage phonétique, Reconnaissance de la parole, Système expert.

Résumé :

Nous présentons dans cet article une technique d'intelligence artificielle utilisée en reconnaissance automatique de la parole. Elle consiste en la représentation et le traitement des connaissances dans le système expert SAPHA que nous avons développé pour le décodage acoustico-phonétique de l'Arabe standard [6]. Le système SAPHA permet la reconnaissance analytique des phonèmes en parole continue et dans un contexte multilocuteur. Il peut être considéré à la fois comme une étape importante d'un système de dialogue oral homme-machine intégrant d'autres informations linguistiques (lexicales, syntaxiques, sémantiques et pragmatiques) ou bien comme un module d'une machine à dicter [4].

Le système reçoit en entrée le signal de parole préalablement digitalisé et renvoie en résultat un treillis de phonèmes. Le module de reconnaissance proprement dit est réalisé sous forme d'un système expert à base de règles de production. En l'absence d'un expert phonéticien, la connaissance est acquise après une étude phonétique de l'Arabe standard faite sur le corpus DJOUMA constitué de 50 phrases prononcées par 11 locuteurs (7 hommes et 4 femmes). Cette étude nous a permis en outre d'adopter une stratégie de segmentation du signal en grandes classes phonétiques et de déterminer les valeurs caractéristiques des paramètres utilisés en reconnaissance [5]. De même, l'étiquetage manuel des phrases du corpus nous a permis de tester les performances du système.

1 Introduction

La langue arabe a fait l'objet de plusieurs études anciennes [11], [8] ou récentes [1], [10] ayant trait soit à l'aspect phonétique, soit à la composante linguistique. Toutefois, le problème de la reconnaissance automatique n'a été que très peu abordé jusqu'à présent. La reconnaissance de l'Arabe parlé pose un certain nombre de problèmes dues aux caractéristiques phonétiques de la langue. Le système consonantique de l'Arabe comprend 28 consonnes. Sa particularité se fonde sur l'existence des consonnes pharyngales, glottales et emphatiques. Les consonnes pharyngales et glottales se distinguent des autres consonnes par le fait qu'elles possèdent des lieux d'articulation verticaux. Les consonnes emphatiques sont décrites comme ayant un second lieu d'articulation au niveau du pharynx [7], [3]. L'Arabe standard comporte six voyelles. À chaque voyelle brève /a/, /i/ et /u/ s'oppose respectivement une voyelle longue /aa/, /ii/ et /uu/. La durée des voyelles et l'opposition temporelle brève longue sont fondamentales aux niveaux grammatical et sémantique et la durée relative d'une voyelle dépend de son environnement et de la vitesse d'élocution.

2 Le système SAPHA

Le système SAPHA est conçu pour le décodage acoustico-phonétique de l'Arabe. Il utilise les fonctions d'éditions et de traitement de la parole du logiciel Snorri [9]. Il est structuré en modules (voir figure 1), il reçoit en entrée le signal de parole d'une phrase et renvoie comme résultat un treillis phonétique. Autour des modules de reconnaissance, nous avons développé des procédures pour l'analyse phonétique et l'affichage graphique ainsi que des modules d'évaluation des performances du systèmes. L'évaluation nécessite un corpus de phrases équilibrées prononcées par plusieurs locuteurs et étiquetées manuellement :

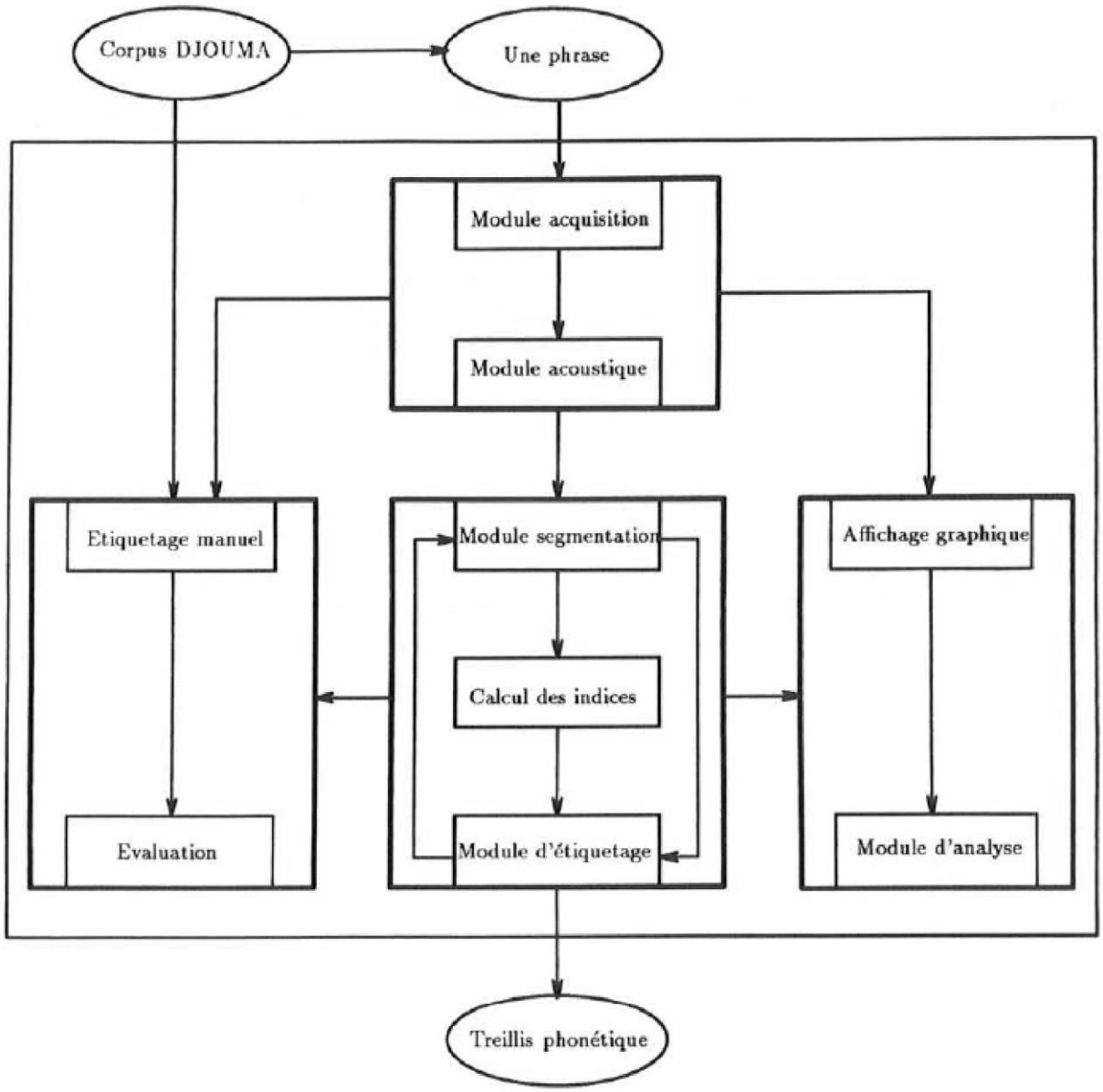

Figure 1 : Architecture de SAPHA

2.1 Le module acquisition

À ce niveau, on fait l'acquisition de la parole directement à partir d'un microphone ou d'une cassette ou bien charger un fichier de parole déjà existant sur disque magnétique. On peut aussi écouter un morceau de parole et jouer sur la valeur de la fréquence d'échantillonnage (par défaut : 16 kHz).

2.2 Le module acoustique

Ce module se charge d'extraire les paramètres acoustiques à partir du signal temporel, il s'agit en particulier de :

- l'amplitude du signal,
- le nombre de passages par zéro en une seconde,

– les pics LPC correspondant aux fréquences de resonnance du conduit vocal,

– les caractéristiques fréquentielles en utilisant un algorithme de transformée de Fourier rapide (FFT),

– la fréquence fondamentale ou pitch..

2.3 Le module de segmentation

Il consiste à segmenter le signal de parole en grandes classes phonétiques en utilisant des algorithmes procéduraux non contextuels et reposant sur des critères simples. Le but essentiel de la segmentation est de réduire l'explosion combinatoire et permettre un cadrage en vue d'un étiquetage automatique. Nous avons retenu trois grandes classes :

– les voyelles { /a/, /i/, /u/, /aa/, /ii/, /uu/ },

– les plosives { /t/, /k/, /?/, /b/, /d/, /q/, /t/ }.

– les fricatives { /z/, /f/,/θ/, /s/, /∫/, /χ/, /ħ/,/z/, /ṣ/ }.

La procédure de segmentation des voyelles utilise les maximas de l'énergie totale et celui de l'énergie dans la bande de fréquences [250-2500 Hz]. L'algorithme de segmentation des plosives utilise l'absence d'énergie au-delà de 600 Hz. Le nombre de passages par zéro et le centre de gravité énergétique sont utilisés pour la segmentation des fricatives. Chacune des trois procédures de segmentation rend une liste des segments détectés. Il faut maintenant traiter les problèmes des inclusions et des intersections. Si deux segments sont inclus l'un dans l'autre, suivant la position des maximas, soit on génère un segment ayant les deux caractéristiques (par exemple voyelle et fricatif pour un /i/) soit on génère deux segments (par exemple plosif puis fricatif pour *z* qui se manifeste comme étant une plosive suivie d'une fricative). Dans cette deuxième hypothèse les limites des segments sont calculées par différence spectrale. Dans le cas de deux segments disjoints, le segment qui existe entre eux sera étiqueté comme "autre" si sa durée est importante, sinon, il sera rattaché à ses voisins. La classe des "autres" comporte donc le vibrant /r/, le latéral /l/, les nasales /m/ et /n/, les semivoyelles /w/ et /y/, ainsi que les fricatives ayant une structure formantique à savoir le /γ/, /h/, /∂/, /∂/ et /ε/.

2.4 Le calcul des indices phonétiques

À chaque indice phonétique est associée une procédure qui le calcule, les valeurs de ces indices phonétiques seront utilisées lors de l'activation d'une règle du module d'étiquetage. Ces indices sont :

– la durée du segment,

– le degré de voisement,

– la position de la barre d'explosion ou burst,

– les caractéristiques de la barre d'explosion,

– les valeurs des formants,

– les transitions formantiques,

– la limite inférieure du bruit et

– le centre de gravité énergétique.

2.5 Le module d'étiquetage

Le module d'étiquetage de SAPHA est un système expert à base de règles de production. En partant des segments fournis par le module de segmentation, le module tente de trouver les bons phonèmes prononcés en utilisant les indices extraits lors de l'étape précédente et les connaissances se trouvant dans la base de connaissances. Le système consiste en une base de connaissances d'identification de phonèmes et un moteur d'inférence. Ce système a été déjà utilisé pour le décodage phonétique du Français dans le cadre du projet APHODEX [2].

2.5.1 La base de connaissances

La base de règles :

Les connaissances acquises sont formalisées sous forme de règles de production.

Une règle se compose de plusieurs parties, pouvant être facultatives :

– un numéro de règle,

– un commentaire en clair,

– une partie contexte gauche (liste de phonèmes),

– une partie contexte droit (liste de phonèmes),

– une partie prémisses,

condition sur le segment en cours d'analyse, du précédent ou du suivant,

– une partie conclusion (une liste de phonèmes pondérés).

Exemple :

```
R223
C Regle pour  /q/ contexte a aa C
CONTEXTE_DROIT [ a aa ]
```

```
SI
burst-present_ACT &
^(burst-freq_ACT 2200 2600)
ALORS [ q 50 ]
```

Le nom des variables est suivi du suffixe ACT (actuel) PRE (précédent) ou SUC (suivant) pour indiquer sur quel segment porte la mesure.

Si deux prémisses sont séparées par la fonction "&" (et) on calcule le minimum de chacun de leurs coefficients de vraisemblance.

Si une prémisse contient des variables qui sont constituées d'une liste de valeurs (par exemple "burst-freq") on va instancier ces variables avec toutes les valeurs possibles de cette liste et cette prémisse se verra affectée d'un coefficient de vraisemblance correspondant au maximum de tous ceux calculés. Les pondérations des résultats vont de -100 (complètement faux) à $+100$ (totalement vrai), le 0 correspond à l'incertitude totale. Actuellement la base contient 187 règles de ce type.

La base de mesures :

À chaque segment correspond une base de mesures qui contient les résultats des procédures de traitement du signal appliquées sur ce segment. Une mesure comprend son nom, sa valeur (booléenne ou liste de valeurs numériques possibles) et le numéro de début et de fin de prélèvement où a été calculée la mesure.

La base de faits :

Au départ, à chaque segment, correspond un nœud du treillis. À chaque nœud est associée une base de faits qui contient la liste des règles déjà appliquées, la liste des contextes droit et gauche supposés à ce stade, les faits qui ont été utilisés et la liste pondérée des phonèmes déduits et la liste des buts à atteindre. À chaque nouveau contexte, on affecte un nouveau nœud.

2.5.2 Le moteur d'inférence

Le moteur du système se charge d'affecter à chaque segment détecté par le module de segmentation une liste d'un ou plusieurs phonèmes. Il est capable de remettre en cause la segmentation à tout moment, de dérouler en parallèle une analyse sur plusieurs segments et de fournir une trace de son raisonnement. Le moteur d'inférence fonctionne en chaînage avant et en chaînage arrière en effectuant une analyse de gauche à droite, segment par segment. L'activation d'une règle est subordonnée à des conditions de comptabilité des contextes gauche et droit, de la conclusion et de la segmentation. Une plausibilité est affectée à chaque phonème hypothétique. La conclusion peut être une action que l'on exécute ou une liste de phonèmes. Pour traiter l'incertitude

et l'imprécision contenues dans les règles, le moteur utilise un raisonnement approximatif basé sur la logique floue et les coefficients de vraisemblance.

3 Évaluation du système

L'évaluation du système consiste à calculer les performances des algorithmes de segmentation et du module d'étiquetage phonétique

3.1 Résultats de la segmentation

Nous avons testé les algorithmes de segmentation sur le corpus DJOUMA, segmenté manuellement. Les résultats que nous indiquons sont calculés par comparaison avec cet étiquetage manuel (voir table 1).

Phon	Nb	Plo	Voy	Fri	Aut	FriVoy	PloFri	Omis	Taux
a	623	1	552	2	14	15	0	39	(91 %)
i	339	3	197	2	5	106	0	26	(89 %)
u	149	1	127	0	4	1	0	16	(85 %)
aa	181	0	163	0	1	9	0	8	(95 %)
ii	58	0	27	2	0	24	0	5	(87 %)
uu	42	1	32	0	2	1	0	6	(78 %)
t	170	165	1	1	1	0	0	2	(97 %)
k	51	49	0	0	0	0	0	2	(96 %)
?	93	66	1	0	23	0	0	3	(70 %)
b	77	73	0	0	3	0	0	1	(94 %)
d	80	75	0	1	4	0	0	0	(93 %)
q	61	60	0	0	1	0	0	0	(98 %)
ṭ	36	33	0	0	1	0	0	2	(91 %)
ḍ	10	2	2	0	5	0	0	1	(50 %)
z	32	2	0	28	1	1	0	0	(96 %)
f	60	4	0	55	0	0	1	0	(93 %)
θ	13	0	0	12	1	0	0	0	(92 %)
s	48	0	0	47	0	0	0	1	(97 %)
$\int$	26	0	0	25	0	1	0	0	(100 %)
χ	10	0	0	10	0	0	0	0	(100 %)
$\hbar$	58	1	0	41	12	1	0	3	(72 %)
∂	8	1	0	0	5	1	0	1	(62 %)
z	12	0	0	12	0	0	0	0	(100 %)
γ	19	2	0	3	12	1	0	1	(63 %)
ε	59	1	2	0	43	2	0	11	(72 %)
h	29	0	2	0	17	0	0	10	(58 %)
ṣ	22	0	0	22	0	0	0	0	(100 %)
$\dot{\partial}$	18	6	0	1	10	0	0	1	(55 %)
m	117	3	5	0	84	0	0	25	(71 %)
n	161	10	2	0	108	1	0	40	(67 %)
l	205	1	4	26	119	2	0	53	(58 %)
r	112	1	4	2	78	1	0	26	(69 %)
w	49	1	1	0	41	0	0	6	(83 %)
j	65	1	1	19	17	2	0	25	(26 %)

Table 1 : Résultat de la segmentation

Les résultats de la segmentation par classe sont résumés dans la table 2

nature	nombre présents	nombre trouvés	nombre insérés
voyelles	1392	1254 (90%)	124 (9%)
plosives	720	672 (93%)	63 (9%)
fricatives	281	256 (91%)	99 (35%)
Autres	852	539 (63%)	193 (21%)

Table 2 : Résultats de la segmentation en grandes classes

3.2 Résultats de la reconnaissance

Le résultat de l'étiquetage phonétique est donné par la matrice de confusion suivante (voir figure 2). L'évaluation a été faite sur le corpus DJOUMA étiquetés manuellement pour 3 locuteurs masculins. Pour chaque segment nous avons gardé les 3 meilleurs étiquettes.

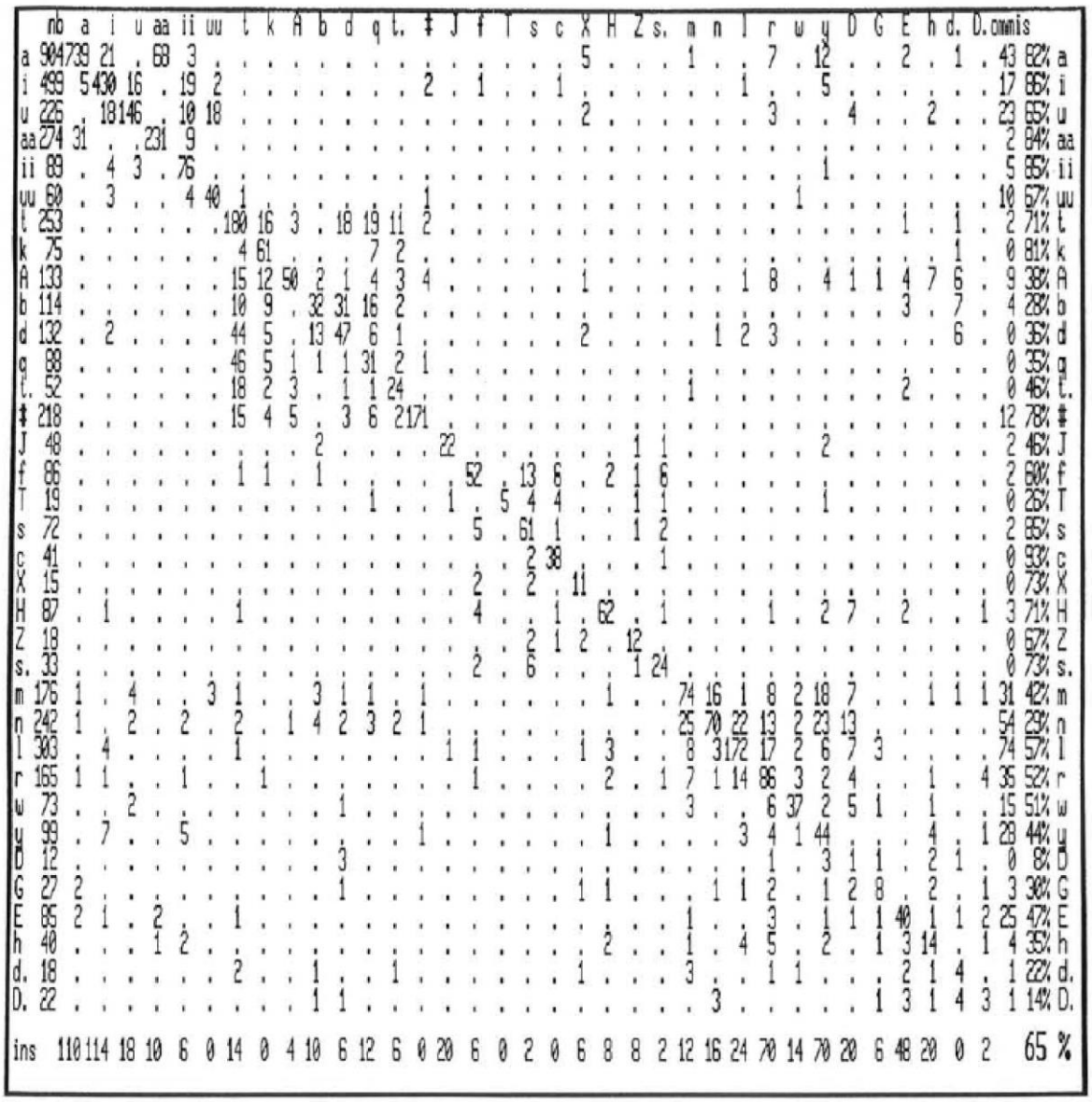

Figure 2 : Matrice de confusion

À l'analyse de la matrice de confusion, nous pouvons remarquer la constitution de blocs autour de la diagonale principale. Les confusions les plus importantes sont entre les phonèmes de la même classe phonétique (voyelle, plosive, fricative ou sonnante). L'analyse des résultats du décodage phonétique de l'Arabe nous a permis d'expliquer certaines erreurs. Les principales causes sont :

- Les erreurs de segmentation, en particulier lorsque deux phonèmes appartenant à une même classe, se suivent dans la phrase, le système rend un seul segment. Ce cas est relativement fréquent dans la classe des sonnantes.
- Les procédures de calcul des indices phonétiques ne fournissent pas toujours les valeurs correctes.
- La base de règles est insuffisante. Il reste un certain nombre de règles à écrire ou à modifier. Une expérience dans le domaine de la lecture des spectrogrammes s'impose pour vérifier la cohérence de la base.

Nous présentons par ailleurs dans la figure 3 un exemple d'une pharse étiquetées automatiquement par le système.

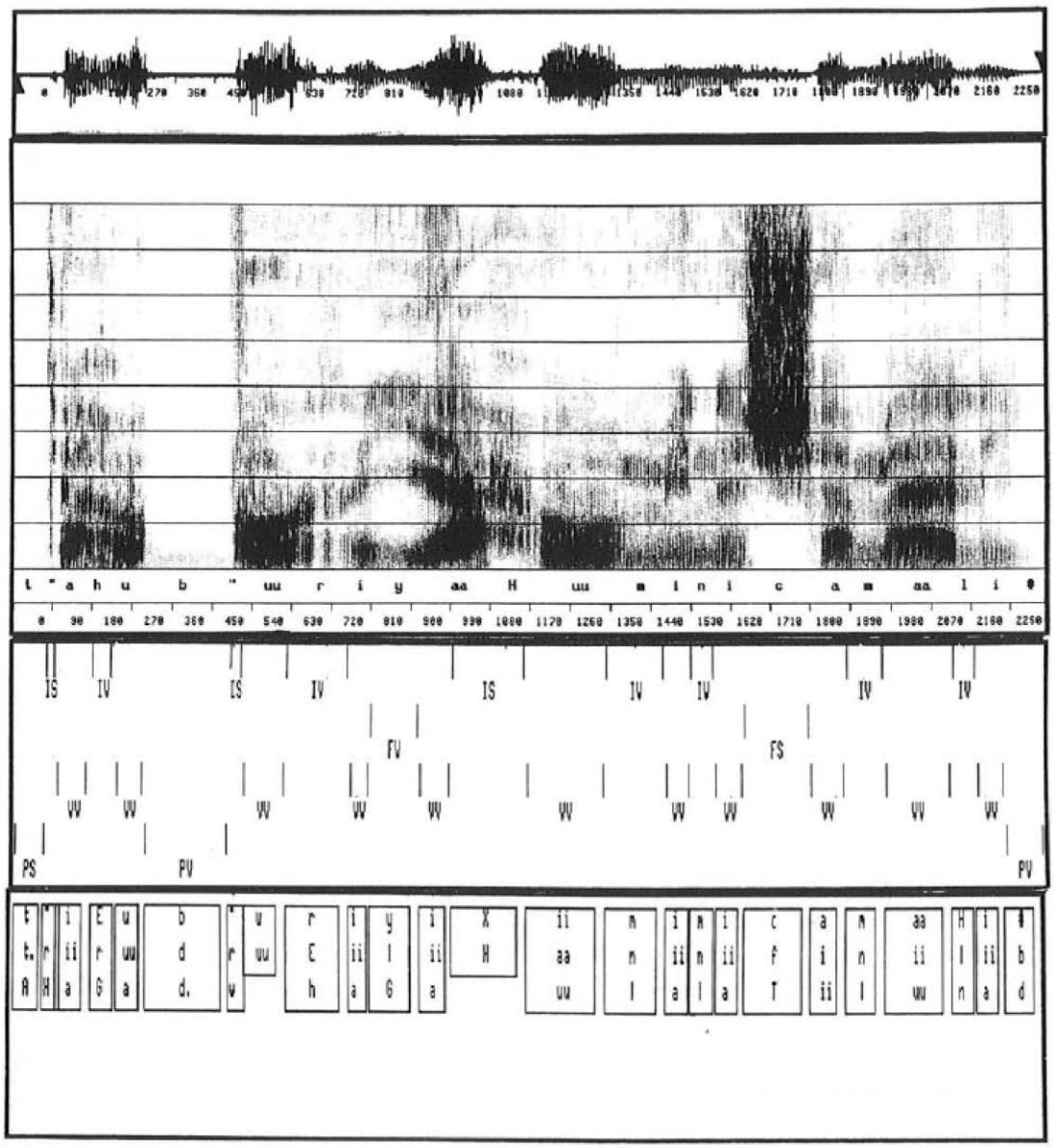

FIGURE 3: Exemple d'étiquetage d'une phrase
traduction de "Les vents soufflent du nord"

4 Conclusion

Nous avons présenté dans cet article une méthode basée sur les techniques d'intelligence artificielle pour le décodage acoustico-phonétique de l'Arabe standard. L'évaluation du système de décodage faite sur un corpus de 50 phrases prononcées par trois locuteurs masculins donne un taux global de reconnaissance de 65%. Les perspectives à donner au présent travail est d'utiliser en plus de l'approche basée sur la connaissance, de nouvelles méthodes de décodage acoustico-phonétique (méthodes connexionnistes et statistiques) afin d'améliorer le pourcentage de reconnaissance et d'intégrer le décodeur phonétique SAPHA dans un système de reconnaissance et/ou de compréhension du langage arabe parlé [4].

Références

[1] J. Cantineau. *Cours de phonétique arabe*. Librairie Klincksieck, 1960.

[2] N. Carbonell, J. P. Haton, D. Fohr, F. Lonchamp, and J. M. Pierrel. APHODEX, design and implementation of an acoustic-phonetic decoding expert system. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1986.

[3] M. Djoudi, H. Aouizerat, and J. P. Haton. Phonetic Study and Recognition of Standard Arabic Emphatic Consonants. In *1990 International Conference on Spoken Language Processing*, Kobe, Japan, 18-22 November 1990.

[4] M. Djoudi, D. Fohr, and J. P. Haton. MARS : Un système de reconnaissance de l'Arabe moderne. In *Actes des* 18ème Journées *d'Études sur la Parole*, pages 217–221, Montréal, Mai, 1990.

[5] M. Djoudi, D. Fohr, and J. P. Haton. Phonetic Study for Automatic Recognition of Arabic. In *Proceedings of European Conference on Speech and Technology*, volume 2, pages 268–271, Paris, September 1989.

[6] M. Djoudi and J. P. Haton. The SAPHA Acoustic Phonetic Decoder System for Standard Arabic. In *1990 International Conference on Spoken Language Processing*, Kobe, Japan, 18-22 November 1990.

[7] A. Giannini and M. Pettorino. *The Emphatic Consonants in Arabic*. Giardini editori e stampatori, 1982.

[8] Ibn Jinni. *Sirr SinaaEat Al IEraab*. Mustapha Al Halabi, 1954.

[9] Y. Laprie. Un système d'étude interactif de la parole. *Actes des* 17ème *Journées d'Etudes sur la Parole*, pages 71–76, Sep 1988.

[10] A. Hadj Salah. Arabic Linguistics and Phonetics. In *Applied Arabic Linguistics and Signal & Information Processing*, pages 3–22. Hemisphere publishing corporation, 1987.

[11] Sibawayh. *EL KITAB, traité de grammaire arabe*. H. Derembourg, 1889.