RECONNAISSANCE AUTOMATIQUE DE LA PAROLE CONTINUE : CAS DE TROIS PHONEMES SPECIFIQUES A LA LANGUE ARABE

Mourad Abbas
Centre de Recherche Scientifique et Technique
pour le Développement de la Langue Arabe

Daoud Berkani
Ecole Nationale Polytechnique

Résumé

La reconnaissance de la parole continue est devenue l'une des préoccupations majeures des chercheurs de cette époque. Les applications diverses y afférentes justifient cet intérêt.

Dans cette étude, nous allons montrer l'importance du choix du vecteur acoustique dans l'efficacité d'un système de reconnaissance des phonèmes arabes dans un flux de parole continue.

L'étude que nous avons élaborée concerne la représentation des signaux de parole, tout en faisant un bon choix des paramètres acoustiques, comme les coefficients cepstraux, l'énergie et le Taux de Passage par Zéro (TPZ). Les différents résultats obtenus en utilisant plusieurs combinaisons de ces paramètres, montrent l'importance du choix d'une meilleure combinaison pour obtenir un taux de reconnaissance amélioré.

Mots clés

Reconnaissance de la parole continue - TPZ - énergie - coefficients cepstraux.

الملخص

أصبح التعرف الآلي على الكلام المتواصل يشكل أحد الاهتمامات الأساسية التي تشغل بال الباحثين في الآونة الأخيرة. وخير دليل على هذا انتشار العديد من التطبيقات المتعلقة بهذا المجال.

سنوضح في هذا المقال أهمية اختيار شعاع المعاملات السمعية للحصول على نتائج جيدة للتعرف الآلي على الحروف العربية في سياق الكلام المتصل.

وتكمن هذه الدراسة في تمثيل إشارات الكلام باختيار صحيح للمعاملات السمعية مثل معدل المرور بالصفر والطاقة والمعاملات الكبسترية. فالنتائج المحصل عليها باستعمال تركيبات مختلفة لهذه المعاملات، توضح جليا أهمية انتقاء التركيبة المثالية القادرة على إعطاء نسب تعرف جيدة.

الكلمات المفاتيح

التعرّف الآلي على الكلام المتواصل - معدل المرور بالصفر - الطاقة - المعاملات الكبسترية.

Abstract

Nowadays, continuous speech recognition becomes one of the researcher's greatest preoccupations. Various applications related to it justify this interest.

In this paper, we will show the importance of an acoustic vector choice to have a good recognition rate of Arabic phonemes in continuous speech.

Our study is about speech signals representation with a good choice of acoustic parameters such as Zero Passage Rate, Energy and Cepstral coefficients. The obtained results, using different combining parameters, show the importance of making the best choice of these parameters in order to reach a better recognition rate.

Keywords

Continuous speech recognition - ZPR - energy - cepstral coefficients.

1. Introduction

L'utilisation de la parole dans un système de communication Homme - Machine présente des avantages dans des situations où la difficulté d'utilisation de la vue ou de la main s'impose, comme elle trouve des applications dans le secteur de communication, où l'on se trouve dans des situations d'accès à distance.

Au cours de ces dernières décennies, des progrès ont été réalisés dans ce domaine. En effet, beaucoup de logiciels destinés à la reconnaissance de la parole continue pour un grand vocabulaire ont été commercialisés dans le but d'approcher les performances humaines. Toutefois, celles-ci sont largement supérieures. Ceci est dû aux facteurs contribuant à la modélisation acoustique et celle de langage. Ce que nous allons appliquer dans ce travail, c'est l'élaboration des vecteurs acoustiques efficaces dans l'amélioration des taux de reconnaissance.

Dans les travaux de l'état de l'art, l'utilisation des coefficients cepstraux et leurs dérivées et même leurs dérivées secondes est courante.

Ceci est justifié par le fait que le signal vocal est décrit par ces paramètres qui sont robustes vis à vis des variations du locuteur et qui permettent de représenter les variations du spectre en fonction du temps [1, 2].

Il a été montré que les coefficients cepstraux sont approximativement considérés comme non corrélés [2], ce qui a pour effet l'absence de la redondance.

Dans cette étude, nous avons enrichi la représentation du signal vocal par d'autres paramètres comme l'énergie et le Taux de Passage par Zéro. Nous allons voir le degré d'influence sur le taux de reconnaissance en faisant simplement la combinaison des différents paramètres.

Nous avons basé notre étude[1] sur la reconnaissance -dans un flux de parole continue- de trois phonèmes arabes en l'occurrence [ḥ], ['] et [h], qui occupent respectivement le sixième, le dix-huitième et le vingt-sixième rang dans l'alphabet arabe.

2. Représentation acoustique

Le signal vocal est une réalisation d'un processus aléatoire non stationnaire. Il est réel, continu et d'énergie finie; sa structure est complexe et variable dans le temps [3].

Il a la caractéristique de transporter plus d'informations que nécessaire, ce qui le rend très résistant au bruit. Le but essentiel du traitement du signal vocal est de donner une représentation moins redondante de ce dernier tout en permettant une extraction précise des paramètres pertinents tels que les formants, le pitch , etc [4].

La grande variété des voix, les variations rapides de la parole, la dualité (source/conduit) de l'appareil vocal, etc., constituent des problèmes essentiels en traitement du signal.

Le signal vocal est quasi stationnaire dans des intervalles de temps qui ne dépassent pas 30 ms, la raison pour laquelle nous effectuons l'analyse sur des trames qui ne dépassent pas cette valeur, en l'occurrence 25.6 ms.

Le fenêtrage est une opération nécessaire si on veut délimiter la durée d'un signal [5]. Les fenêtres spectrales sont caractérisées par la largeur de base du pic

[1] Nous avons fait une étude similaire à celle présentée dans cet article, en utilisant des données et des outils différents [8].

central et le rapport de l'amplitude de ce lobe avec les lobes secondaires [6,7]. Nous avons fait un recouvrement de moitié pour assurer un bon traitement.
La fenêtre que nous avons utilisée est la fenêtre de Hamming. Elle se caractérise par la largeur de base du pic central et la présence limitée des lobes secondaires. Ceci a pour conséquence l'atténuation de l'effet de Gibs.

La figure (1) présente les différences existantes entre la fenêtre triangulaire et celle de Hamming.

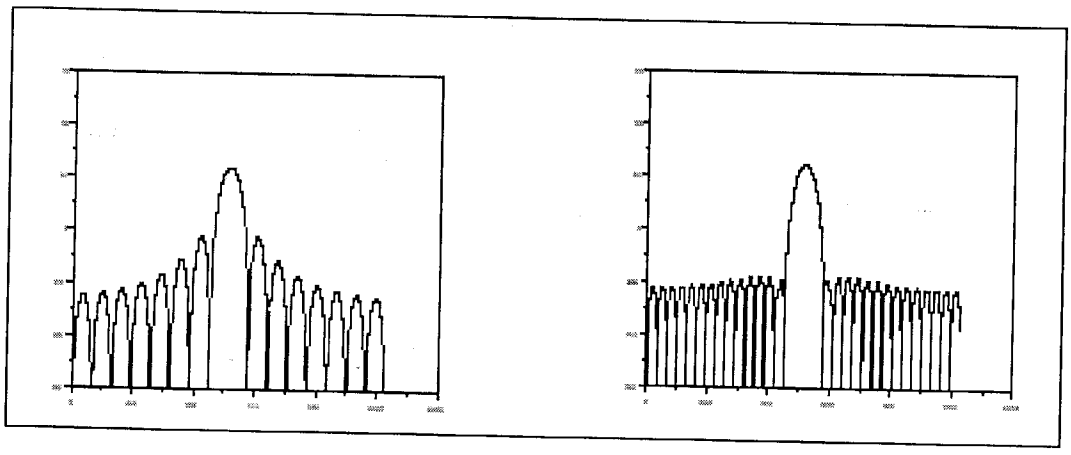

Figure 1 : *Fenêtre triangulaire (à gauche) et fenêtre de Hamming (à droite) [11]*

Dans ce qui suit nous allons citer les différents paramètres acoustiques utilisés dans nos expériences.

2.1. Taux de Passage par Zéro (TPZ)

On peut affirmer qu'il y a un passage par zéro lorsque le signe d'un échantillon est opposé à celui qui le succède. La formule (1) présente le Taux de Passage par Zéro court terme :

$$K[n]=\frac{1}{2}*\sum\left(Sgn*x(n)-Sgn*x(n-1)\right)*W(n-l) \qquad (1)$$

avec

$$W(n)=\begin{cases}\dfrac{1}{n} & \text{si} \quad 0\le n\le N\\ 0 & \text{ailleurs}\end{cases}$$

On peut décider le voisement ou le non voisement d'un son à partir de k(n). Autrement dit k(n) est plus élevé pour les sons non voisés que pour les sons voisés.

2.2. Energie

Le paramètre de l'énergie est utilisé dans la différentiation des sons voisés de ceux non voisés. Il est utilisé aussi dans la détection parole/silence, étape nécessaire et

préliminaire dans un système de reconnaissance automatique de la parole. Ce paramètre peut être obtenu par la formule suivante :

$$Energie = 10G \log\frac{1}{N}\sum x^2(n) \qquad (2)$$

N : nombre d'échantillons
G : gain

2.3. Les coefficients cepstraux

Le signal vocal x(n) est produit par un signal excitateur (la source glottique) qui traverse un système de réponse impulsionnelle b_n (conduit vocal). Le but du cepstre est la déconvolution de a_n et b_n ce qui donnera les coefficients cepstraux approchés :

$$\tilde{c}_n = \tilde{a}_n + \tilde{b}_n \ . \qquad (3)$$

Nous avons calculé les coefficients cepstraux en utilisant les coefficients LPC. L'égalité (4) montre la relation entre les coefficients LPC et les coefficients cepstraux C_q .

$$\ln\left(\frac{1}{A_p(z)}\right) = \sum_{n=1}^{\infty} C_q(n)\, z^{-n} \qquad (4)$$

$$C_q(i) = -a_p(i) - \sum_{n=1}^{i-1}(1-\frac{n}{i})\, a_p(n)\, C(i-n)$$

$avec \quad i \succ 0$

$$C_q(0) = \ln\sigma^2 \qquad pour\ 0 \le i \le p$$

2.4. Les coefficients cepstraux dérivés

Les dérivées des coefficients cepstraux se calculent par rapport à (2k+1) trames de part et d'autres de la trame courante.

$$\Delta C_n(i) = G.\sum_{k=-K}^{+K} K.C_{n-k}(i) \qquad (5)$$

avec G : gain.

2.5. Les dérivées secondes

Ce sont les dérivées secondes des coefficients cepstraux. Ils se calculent par la formule (6) :

$$\Delta(\Delta C_n(i)) = G.\sum_{k=-K}^{+K} K.\Delta C_{n-k}(i) \qquad (6)$$

3. Outil de reconnaissance

La reconnaissance automatique de la parole consiste à déterminer la probabilité d'observer la séquence de vecteurs acoustiques y sachant une suite de mots. Cette probabilité est obtenue en utilisant des modèles acoustiques.

Les modèles qui sont bien connus pour leur efficacité en reconnaissance, et que nous avons utilisé dans notre travail, sont les modèles de Markov Cachés appelés HMM (Hidden Markov Models) [8, 9, 10]. Les étapes de reconnaissance sont schématisées par la figure (2).

Nous utilisons la quantification vectorielle, pour extraire le dictionnaire dont la taille influe sur le taux de reconnaissance obtenu. Ce dictionnaire se constitue de vecteurs qui ont comme composantes les paramètres cités ci-dessus.

La reconnaissance par les HMM nécessite en premier lieu l'apprentissage des modèles réalisé par l'algorithme de Baum-Welch et en second lieu l'identification par l'algorithme de Viterbi. Ces deux algorithmes seront définis dans les prochaines sections.

La phase d'apprentissage est une étape nécessaire dans les systèmes de reconnaissance basés sur les HMM, elle consiste à construire un modèle pour chaque phonème, à partir d'un modèle initial et d'un nombre considérable de vecteurs, calculés à partir d'un grand ensemble de phrases segmentées manuellement au préalable.

La reconnaissance des phonèmes se fait par leur comparaison aux modèles déjà construits qui constituent des références, par le biais d'une technique probabilistique qui est l'algorithme de Viterbi.

3.1. Modèles de Markov Cachés

Les modèles de Markov cachés (Hidden Markov Models : HMM) sont utilisés pour capturer les variabilités du signal vocal. Chaque unité de parole est représentée par un modèle. Dans notre cas nous en avons trois qui correspondent aux trois phonèmes en question.

Un HMM est défini par les éléments suivants [8, 9] :

- Un ensemble de n états $S=\{ S1, S2,...SN \}$.
- Une matrice $A= [a(i,j)]$ de dimension (n,n), où a(i,j) représente la probabilité de transiter de l'état i à l'état j qui s'écrit comme suit :

$$a(i,j) = p(x(t+1) = Sj / x(t) = Si) \quad (7)$$
$$i \le i, j \le n$$

avec x(t) un état au temps t.

- Une matrice $B = [b(j,k)]$ de dimension (n,m) ou b(j,k) est la probabilité d'observer le symbole k quand le processus est à l'état j.

$$b_j(k) = p(y(t) = V_k / q_t \in s_j) \quad (8)$$
$$1 \le k \le M ; 1 \le i, j \le n$$

y(t) étant l'observation au temps t, (y(t) appartient à V)

$V = \{ V_1,...V_m \}$ étant l'ensemble des symboles d'observations.

- La distribution de la probabilité de commencer à l'état i :

$$\pi = [\pi_i] \quad \pi_i = p(q_i = s_i) \text{ avec } 1 \le i \le n$$

- Deux états particuliers dits initial et final I et F qui n'émettent aucun symbole.

3.2. Algorithme de Baum-Welch

C'est un algorithme de construction d'un modèle en utilisant un ensemble d'élocutions du même mot. Il consiste, à partir d'un modèle initial, à recalculer les matrices a(i,j) et b(j,k) pour définir un nouveau modèle où les probabilités d'observation des élocutions utilisées seraient supérieures à ce qu'elles étaient dans le modèle initial.
Soit la probabilité :

$$P_t = p(x(t)=i, x(t+1)=j / Y(1)-Y(t), M) \quad (9)$$

$$P_t(i,j) = \alpha(i,t).a(i,j).b(j,Y(t+1)) \Big/ p(o/\ M) \quad (10)$$

$$\text{avec} \quad p(o/M) = \sum_{i=1}^{N} \alpha(i,t) = \sum \pi_i \beta(i,t) \quad (11)$$

Les paramètres du modèle sont obtenus par ce qui suit :

a) $$\pi_i = Y_t(i) \quad \text{avec } Y_t(i) = \sum_{j=1}^{N} P_t(i,j) \quad (12)$$

b) $$a(i,j) = \sum_{t=1}^{T-1} P_t(i,j) \Big/ \sum_{t=1}^{T} Y_t(i) \quad (13)$$

c) $$b(j,k) = \sum Y_t(j) \Big/ \sum_{t=1}^{T} Y_t(i) \quad (14)$$

3.3. Algorithme de Viterbi

Dans un modèle de Markov une suite d'observations y(1),...,y(t) peut être produite par plusieurs suites d'états. Par conséquent on veut connaître la suite d'états x(1),...x(T) la plus probable qui a engendré cette suite d'observations.

Cela est obtenu en maximisant la probabilité suivante :

p(X(1),..X(T)=x(1),..x(T) , Y(1),..Y(T)=y(1),..y(T) /M) (15)

qui n'est rien d'autres que la variable :

GAMA(t,j)= MAX (GAMA(t-1, i) * a(i, j) * b(j, y(t)) (16)

GAMA(0, 0)=1 , GAMA(0, j)=0 pour j non nul.

Cette formule permet de déterminer, à l'instant t, le meilleur chemin jusqu'à l'état j à partir des meilleurs chemins trouvés au temps t-1. Nous voulons dire, par meilleur chemin, le meilleur alignement entre la suite d'états et la suite d'observations.

F est l'état final alors que GAMA(T,F) est la probabilité de la meilleure suite d'états associée à la suite d'observations donnée.

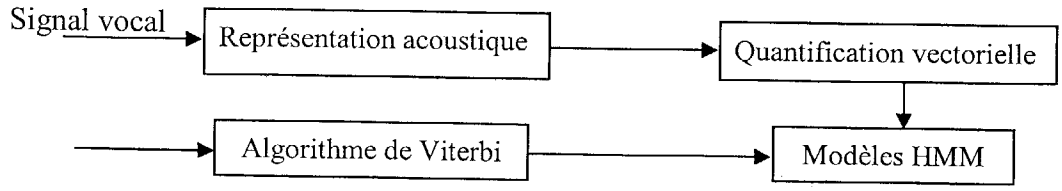

Figure 2 : Les étapes de reconnaissance automatique de la parole

4. Expériences et évaluation des résultats

La reconnaissance automatique de la parole nécessite un corpus sur lequel se fait l'apprentissage. Notre corpus est constitué de cinquante phrases qui contiennent les phonèmes à reconnaître. L'enregistrement est réalisé quatre fois par cinq locuteurs, ce qui donne un corpus de taille 1000 (1000 enregistrements). La fréquence d'échantillonnage utilisée est 10 khz.

Dans cette section nous allons montrer qu'avec un choix judicieux des paramètres d'entrée le taux de reconnaissance est amélioré. Les tables dressées ci-dessous présentent les matrices de confusion pour chacune des expériences effectuées, et cela pour les différentes combinaisons des paramètres.

Phonème reconnu	[ḥ]	[h]	[ʿ]
[ḥ]	285	21	41
[h]	33	109	30
[ʿ]	43	18	163
Taux de reconnaissance	78.94 %	73.65 %	69.66 %

Table 1 : Matrice de confusion pour huit dérivées et huit accélérations

Dans la table (1) nous présentons les résultats issus de l'expérience où nous avons utilisé la combinaison de huit coefficients cepstraux dérivés et huit dérivées secondes des coefficients cepstraux appelées accélérations. Ces résultats ne sont pas satisfaisants. En effet la moyenne des taux de reconnaissance des trois phonèmes ne dépasse pas 74.08 %. Le phonème [ʿ] est le moins reconnu dans cette expérience : 69.66 %.

Phonème reconnu	[ḥ]	[h]	[']
[ḥ]	171	27	26
[h]	28	107	21
[']	22	13	211
Taux de reconnaissance	77.37 %	72.79 %	81.78 %

***Table 2** : Matrice de confusion pour six cepstraux, six dérivées et six accélérations*

La deuxième expérience effectuée en choisissant des vecteurs acoustiques constitués de six coefficients cepstraux, six dérivées cepstrales et six accélérations fait améliorer le taux de reconnaissance du phonème ['] qui est de 81.78 % par rapport à l'expérience précédente où l'on avait un taux de 69.66 %, toutefois ceux des deux autres phonèmes n'augmentent pas voire diminuent. La table (2) présente les résultats relatifs en détail.

Phonème reconnu	[ḥ]	[h]	[']
[ḥ]	185	26	20
[h]	24	101	17
[']	23	21	217
Taux de reconnaissance	79.74 %	68.24 %	85.43 %

***Table 3** : Matrice de confusion pour huit cepstraux, huit dérivées et huit accélérations*

La table (3) montre la matrice de confusion pour une combinaison de huit coefficients cepstraux, huit coefficients cepstraux dérivés et huit accélérations. Nous remarquons toujours une amélioration au niveau du phonème ['] : 85.43 %, avec presque les mêmes résultats que l'expérience précédente par rapport aux phonèmes [h] et [ḥ].
Ces derniers, par contre, sont mieux reconnus en utilisant un vecteur acoustique composé de huit coefficients cepstraux et huit cepstraux dérivés. Ceci est bien clair dans la table (4).

Phonème reconnu	[ḥ]	[h]	[']
[ḥ]	199	28	64
[h]	32	146	26
[']	9	4	200
Taux de reconnaissance	82.57 %	82.02 %	68.96 %

Table 4 : *Matrice de confusion pour huit cepstraux et huit dérivées*

Phonème reconnu	[ḥ]	[h]	[']
[ḥ]	199	28	64
[h]	38	100	10
[']	9	4	200
Taux de reconnaissance	80.89 %	75.75 %	73 %

Table 5 : *Matrice de confusion pour huit cepstraux, huit dérivées et quatre accélérations*

La combinaison précédente a amélioré, certes la reconnaissance de certains phonèmes en l'occurrence [ḥ] et [h] : 82.57 % et 82.05 %, mais a diminué celle du phonème ['] : 68.96 %. Pour faire augmenter les taux de reconnaissance des trois phonèmes, il a fallu essayer d'autres combinaisons pour arriver à des résultats satisfaisants, comme dans le cas du vecteur constitué de huit dérivées cepstrales, huit accélérations et le paramètre énergie, présentés dans la table (6).

Phonème reconnu	[ḥ]	[h]	[']
[ḥ]	184	10	15
[h]	10	140	23
[']	5	9	213
Taux de reconnaissance	92.46 %	88.05 %	84.86 %

Table 6 : *Matrice de confusion pour huit dérivées, huit accélérations et le paramètre Energie*

Une combinaison qui parait plus représentative et plus efficace est celle qui a donné des taux de reconnaissance supérieurs à 91 %, et ce pour les trois phonèmes. Cette représentativité est expliquée par la présence des différents paramètres pertinents qui forment le vecteur acoustique en l'occurrence : huit coefficients cepstraux dérivés, huit accélérations, le Taux de Passage par Zéro et le paramètre énergie. Cette pertinence n'aurait pu être atteinte si nous avions utilisé chaque paramètre à part. Les résultats correspondants à l'utilisation de cette combinaison est montrée dans la table (7).

Phonème reconnu	[ḥ]	[h]	[']
[ḥ]	180	8	10
[h]	6	139	4
[']	8	4	160
taux de reconnaissance	92.78 %	92.05 %	91.95 %

***Table** 7. Matrice de confusion pour huit dérivées, huit accélérations et les deux paramètres Energie et TPZ*

5. Conclusion

Nous remarquons qu'en utilisant les dérivées et les dérivées secondes, les taux de reconnaissance des trois phonèmes varient entre 70 % et 79 %. L'ajout de six paramètres qui sont les dérivées secondes augmente le taux de reconnaissance du phonème ['] pour atteindre 81 %. Et en ajoutant les mêmes paramètres en nombre de huit le taux est amélioré : 85 %.

Le vecteur acoustique constitué de huit cepstraux et huit dérivées augmente le taux de reconnaissance à 82 % pour les deux phonèmes[ḥ]et [h] mais fait retomber celui du phonème ['] à 69 %.

La combinaison de huit dérivées et huit dérivées secondes et le paramètre énergie donnent respectivement 92 %, 88 % et 84 % pour les phonèmes [ḥ], [h] et ['].
L'introduction d'un autre paramètre, qui est le taux de passage par zéro, sur la combinaison précédente, a encore amélioré les résultats pour atteindre les taux respectifs des phonèmes : 92.78 %, 92.05 % et 91.95 %.

En réalisant cette expérience, nous avons pu montrer l'importance du choix du vecteur acoustique dans l'amélioration de la reconnaissance des phonèmes arabes cités ci-dessus.

BIBLIOGRAPHIE

[1] S.D. Furui, "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum", IEEE Trans on Acoustics, Speech and Signal Processing, vol. ASSP-34, n° 1, February 1986.

[2] J.W. Picone, "Signal Modeling Techniques in Speech Recognition", Proceedings of the IEEE, vol 81, n° 9, September 1993.

[3] Calliope, "La parole et son traitement automatique", Editeur J.P.TUBACH, Masson, Paris, 1989.

[4] R. Boite, H. Bourlard, T. Dutoit, J. Hancq et H. Leich, "Traitement de la parole", Presses Polytechniques et Universitaires Romandes, 2000.

[5] A. V. Oppenheim, R. W. Schafer, "Discrete Time Signal Processing", Prentice-Hall, Inc. Englewood cliffs, New Jersey, 1988.

[6] L. R. Rabiner, R. W. Schafer, "Digital Processing of Speech Signals", Prentice-Hall, Inc., Englewood Cliffs, New Jersey 07632, 1978.

[7] L. Rabiner, B. Juang, "Fundamentals of Speech Recognition", Englewood Cliffs NJ: PTR Prentice Hall (Signal Processing Series), c1993, ISBN 0-13-015157-2.

[8] M. Abbas, M. Debyeche, "Reconnaissance Automatique des Phonèmes Arabes dans la Parole Continue", Seventh Magrebian Conference on Computer Sciences, vol.1, pp. 79-87, 6-8 mai 2002, Laboratoire de Recherche en Informatique (LRI), Université de Annaba, Algérie.

[9] J. P. Haton, « Reconnaissance automatique de la parole », Bordas, Paris, 1991.

[10] M. Abbas, M. Debyeche, " Un système de reconnaissance automatique de la parole en Multibandes", Conférence Internationale sur les Systèmes Complexes CISC'04, du 6 au 8 septembre 2004, Département Informatique, Faculté des Sciences de l'Ingénieur, Université de Jijel, Algérie. Http://www.univ-jijel.dz/(2004)

[11] Yves Laprie, "Analyse spectrale de la parole", Loria, France, 16 octobre 2002. Http://parole.loria.fr/(2005)