Classification Automatique des Consonnes Arrières Arabes en vue de la Correction de Substitution Phonémique

Ahcène Abed, Mhania Guerti

الملخّـص

في هذا العمل نقدم نظاما لتصنيف آلي يتعلق بالأصوات العربية المتأخرة في الترتيب المخرجيّ الحديث -من الفم إلى الحلق- : /ق/ /ء/ و /خ/ /ح/. يكمن الهدف الأساس في إنجاز نظام مساعد لتقويم النّطق بالنسبة للأطفال الجزائريين الذين يعانون من اضطراب في الإبدال الحرفيّ الوظيفيّ. هذا النظام عبارة عن تطبيق للتّعرّف الآليّ على الكلام (RAP)، يرتكز على نموذج «ماركوف» المخفي (HMM). بالنسبة للتحليل الصوتي فإنّه يعتمد على المعاملات (MFCC) مع مشتقاتها الأولى والثانية. وقد قمنا في هذا العمل بدراسة موضوعيّة لمعرفة مدى قدرات النّظام المقترح، والذي أثبت نسبا جيدة تصل إلى 90.46 % بالنسبة للصّوت /ق/ و 91.71 % للصّوت /خ/.

الكلمات المفاتيح : الأصوات العربيّة المتأخّرة، الإبدال الحرفيّ، التّعرّف الآلي على الكلام، MFCC، HMM.

Classification Automatique des Consonnes Arrières Arabes en Vue de la Correction de la Substitution Phonémique

Ahcène Abed[1, 2], Mhania Guerti[2]

[1] Centre de Recherche Scientifique et Technique pour le Développement de la Langue Arabe (CRSTDLA), Alger, Algérie.
[2] Laboratoire Signal et Communications, Ecole Nationale Polytechnique, Alger, Algérie.

abedahcene@gmail.com, mhania.guerti@enp.edu.dz

Résumé

Dans cet article, nous présentons un système de classification automatique des consonnes arrières arabes [q][ʔ] et [x][ḥ]. Le but principal est de construire un système d'aide orthophonique pour les enfants algériens soufrant des problèmes de substitution phonémique fonctionnel. Ce système est une application de la Reconnaissance Automatique de la Parole (RAP) basé sur le Modèle de Markov Caché ou HMM (Hidden Markov Model). La paramétrisation des signaux de parole repose sur une représentation cepstrale utilisant les MFCC (Mel Frequency Cepstral Coefficients) avec ses dérivées premières et secondes. Une évaluation objective a été appliquée à notre travail. Cette dernière montre de bonnes performances, avec des Taux de Reconnaissance de 90.46 % pour le son [q] et 91.71 % pour le [x].

Mots clés : les consonnes arrières arabes, la substitution phonémique, RAP, HMM, MFCC.

1. Introduction

Un grand nombre d'enfants algériens rencontrent durant leurs périodes d'apprentissage langagier des difficultés de prononciation. Celles-ci sont montrées comme l'addition, la distorsion, l'omission et la substitution phonémiques [1-2]. Pour cette dernière l'enfant remplace un phonème non encore acquis par un autre phonème très proche dans les zones articulatoires. Les deux sons uvulaires [q] et [x] sont substitués respectivement par la glottale [ʔ] et la pharyngale [ḥ].

Pour la rééducation phonétique (orthophonie) nous remarquons, d'une part, l'absence des cliniques spécialisées dans ce domaine, et d'autre part, les méthodes utilisées sont classiques, lentes et fatigantes pour un enfant. Avec le développement remarquable du matériel informatique, tels que les ordinateurs, les tablettes et les

téléphones portables, il est possible de construire des logiciels sous formes des jeux éducatifs pour ce type de rééducation. Ces logiciels peuvent être utilisés par l'enfant indépendamment des contraintes du temps et du lieu.

La construction de ces applications est basée sur les systèmes de RAP. Celle-ci est l'une des technologies les plus réussies dans le domaine de la Communication Homme-Machine. Grâce à cette technologie, on peut effectuer plusieurs tâches utilisant la communication orale. On désigne habituellement par RAP, tout processus de décision consistant à extraire des informations contenus dans un signal de parole. Le Modèle de Markov Caché ou HMM est largement utilisé dans ces systèmes.

Dans ce travail, nous nous intéressons à manipuler ces applications pour construire un système de classification automatique des phonèmes [q][ɔ] et [x][ḥ]. Le système utilisé est basé sur les HMM, dont la paramétrisation des signaux de paroles repose sur les MFCC.

2. Problème de la substitution phonémique

Les enfants souffrants des troubles du langage oral fonctionnelles montrent durant leurs prononciations les erreurs suivantes :

- l'addition phonémique : l'enfant ajoute un son ou une syllabe au mot (سسمكة [ssamaka] au lieu de سمكة [samaka]) ;
- l'omission phonémique : l'enfant omet un ou plusieurs sons lors de la prononciation d'un mot (تاب [ta:b] au lieu de كتاب [kita:b]) ;
- la distorsion phonémique : elle consiste à prononcer le phonème d'une façon irrégulière ;
- la substitution phonémique : l'enfant remplace un phonème non encore acquis par un autre très proche dans les zones articulatoires. Exemple : (قمر [qamar] est prononcé comme ئمر [ɔamar] et خيمة [xayma] devient حيمة [ḥayma]).

La substitution phonémique est le résultat d'une fausse articulation. Chaque phonème possède un lieu et un mode d'articulation différents (Table 1) :

Sons Arabes	TOP	Mode d'articulation	Lieu d'articulation
ق	[q]	Occlusive	Uvulaire
ء	[ɔ]	Occlusive	Glottale
خ	[x]	Fricative	Vélaire
ح	[ḥ]	Fricative	Pharyngale

Table 1. Modes et lieux d'articulation de sons traités et leurs Transcription Orthographique Phonétique (TOP).

3. Reconnaissance Automatique de Parole

La RAP a pour but d'extraire le message linguistique contenu dans un signal de parole et de le présenter dans une suite de phonèmes ou de mots, indépendamment du dispositif utilisé (type de microphone), l'environnement acoustique (bureau, salle

bruitée ou hôpital, etc.) et le locuteur (genre, âge, etc.). Le système est composé respectivement de processus :

- acoustique : transformant le son articulé en une suite de vecteurs acoustiques ;
- d'analyse du message correspondant selon le critère de maximum de vraisemblance.

3.1. Les paramètres acoustiques utilisés

Les paramètres MFCC sont les plus utilisés dans les systèmes de RAP [3, 7], en exploitant les propriétés du système auditif humain par la transformation de l'échelle linéaire des fréquences en échelle Mel. Cette dernière est codée à partir d'un banc de filtres triangulaires espacés linéairement pour les fréquences inférieures à 1 KHz et utilisant une échelle logarithmique dans le cas contraire. Ces coefficients sont plus discriminants, plus robustes au bruit ambiant et moins corrélés entre eux (Figure 1).

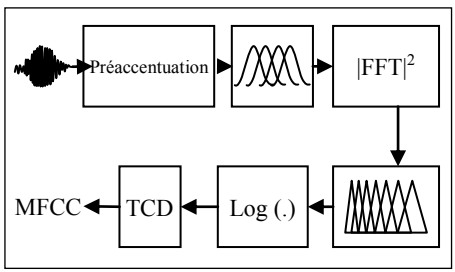

Figure 1: Extraction des paramètres MFCC.

Pour renforcer la contribution des hautes fréquences, le signal de parole est passé par un module de préaccentuation en filtrant ce signal par un filtre passe-haut :

$$x_p(t) = x(t) - a.x(t-1) \qquad (1)$$

Avec : $0.9 \leq a \leq 1$

Le signal accentué est analysé par une fenêtre glissante de Hamming de durée 25 ms avec un recouvrement de 50%. Dans cet intervalle de temps le signal de parole est considéré comme quasi stationnaire :

$$H(n) = 0.54 - 0.46\, cos\frac{2\pi n}{N} \qquad (2)$$

avec $0 \leq n \leq N-1$.

La transformée de Fourier est calculée pour chaque trame pour obtenir le spectre du signal. Ce dernier est multiplié par un banc de filtres triangulaires équidistants dans l'échelle Mel. La localisation des fréquences centrales des filtres est donnée par [4] :

$$f_{mel} = 2595\, log_{10}\left(1 + \frac{f}{700}\right) \qquad (3)$$

Les coefficients cepstraux sont obtenus par une transformation en cosinus décrète à partir du logarithme de l'énergie du banc de filtres. L'expression de ces coefficients est donnée par :

$$C_l(i) = \sqrt{\frac{2}{K}} \sum_{j=1}^{K} S_j\, cos\left[(j - 0.5)\frac{i.\pi}{K}\right] \quad (4)$$

La taille de ces coefficients est augmentée en ajoutant ces dérivées premières Δ et secondes ΔΔ :

$$\Delta C_l(i) = 0.375 \sum_{k=-K}^{K} k\left(\Delta MFCC_{l-k}(i)\right) \qquad (5)$$

$$\Delta\Delta C_l(i) = [\Delta C_{l+1}(i) - \Delta C_{l-1}(i)] \qquad (6)$$

3.2. Modèle de Markov Caché

Les HMM reposent sur l'hypothèse qu'un signal vocal peut être caractérisé par un processus aléatoire paramétrique dont les paramètres peuvent être déterminés avec précision par une méthode bien définie. Un HMM est constitué par des états reliés par des transitions [5, 8]. Généralement ce modèle est donné par :

$$\lambda = \{A, B, \pi\} \quad (7)$$

avec :

- A : est la matrice de transition incluant l'ensemble des états q_i ;
- B : est la matrice d'observation ;
- π : est le modèle initial du HMM.

La probabilité des séquences d'observation connaissant le modèle λ est donnée par :

$$P(X|\lambda) = \sum_S P(X, S|\lambda) \quad (8)$$

$$P(X, S|\lambda) = P(S|\lambda)P(X|S, \lambda) \quad (9)$$

La probabilité de la séquence d'états S est donnée sous la forme :

$$P(S|\lambda) = \pi_{q(1)} a_{q(t-1)q(t)} \quad (10)$$

Donc :

$$P(X|S, \lambda) = \prod_{t=2}^{T} b_{q(t)}(x_t) \quad (11)$$

Le calcul de cette probabilité est très complexe, pour un modèle de N états et une séquence d'observation de durée T cela correspond à $(2T-1)N^T$ multiplications et $N^T - 1$ additions. La solution à ce problème est d'utiliser l'algorithme avant - arrière [6] :

La probabilité avant est donnée par :

$$\alpha_t(i) = P(x_1, \dots, x_t, q(t) = q_i|\lambda) \quad (12)$$

Peut-être calculée récursivement comme suit :

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{q_i q_j}\right] b_{q_j}(x_{t+1}) \quad (13)$$

$$\alpha_1(i) = \pi(i) b_{q_i}(x_1) \quad (14)$$

$$P(X|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \quad (15)$$

La probabilité arrière est donnée par :

$$\beta_t(j) = P(x_{t+1}, \dots, x_T | q(t) = q_j, \lambda) \quad (16)$$

Et récursivement peut-être aussi, calculée par :

$$\beta_t(i) = \sum_{j=1}^{N} a_{q_i q_j} b_{q_j}(x_{t+1}) \beta_{t+1}(j) \quad (17)$$

$$\beta_T(i) = 1 \quad (18)$$

$$P(X|\lambda) = \sum_{i=1}^{N} \alpha_t(i) \beta_t(i) \quad (19)$$

Ce modèle utilise l'algorithme de Viterbi [9], [10] pour la résolution du problème de décodage. Cet algorithme consiste à construire de façon itérative la meilleure séquence d'états possible à partir d'un tableau $T \times N$ contenant les vraisemblances du meilleur chemin $\delta_t(i)$. Les valeurs $\delta_t(i)$ peuvent être calculées par récurrence:

- Initialisation :

  $\delta_0(i) = \pi_i$, probabilité d'être dans l'état i à l'instant initial ;

- Itération:

$\delta_t(i) = \arg\max_j(\delta_{t-1}(i)a_{ij})\, b_i(x_t)$,
dont $b_i(x_t)$ est une distribution gaussienne qui modélise la probabilité de sortie ;

- Terminaison :

$P = \arg\max_i \delta_T(i)$, cette valeur détermine la vraisemblance donnée par la formule précédente.

4. Méthodes et matériels utilisés

Huit enfants des deux sexes, âgés entre 4 et 6 ans, prononcent les mots demandés (Table 2). Ceux-ci portent les phonèmes cibles dans différentes positions possibles : initiale (قلم), médiane (سقوط) et finale (شقيق). Ces signaux sont échantillonnés à 16 kHz, ceci convient pour conserver la quasi-totalité de l'information (théorème d'échantillonnage de Nyquist/Shannon) et sont quantifiés à 16 bits afin d'obtenir une bonne qualité.
Le signal d'entrée, $s[n]$, est transformé en une suite de vecteurs acoustiques (MFCC) :

$$X = \{x_1, x_2, \dots, x_T\} \qquad (20)$$

Son	Initiale	Médiane	Finale
قَ	قلم [qalam]	مقال [maqa:l]	سرق [saraq]
قُ	قنفد [qunfud]	سقوط [su-qu:ṭ]	يتسلق [yatasa-laqu]
قِ	قنديل [qin-di:l]	شقيق [ʃaqi:q]	للفريق [lilfa-ri:qi]
ءَ	أسد [ʔsad]	سؤال [suʔa:l]	ساء [sa:ʔa]
ءُ	أنبوب [ʔunbu:b]	زئير [zaʔi:r]	ينوء [yanu:ʔu]
ءِ	إسلام [ʔsla:m]	مسئول [masʔu:l]	لإجلاء [liʔiʒla:ʔi]
خَ	خليل [xali:l]	فخار [faxa:r]	سلخ [salaxa]
خُ	خضوع [xuḍaʔ]	صخور [suxu:r]	ينسخ [yan-saxu]
خِ	خنجر [xinʒar]	بخيل [baxi:l]	للنسخ [linnas-xi]
حَ	حليب [ḥali:m]	سحاب [saḥa:b]	فرح [fariḥa]
حُ	حروف [ḥuru:f]	لحوم [luḥu:m]	يلمح [yulamiḥu]
حِ	حصان [ḥisa:n]	رحيق [raḥi:q]	للربح [lirribḥi]

Table 2. Mots utilisés pour la construction du corpus de parole

Le bloc de classification décode ces vecteurs dans une représentation symbolique, selon le sens de maximum de vraisemblance, qui pourrait produire l'ordre de ces derniers, en utilisant l'ensemble des modèles de références (HMM). Le modèle reconnu λ est celui qui donne le maximum de vraisemblances parmi les différents modèles (Figure 2).

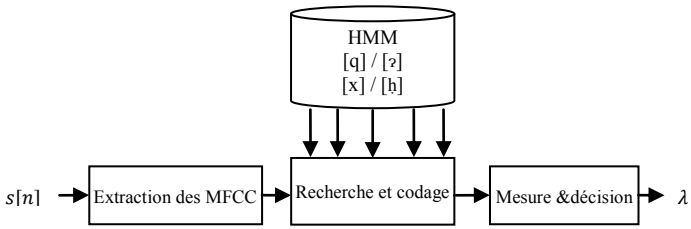

Figure 2: Schéma fonctionnel du système proposé.

5. Résultats et discussion

Dans cette expérience nous avons étudié les performances du système proposé en fonction du nombre d'états, en utilisant un HMM de 3, 4, 5 et 6 états. Pour les paramètres acoustiques nous employons 12 MFCC.

Ce système montre des performances adéquates pour le problème traité (Figure 3). Nous remarquons qu'un HMM de 6 états donne les meilleures performances pour les deux cas. Considérons le phonème [q], le système atteint un taux de classification correcte de 89.58% et de 90.15% pour le phonème [x].

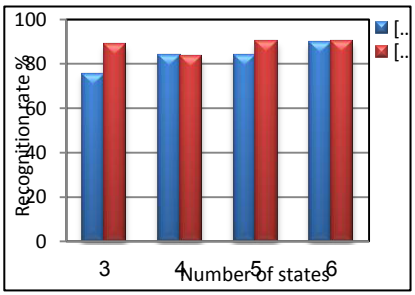

Figure 3 : Taux de classification correct en fonction du nombre d'états avec 12MFCC.

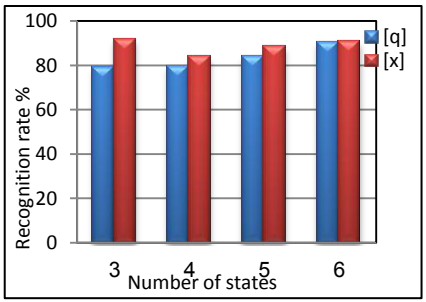

Figure 4 : Taux de classification correcte en fonction du nombre d'états avec 12 MFCC + 3E + 12Δ + 12ΔΔ

Pour voir l'effet des paramètres dynamiques nous avons fait une deuxième expérience utilisant 12 MFCC avec ses dérivées premières Δ et secondes ΔΔ. Les différents états sont : 3, 4, 5 et 6.

Le taux de classification correcte montre une amélioration sur les performances du système (Figure 4). Pour un HMM de 6 états, la classification du premier son [q] a atteint 90.46%. En outre, pour le deuxième son [x], le meilleur résultat de 91.71% est obtenu avec un HMM de 3 états.

Pour confirmer la possibilité d'utiliser ce type de traitement pour le problème de la substitution phonémique, nous avons testé le système proposé avec deux garçons de 5 ans souffrant des problèmes concernant les deux sons traités. Nous avons réalisé une interface graphique sous forme d'un jeu éducatif installé sur un ordinateur, l'enfant utilise un casque pour écouter et répéter les mots demandés, plusieurs fois. Le processus prend une séance de 20 minutes chaque semaine. Nous avons obtenu de bonnes performances après 4 mois de rééducation.

6. Conclusion

Dans ce travail nous avons introduit un système de RAP, basé sur les HMM, pour le problème de la rééducation orthophonique. Le système est orienté vers le trouble de la substitution phonémique pour les enfants algériens. D'après les résultats obtenus nous avons montré la possibilité de concevoir un système d'aide. Ce système résout beaucoup de problèmes liés à l'absence des cliniques spécialisés dans ce domaine en plus du gain du temps et du coût nécessaire pour ce type de traitement. Généralement cette technique de rééducation peut être plus efficace en conjonction avec les méthodes classiques.

7. Références

[1] Abed, A. and Guerti, M., "Application des HMM à la substitution Phonémique dans l'Arabe Parlé", Journées d'Etudes Algéro-Françaises de Doctorants en Signal-Image & Applications, JEAFD'2012, Alger, 18-23, 2012.

[2] Abed, A. and Guerti, M., "Errors Classification of Phonemic Substitution in Arabic Speech", International Congress on Telecommunication and Application'14. Bejaia, Algeria, 23-24 Apr 2014.

[3] Amrouche, A., Debyeche, M., Ahmed, A.T., Rouvaen, T.M. and Yagoub, M.C.E., "An efficient speech recognition system in adverse conditions using the non parametric regression", Engineering Applications of Artificial Intelligence, Elsevier, 23: 85-94, 2010.

[4] Kumari, R.S.S., Nidhyananthan, S.S. and Anand, G., "Fused mel feature sets based text-independent speaker identification using gaussian mixture model", Elsevier Procedia Engineering, editor, International Conference on Communication Technology and System Design, 30: 319-326, 2012.

[5] Zeng, J., Duan, J. and Wu, C., "A new distance measure for hidden markov models", Expert Systems with Applications, Elsevier, 37:1550-1555, 2010.

[6] Bilmes, J.A., "A gentle Tutorial of the EM Algorithm and its applications to Parameter Estimation for Gaussian Mixture and Hidden Markov Models", Technical report, ICSI-TR-97-021, 1998.

[7] Davis, S. P. and Mermelstein, P., "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Transactions on Acoustics, Speech, and Signal Processing, 28: 357-366, 1980.

[8] Hacine-Gharbi, A., "Sélection de paramètres acoustiques pertinents pourla reconnaissance de la parole", Thèse de doctorat, université de Sétif Algérie, 2012.

[9] Haton, J.P., Cerisara, C., Fohr, D., Laprie, Y. and Smaili, K., "Reconnaissance automatique de la parole : du signal à son interprétation", Paris: Dunod, 2006.

[10] Rabiner, L.R. and Juan, B.H., "Fundamentals of speech recognition", Englewood Cliffs, N.J., USA: Prentice Hall, 1993.