SAPHA : Un système expert pour le décodage acoustico-phonétique de l'Arabe standard

M. Djoudi D. Fohr J. P. Haton

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 le système expert SAPHA que nous avons développé pour le décodage acoustico-phonétique de l'Arabe standard. Nous décrivons tout d'abord l'architecture générale du système et les différents modules qui le composent, ensuite nous développons les principaux problèmes à résoudre, à savoir :

- La segmentation du signal de parole en grandes classes phonétiques en utilisant des méthodes adaptées du système APHODEX développé dans notre équipe pour le décodage phonétique du français [Foh86] [CFH+84].
- L'extraction automatique des indices pertinents à la reconnaissance phonétique.
- L'identification des phonèmes en utilisant un système expert à base de règles de production.

Les résultats obtenus pour 3 locuteurs seront présentés et discutés.

1 Introduction

Nous décrivons un système de décodage acoustico-phonétique de l'Arabe standard qui 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 (lexique, syntaxe, sémantique et pragmatique) ou bien comme un module d'une machine à dicter commandée par une voie en Arabe standard.

Le système est structuré en modules, il reçoit en entrée le signal de parole préalablement digitalisé et renvoie en résultat un treillis phonétique. 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 [Ani70], [Sal 87], faite sur le corpus DJOUMA constitué de 50 phrases prononcées par 10 locuteurs (7 hommes et 3 femmes) [Djo89]. 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) [Djo89]. De même, l'étiquetage manuel des phrases du corpus nous a permis de tester les performances du système.

2 Architecture du système

Le système SAPHA est composé d'un ensemble de modules et englobe les trois étapes : acoustique, phonétique et phonologique que comporte le niveau inférieur d'un système de reconnaissance automatique de la parole [Gea84]. Ces modules sont :

2.1 Le module acquisition

C'est la première étape dans tout processus de reconnaissance de la parole. A 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. Les fonctions de sauvegarde, de coupure et d'affichage du signal sur une console graphique sont prévus à ce niveau.

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'énergie du signal.

$E_a = \frac{1}{N} \sum_{n=0}^{N-1} |x(n)|$

Ce paramètre peut servir pour distinguer entre parole et non parole et entre sons voisés et sons non voisés, de même, il fournit une information sur l'intonation.

- Le nombre de passage par zéro en une seconde.

Ce paramètre, facilement calculable, est utilisé pour distinguer entre parole et non parole et permet de différencier les sons voisés des sons non voisés.

- Les coefficients LPC qui nous permettent de détecter les pics correspondant aux fréquences de resonnance du conduit vocal.

– Les caractéristiques fréquentielles en utilisant un algorithme de transformée de Fourier rapide (FFT) directement du signal ou bien à partir des coefficients LPC. Le résultat est un spectrogramme numérique.
– La fréquence fondamentale ou pitch, qui correspond aux vibrations des cordes vocales et qui permet de séparer les sons voisés des sons non voisés.

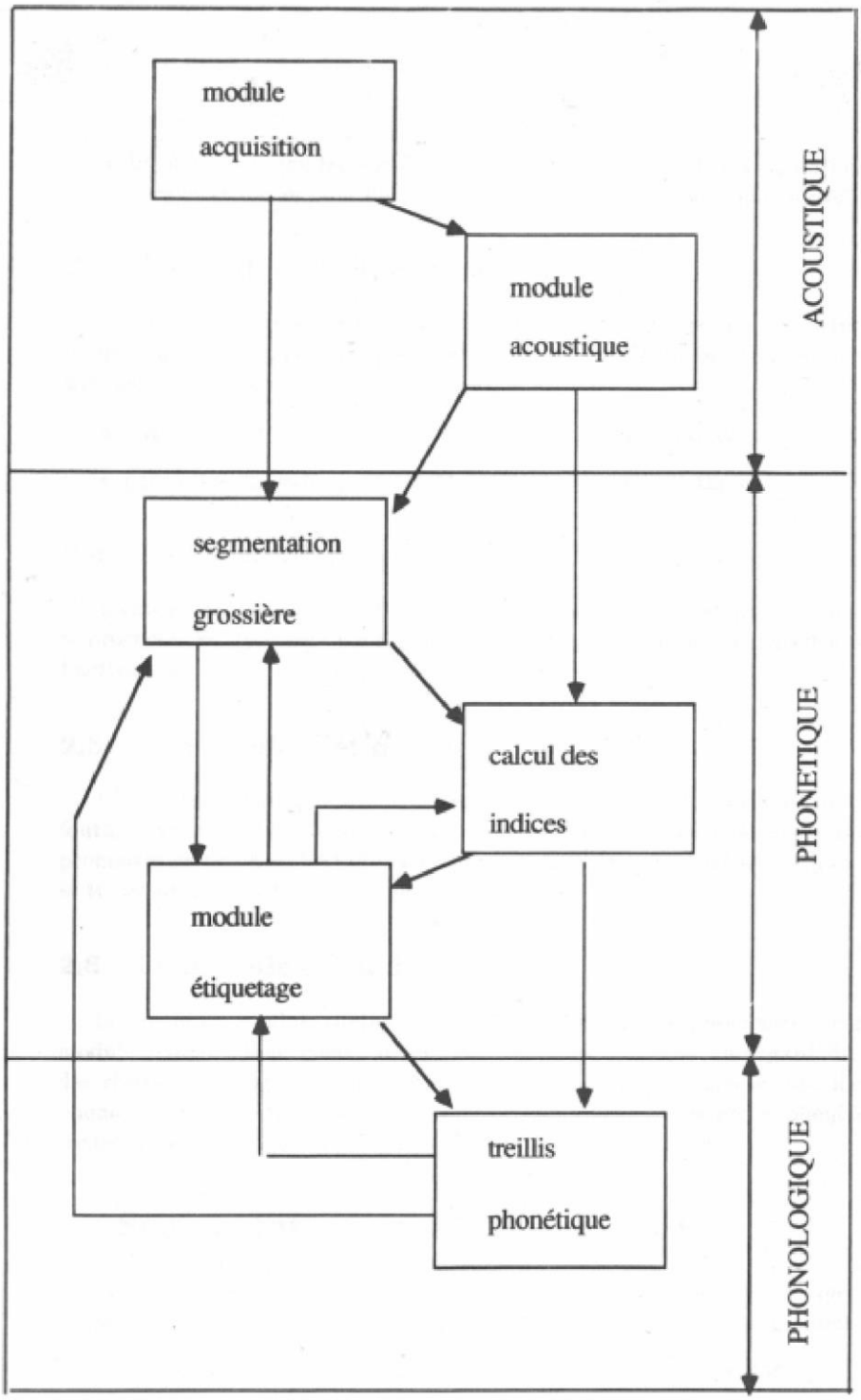

FIG. 1: Architecture de SAPHA

2.3 Le module de segmentation

Il consiste à segmenter le signal de parole en grandes classes phonétiques [CMQ+82] 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 lors de la reconnaissance,
- permettre un cadrage en vue d'un étiquetage automatique.

2.4 Le calcul d'indices

L'extraction des indices phonétiques pertinents est une étape très importante dans le processus de décodage phonétique. Les valeurs de ces indices seront utilisées lors de l'activation d'une règle du module d'étiquetage.

2.5 Le module d'étiquetage

C'est à ce niveau que se fait le décodage proprement dit. 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 règles.

2.6 Le module phonologique

La sortie du module d'étiquetage est un ensemble de phonèmes par segment. Le module phonologique essaye d'éliminer certaines solutions au regard du contexte et des règles phonologiques qui régissent la langue arabe. Dans le système, le module phonologique est intégré dans le module d'étiquetage, et les règles phonologiques sont contenues dans la base de connaissances.

3 Segmentation en grandes classes

Elle est réalisée par un ensemble de trois modules et consiste à segmenter le signal de parole en grandes classes phonétiques. Nous avons retenu trois grandes classes :

- les fricatives { /z/,/f/,/θ/, /s/, /ʃ/,/χ/, /h/,/z/, /s/ } + burst,
- les voyelles { /a/, /i/, /u/, /aa/, /ii/, /uu/ }
- les plosives { /t/, /k/, /?/, /b/, /d/, /q/, /t/ }.

Nous n'avons pas retenu le reste des phonèmes habituellement considérés comme fricatives ainsi que la plosive /d/ car ils présentent, pour tous les locuteurs, des caractéristiques proches de celles des sonnantes.

3.1 Les voyelles

Pour déterminer les voyelles, nous utilisons :

– la courbe de l'énergie dans une bande de fréquence comprise entre 250 et 2500 Hz (zone où sont localisés les deux premiers formants),
– la courbe de l'énergie totale.

La première courbe est obtenue en sommant parmi les canaux correspondant aux fréquences comprises entre 250 et 2500 Hz ceux qui atteignent le seuil de visibilité sur le spectrogramme [CHF+86]. La seconde est obtenue en calculant l'énergie du signal temporel.

Nous recherchons sur ces deux courbes les maximas qui vérifient :

– une intensité au moins égale à la moitié de l'énergie du pic précédent,
– une vallée droite et gauche suffisante,
– la présence de voisement.

Cette procédure permet en outre de calculer une durée vocalique moyenne, qui nous donne une indication sur la vitesse d'élocution.

3.2 Les fricatives

Nous calculons deux courbes :

– une courbe des passages par zéro sur un signal filtré par un filtre passe haut dont la fréquence de coupure est de 800 Hz,
– une courbe du centre de gravité, calculé sur les parties du spectre visibles sur nos spectrogrammes numériques :

$G=(\sum i*S(i) /(\sum S(i))$ pour i de 1 à nombres de canaux

S(i)= intensité en dB du ième canal s'il est visible, 0 sinon.

Une fricative est détectée si on met en évidence un maximum local sur ces deux courbes.

3.3 Les plosives

Nous calculons une courbe d'énergie sur le signal temporel préaccentué et filtré par un passe haut dont la fréquence de coupure a pour valeur 600 Hz. Les plosives correspondent à un minimum local sur cette courbe.

3.4 Traitement des inclusions

Chacune des trois procédures décrites ci-dessus rend une liste des segments détectés (début, fin, position de l'extremum, coefficient de certitude). Il faut maintenant fusionner ces trois listes pour obtenir une première segmentation sous forme de treillis. On commence par traiter le problème des inclusions. 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 fricatif et plosif pour un /f/) soit on génère deux segments (par exemple plosif puis fricatif lorsque la plosive possède un burst fricatif

/ti/). Dans cette deuxième hypothèse les limites des segments sont calculées par différence spectrale.

4 Extraction des indices

Nous avons prévu une procédure pour chaque indice figurant dans la partie prémisse des règles du système.

4.1 La durée du segment

C'est tout simplement la longueur du segment convertie en millisecondes

4.2 Le degré de voisement

C'est le rapport entre le nombre de prélèvements voisés et le nombre total des prélèvements d'un segment. Le voisement d'un segment est déterminé lors du calcul de la fréquence fondamentale.

4.3 La position de la barre d'explosion

Servant à identifier les plosives [DDLO83], [FT82], La barre d'explosion ou burst est une explosion d'énergie de durée brève, ce qui fait que l'algorithme de son d'extraction est très difficile à concevoir. L'idée de base [Djo86] est de détecter le burst dans différentes bandes de fréquence comme étant la position où l'énergie partielle est maximale. La position finale du burst est le numéro de prélèvement qui est détecté le plus de fois sur l'ensemble des bandes de fréquences et qui correspond bien à sa visibilité sur le spectrogramme.

4.4 Les caractéristiques de la barre d'explosion

Si la barre d'explosion existe, il est intéressant de l'analyser et de calculer ses paramètres. L'analyse consiste à dire si elle est compacte ou diffuse, c'est-à-dire concentrée dans une bande de fréquence ou bien continue et répartie sur l'ensemble des bandes. Pour cela nous avons calculé le rapport entre l'énergie maximale et moyenne et projeté sa valeur dans l'intervalle [0,1]. De plus nous avons calculés la position du centre de gravité et les extrémités du pic.

4.5 Le suivi de formants

Les formants des voyelles jouent un rôle très important dans l'identification des voyelles mais aussi dans la discrimination des autres phonèmes à l'intérieur de leurs classes phonétiques [**Erreur ! Source du renvoi introuvable.**].

Pour chaque prélèvement du segment vocalique et à partir des coefficients LPC (ordre 20) nous calculons les premiers pics significatifs qui seront des candidats à des positions de formants.

Ensuite, nous calculons les trois premiers formants comme étant des pics visibles dans les bandes de fréquences [250-850], [750-2300] et [1800-2800] Hz respectivement pour F1, F2 et F3.

4.6 Les transitions formantiques

Pour prendre en compte les phénomènes de coarticulation, on doit étudier les transitions formantiques CV ou VC au frontière entre la voyelle et la consonne adjacente et dire pour chaque formant si la transition est montante, descendante ou plate. Pour se faire, nous prenons un intervalle à la frontière de la voyelle et nous passons les valeurs du formant par une procédure de régression linéaire qui approxime un ensemble de points par une droite en utilisant la méthode des moindres carrées ; le signe du coefficient directeur (la pente) de la droite permet de déterminer la nature de la transition.

4.7 La limite inférieure du bruit

Ce paramètre est très important pour différencier les fricatives, Pour calculer cette limite nous évaluons sur chaque prélèvement du segment le seuil de visibilité inférieur et puis sur l'ensemble du segment nous calculons la moyenne.

5 L'étiquetage phonétique

Le module d'étiquetage de SAPHA est un système expert à base de règles de production. Il consiste en une base de connaissances d'identification de phonèmes sous forme de règles de production et un moteur d'inférence.

5.1 La base de connaissances

5.1.1 La base de règles

Les connaissances acquises sont actuellement formalisées sous forme de règles de production [CHF^{+}86]

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 règle ne pouvant s'appliquer que dans un contexte particulier,

- une partie prémisses, condition sur les mesures du segment en cours d'analyse, du précédent ou du suivant,
- une partie conclusion

soit une action à déclencher pour modifier la segmentation ou le treillis,

soit une liste de phonèmes pondérés.

Exemple : Le nom des variables est suivi du suffixe ACT (actuel) PRE (précédent) ou SUIV (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 deux prémisses sont séparées par la fonction “|” (ou) on calcule le maximum 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.

5.1.2 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.

5.1.3 Base de faits

Au départ, à chaque segment, correspond un nœud du treillis. A 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.

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.

6 Résultats et commentaires

Nous avons testé les algorithmes de segmentation sur le corpus DJOUMA que nous avons segmenté manuellement. Les résultats que nous indiquons sont calculés par comparaison avec cet étiquetage manuel pour 3 locuteurs.

nature	nombre présents	nombre trouvés	nombre insérés
fricatives	200	185 (93%)	8 (4%)
plosives	288	279 (97%)	9 (3%)
voyelles	676	642 (95%)	27 (4%)

Le taux relativement important d'insertion de plosives s'explique par le fait que beaucoup de /f/ sont étiquetés à la fois plosifs et fricatifs et surtout par le fait que le /m/ est étiqueté plusieurs fois comme plosif

Les omissions des voyelles se produisent dans les contextes "Voyelle-Sonnante-Voyelle' dans lesquels les variations d'énergie sont très faibles. Dans ce cas le système ne trouve qu'une seule des deux voyelles ou regroupe celle-ci dans un seul grand noyau.

Les omissions de fricatives sont dues aux phonèmes /f/ et /h/.

Pour la partie reconnaissance, le pourcentage d'étiquetage correct est donné par le tableau suivant :

Classe	pourcentage
Voyelles	94
Fricatives	83
Plosives	80
Inconnues	71

Le pourcentage élevé des voyelles vient sans doute de la simplicité du système vocalique de l'Arabe et l'absence des voyelles nasales. Le taux relativement moyen des plosives et des fricatives s'explique par le fait qu'il manque beaucoup de règles dans la base et que les règles existantes sont peut-être mal formulées. Pour ce qui de la classe des inconnues, la grande diversité dans la nature de ses phonèmes est l'origine essentielle du faible taux de reconnaissance.

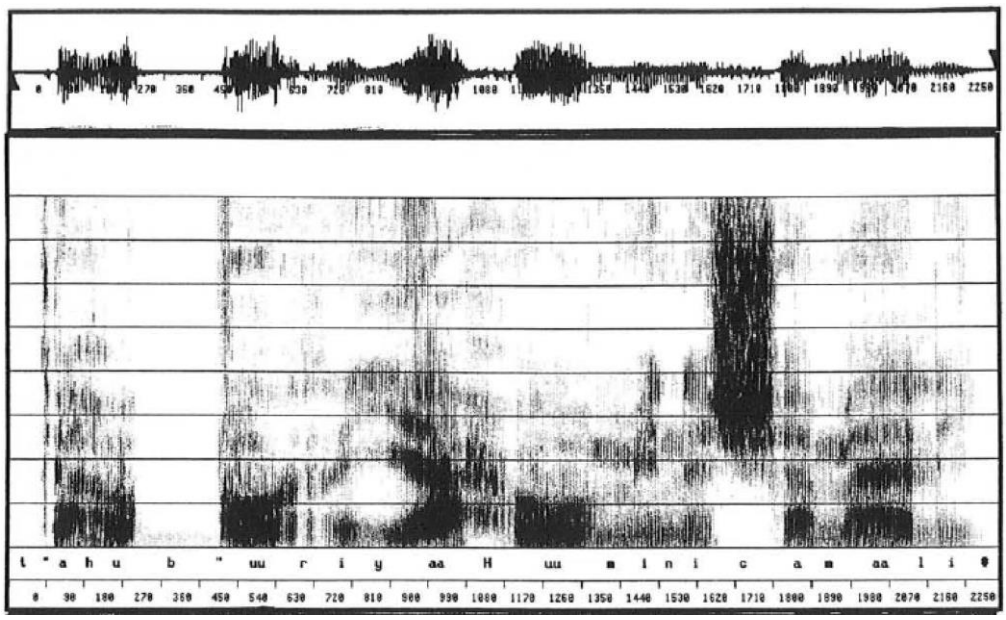

FIG. 2: Signal temporel et spectrogramme
Phrase : traduction de "Les vents soufflent du nord"

7 Conclusion

Nous avons présenté dans cet article un système de décodage acoustico-phonétique de l'Arabe standard basé sur des techniques à base de connaissances. La validation et l'élaboration de nouvelles règles est en cours. Les perspectives à donner au présent travail est de segmenter plus finement la classe des inconnues, d'opter pour une nouvelle forme de représentation des connaissances afin de pouvoir tenir compte réellement du contexte et de travailler en étroite collaboration d'un expert phonéticien pour avoir des connaissances plus sûres et plus cohérentes.

Références

[Ani70] S. H. Al. Ani. *Arabic Phonology. An Acoustical and Physiological Investigation*. Mouton & Co N.V., 1970.

[CFH+84] N. Carbonell, D. Fohr, J. P. Haton, F. Lonchamp, and J. M. Pierrel. *An expert system for the automatic reading of French spectrograms*. In IEEE International Conference on Acoustics, Speech and Signal Processing, San Diego, 1984.

[CHF+86] 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.

[CMQ+82] A. Callec, S. Monne, M. Querre, O. Travarain, and G Mercier. *Automatic segmentation of phonetic units and training in the keal speech recognition system*. In IEEE International Conference on Acoustics, Speech and Signal Processing, pages 2000–2003, Paris, 1982.

[DDLO83] P. Demichelis, R. DeMori, P. Laface, and M. O'Kane. *Computer recognition of plosive sounds using contextual information*. IEEE Trans. Acoust., Speech, Signal Processing, ASSP-31(2), 1983.

[DFH89] 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.

[Djo86] M. Djoudi. Détection et localisation de la barre d'explosion en parole continue et dans un contexte multilocuteur. Rapport de D.E.A, Centre de Recherche en Informatique de Nancy, 1986.

[Djo89] M. Djoudi. *Étude phonétique de l'Arabe standard*. Technical Report 89-R-057, Centre de Recherche en Informatique de Nancy, 1989.

[Foh86] D. Fohr. APHODEX : Un système expert en décodage acoustico-phonétique de la parole continue. Thèse de Doct. Univ. de NANCY 1, 1986.

[FT82] H. Fujisakti and M. Tominga. *Automatic Recognition of Voiced Stop Consonants in CV and VCV Utterances*. In IEEE International Conference on Acoustics, Speech and Signal Processing, pages 1986–1999, 1982.

[Gea84] Gillet and et al. *SERAC : Un système expert en reconnaissance acoustico-phonétique*. Actes du congrès AFCET Reconnaissance des Formes et Intelligence Artificielle, 1984.

[Nor87] K. Norlin. *A Phonetic Study of Emphasis and Vowels in Egytian Arabic*. Lund University Departement of Linguistics, 1987.

[Sal87] A. Hadj Salah. *Arabic Linguistics and Phonetics*. In Applied Arabic Linguistics and Signal & Information Processing, pages 3–22. Hemisphere publishing corporation, 1987.