

UNE APPROCHE POUR LA REALISATION DE PHRASES EN LANGUE ARABE PHONETIQUEMENT EQUILIBREE

Handoura Abdallah

Laboratoire de l'Informatique des Télécommunications

Ecole Nationale Supérieure des Télécommunications de Bretagne France

Abdallah.Handoura@enig.rnu.tn

Abstract: this paper concern the methodology of style of some Arabic sentences phonetically equilibrated, the realisation of these sentences is based on statistic studies realised by A.H.MOUSSA and M.MRAYATI. These studies concentrated on the etymology and words in Arabic language. A similar study was done by P.COMBESURE, that concern the French language .

these sentences realised are use like a data base for a speech recognition system.

Résumé:

Ces papiers propose une méthodologie de mise en ouvre de phrases arabes phonétiquement équilibrées. La réalisation de ces phrases est basée sur une étude statistique effectuée par A.H.MOUSSA et M.MRAYATI portant sur les racines et les mots de la langue arabe.

les phrases réalisées peuvent être utilisées comme un corpus d'apprentissage pour un système de reconnaissance de la parole

mots clés : phonèmes, diphonèmes, équilibre phonétique, test statistique,

1. Introduction

Le traitement automatique de la parole sur le plan reconnaissance et synthèse fait appel à d'importants corpus de parole. Chaque équipe travaillant dans le domaine peut être amenée à en créer, mais dès qu'il s'agit de comparer des résultats de recherche ou d'évaluer les performances des corpus d'apprentissage pour les systèmes de reconnaissance ou de synthèse de la parole, un corpus commun ou une méthodologie de test et de validation devient nécessaire.

Le taux de reconnaissance d'un système de reconnaissance de la parole dépend de nombreux paramètres dont certains sont difficilement

quantifiables. Parmi ceux celui du corpus d'apprentissage. Lequel doit contenir tous les phonèmes. Ces différents phonèmes doivent être bien repartis sur eux-mêmes, c'est cette raison le corpus doit contenir des phrases et non plus des diphonèmes ou des mots isolés.

Les travaux de A.H.MOUSSA et M.MRAYATI[1] pourtant sur des études statistiques sur les racines et les mots de la langue arabe, nous ont permis d'appliquer le test de Kolmogorov-smirnov pour mettre au point des listes de phrases phonétiquement équilibrées.

2. Caractères phonétiques de la langue Arabe

Pour tout arabophone, existe une identité de langue appelée "fusHa". Cette langue qui est comprise par tous les arabophones, est régie par des règles uniques de syntaxe, de morphologie et de phonologie que l'on retrouve dans les réalisations dialectales.

Pratiquement il est impossible de stocker tous les mots de la langue arabe sous forme de parole numérisée dans le mémoire d'un ordinateur.

La base de la langue "fusHa" est constituée de 28 phonèmes consonantiques et de

L'ordre de 10 phonèmes vocaliques, dont certaines peuvent être brèves ou longues.

si on met l'Arabe en regard avec les autres langues du monde la plus connue on aura (Tab1) :

	arabe	français	anglais	allemand	italien	espagnol
inflexion	+	+	-	+	+	+
composition	-	±	-	+	-	-
apostrophe	-	+	-	-	+	-
Liaison	+	+	-	-	-	-
Dérivation	+	+	+	+	+	+
ordre libre	-	-	-	+	-	+
Agglutination	+	-	-	-	±	+

Tab1 : Langue Arabe comparée avec d'autres langues

Pour ce qui est de l'Arabe, son originalité réside dans la combinaison de quatre caractéristiques : inflexions, liaison, dérivation et agglutination.

3. Equilibre phonétique en langue arabe

L'accès aux différents phonèmes joue un rôle crucial pour des applications comme la reconnaissance de la parole. Si un mot (ou un

phonème) est inconnu dans le corpus d'apprentissage, le système est mis en échec. En effet, l'accès étant tolérant aux fautes ou aux imprécisions de reconnaissance et le système trouvera toujours un mot, plus ou moins proche, pour remplacer celui qui est observé. Pour cela on définit un corpus de parole phonétiquement équilibrée, un ensemble des phrases continent tous les phonèmes d'une langue sous les différentes possibilités de répartition (cvc, ccvc,...). P.C. vu qu'elle demande un nombre des cellules mémoires très grandes.

A titre d'exemple :

Un vocabulaire de 10 chiffres représenté par 25 vecteurs de référence nécessite 2000 cellules mémoires de 8 octets. [2]

Pour résoudre ce problème on a exploité une étude statistique sur les racines et sur

Les mots de la langue arabe faite par A.HAMOUSSA et M.MRAYATI portant sur 12829 phonèmes consonantiques et vocaliques, et avec un test statistique efficace, on cherche si nos phrases ont la même distribution que celle définie par A.H.MOUSSA ou non.

Une étude multidimensionnelle sur le tableau (Tab2) par la méthode d'analyse factorielle des correspondances offre l'avantage de visualiser les résultats d'une distribution statistique en représentant les forces d'attractions des éléments les uns par rapport aux autres.

Il s'agit là de mesurer à l'intérieur d'un espace multidimensionnel, des distances entre des lignes (les consonnes) et des colonnes (les voyelles); et déterminer de manière systématique des axes les plus pertinents.

si X_{ij} est l'élément situé à l'intersection de $i^{\text{ème}}$ ligne et de la $j^{\text{ème}}$ colonne dans la matrice X, la distance (quadratique) d'un point par rapport à l'origine s'exprime par :

$$X_i^2 = \sum_{j=1}^p X_{ij}^2 = \text{dist}^2(X_i / o)$$

$$= \text{proj}^2(X_i / F_1) + \text{dist}^2(X_i / F_1)$$

Le meilleur axe factoriel F_1 s'obtient par minimisation de $\text{dist}(X_i / F_1)$ qui se ramène à appliquer la méthode classique de multiplicateur de

Lagrange pour déterminer le vecteur directeur de F_1 , [2] qui est le vecteur propre relatif au valeur propre max de $X'X$? (fig 1)

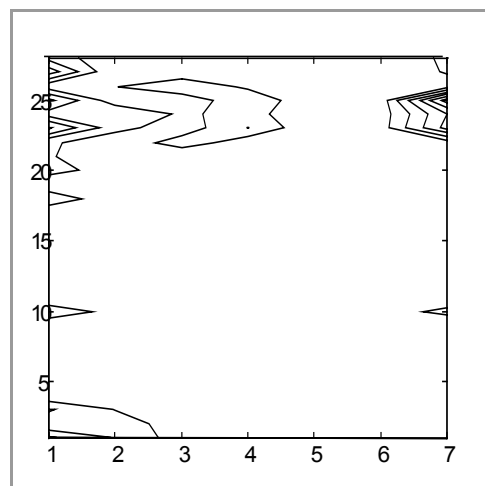


Figure 1 : valeurs propres.

Le traitement par la méthode AFC nous fournie la répartition des phonèmes sur le graphe; cinq zones se démarquent sur le graphe. Chacune d'elles est caractérisée par la présence d'une certaine concentration des traits consonantique.

4.METHODOLOGIE DE TEST

Pour définir une règle de décision concernant la validité d'une hypothèse relative à l'accord global d'une observation empirique (A) faite par A.HA. MOUSSA avec notre distribution théorique à priori ajusté sur des observations (B), on a utilisé le test non paramétrique (test de Kolmogorov-Smirnov)[5] qui consiste à comparer deux observations (A) et (B) et chercher si elles ont la même distribution.

	a	i	u	Aa	ii	uu	pause	total
A	3.141	1.91	.514	.67	.047	.047	.53	6.859
b	1.115	1.66	.374	.218	.218	.218	.974	4.777
t	2.097	.967	.834	.164	.094	.086	.631	4.873
t^	.179	.031	.133	.023	.023	.023	.094	.506
j	.499	.171	.117	.203	.023	.039	.265	1.317
H	.592	.156	.055	.133	.125	.062	.327	1.45
x	.444	.086	.14	.125	.031	00	.203	1.029
d	.46	.304	.203	.148	.171	.304	.616	2.206

D	.327	.171	.062	.398	.522	.086	.437	2.003
r	1.653	.655	.390	.242	.218	.514	1.274	4.919
Z	0.164	.062	.008	.016	.039	.00	.07	.359
s	.787	.398	.195	.304	.023	.14	.865	2.712
c	.208	.062	.07	.109	.008	.023	.265	.818
S	.164	.133	.07	.171	.031	.016	.32	.905
d^	.187	.179	.023	.008	.008	.008	.117	.53
T	.195	.055	.008	.094	.047	.016	.109	.524
D^	.078	.039	.094	.055	.023	.008	.078	.375
E	1.723	.249	.156	.156	.039	.257	.631	3.211
G	.179	.023	.047	.094	.008	.023	.164	.538
f	1.419	.483	.171	.109	.335	.078	.288	2.883
q	1.099	.218	.335	.772	.094	.148	.288	2.954
k	1.208	.203	1.52	.413	.008	.14	.312	3.804
l	4.116	1.356	.429	2.027	.133	.522	4.373	12.956
m	1.59	1.98	.818	1.387	.218	.405	4.007	10.468
n	3.219	.39	.32	1.769	.218	.421	6.462	12.862
h	.335	.967	1.918	.795	.327	.468	.249	5.059
w	3.531	.016	.031	.203	.023	.008	1.13	4.942
y	1.691	.171	.452	.546	.023	.039	1.239	4.161

Tab2 : Redondance des lettres dans les phrases en annexe.

Pour ce qui est des voyelles il est utile de rappeler que l'Arabe oppose aux voyelles brèves[a,u,i] les voyelles longues[aa,uu,ii].Les voyelles brèves peuvent apparaître en syllabe ouverte et fermée, mais les voyelles longues peuvent apparaître qu'on syllabe ouvert sauf pour [aa][4].

Un test statistique paramétrique ou non ne peut être appliqué qu'avec ces conditions d'utilisations, pour celui de notre test, il définit deux conditions:

-(A) et (B) doivent avoir les mêmes limites de classes, dans notre cas c=28

La variable étudiée doit être une grandeur continue : en langue arabe il est impossible de composé un mot par un seul diphonème, C'est pour cela on peut supposé que la langue arabe est continu en fonction des diphones

Soient :

H0: hypothèse nulle relative à une population considérée

α : la probabilité de rejeter H0 alors qu'elle est vraie
1- α : la probabilité d'accepter H0 alors qu'elle est vraie.

S , F: les effectifs cumulés ou les fréquences cumulés:

test bilatéral

H0[(A)=(B)] contre H1[(A)≠(B)]

Si $d \leq d_0$ on accepte H0

Sinon on rejette H0

test unilatéral

Soit le tableau des fréquences observées dans notre corpus :

A	b	t	t^	J	H	X	d	D	r
5.29	4.62	9.01	.676	1.91	3.38	.676	4.96	.676	6.76
Z	s	c	S	d^	T	D^	E	G	f
.789	2.93	1.69	1.35	1.01	2.03	.676	3.6	1.12	3.6

d avec

$$d = \max|FA - FB|$$

avec une limite qui prend la valeur de :

$$d_0 = c \sqrt{\frac{n_1 + n_2}{n_1 n_2}}$$

ou c prend les valeurs suivantes en fonction de α

α	test bilatéral	test unilatéral
5%	1.36	1.22

H0[(A)=(B)] contre H1[(A)>(B)]

cette hypothèse (A) > (B) équivaut a considéré que la distribution (A) est décalée vers la droite par rapport à (B) lorsqu'on envisage les fréquences (ou les effectifs) cumulées pour une valeur x de la variable X des distributions (A) et (B)

si $d \leq d_0$ et $FA - FB > 0$ on accepte H0
sinon on rejette H0

5.DISTRUBITION EXPERIMENTALE

q	k	l	M	n	H	w	y
3.72	3.26	12.5	8.11	5.97	2.14	3.49	3.94

Pour vérifier que la fréquence des phonèmes réalisées dans notre corpus s'approchent de celles données par A.H.MOUSSA nous avons utilisé le test

déjà énoncée et à fin d'augmenter la précision [6] statistique de notre travail on a construit 60 phrases.

éléments de calcul :

SA	6.859	11.636	16.509	17.015	18.332	19.782	20.811	23.017	25.02
SB	5.29	9.91	18.92	19.596	21.512	24.894	25.57	30.53	31.206
100FA	6.859	11.636	16.509	17.015	18.332	19.782	20.811	23.017	25.02
100FB	5.29	9.91	18.92	19.596	21.512	24.894	25.57	30.53	31.206
100(FA-FB)	1.569	1.726	-2.411	-2.581	-3.18	-5.112	-4.759	-7.553	-6.186

SA	29.939	30.298	33.01	33.828	34.733	35.263	35.784	36.162	39.373
SB	37.966	38.755	41.685	43.376	44.728	45.738	47.767	48.443	52.05
100FA	29.939	30.298	33.01	33.828	34.733	35.263	35.784	36.162	39.373
100FB	37.966	38.755	41.685	43.376	44.728	45.738	47.767	48.443	52.05
100(FA-FB)	-7.977	-8.457	-8.675	-9.548	-9.995	-10.475	-11.983	-12.281	-12.677

SA	39.911	42.794	45.748	49.552	62.508	72.976	85.838	90.897	95.839	100
SB	53.177	56.784	60.504	63.773	76.287	84.404	90.379	92.512	96.613	100
100FA	39.911	42.794	45.748	49.552	62.508	72.976	85.838	90.897	95.839	100
100FB	53.177	56.784	60.504	63.773	76.287	84.404	90.379	92.512	96.613	100
100(FA-FB)	-13.266	-13.99	-14.756	-14.221	-13.779	-11.428	-4.521	-1.632	-0.794	0

$$\alpha=0.05 \quad d=\max|FA-FB|=0.14221$$

test bilatéral :

$$H_0[(A)-(B)] = \text{contre } H_1[(A) \neq (B)]$$

$$d=0.14221 < d_0=1.63\sqrt{\frac{200}{1000}}=0.2305$$

on accepte l'hypothèse d'identité des distributions (A) et (B)

test unilatéral :

$$H_0[(A)=(B)] \text{ contre } H_1[(A) > (B)]$$

$$d=0.14221 < 1.52\sqrt{\frac{200}{10000}}=0.2149$$

on accepte l'hypothèse d'identité des distributions (A) et (B)

NB : on peut faire un seul test

6. CONCLUSION

Ce travail nous a permis de mettre au point une base de données phonétiquement équilibrées qui sert actuellement comme un corpus d'apprentissage pour un système de reconnaissance de la parole continue (VINICS)[7],

les enregistrements sont faits au sein de laboratoire de communication parlée de l'IRSIT sous une station SUN équipée d'un logiciel de traitement de la parole waves+

bibliographie:

[1]Boudraa.M, Boudraa.B, Guerin.b:mise en œuvre de phrases arabes phonétiquement équilibrées: 19esJEP-bruxelles,19au22mai 1992. p 451-456

[2]Calliope: la parole et son traitement automatique.1989

[3]AbouHaider.l, Lhote.e, Lefevre.f, Dupret.j.p :les invariants phonétiques en arabe en vue de l'élaboration automatique d'un audiogramme vocal pour tout publics arabophones: 16es JEP 5-9 oct 1987 p270-273.

[4] Ghazali.s,Braham.a:voyelles longues et voyelles brèves en arabe standard, organisation temporelle.: 19es JEP 19-22 mai 1992 p 19-24

[5]Ceresta: aide mémoire pratique des techniques statistiques

[6] Aube.r, Boe.l.j, Lefevre.j.p: lexiques et groupes consonantiques :17es JEP,20-22septembre 1988 p55-60.

SEIRIE A

١. تهب الرياح من الشمال
٢. لقد احتفظوا بعادات افطارهم
٣. غضب الاب على ابنه الصغير
٤. كثرة الكلام تميت القلب
٥. اصبحت الفتاة من ابرز السدات
٦. تحول الموقع الى محطة للقطار
٧. وجدت نفسها في عالم الاقتصاد
٨. قرأت في المجلة خبرا سارا
٩. ان الشاب في حالة جيدة
١٠. كان يود دراسة الطب

SEIRIE B

١١. حاول دائما استغلال الوقت
١٢. تمكن فريقنا من الفوز بالمقابلة
١٣. كنت اسير بجوار قصر عظيم
١٤. تسللت الى عالم الفن فتاجميلة
١٥. لقد ادرك انه غلب على امره
١٦. اجرى الطبيب عملية جراحية
١٧. خطرت لي افكار كثيرة
١٨. غدا سيكون الطقس باردا
١٩. القط تحت المنضدة
٢٠. ابتعدت عن ذكريات شبابي

SEIRIE C

٢١. سقطت الامطار بالمناطق الوسطى
٢٢. تقوم الحياة العائلية على المودة
٢٣. كن متفائلا في تحقيق مشاريعك
٢٤. لقد عاشت في منزل صغير
٢٥. عثر الشرطي على السيارة المسروقة
٢٦. الشمس كانت جد ساخنة
٢٧. حضر الوالد بعد انتهاء الحفل
٢٨. العلم نور والجهل ظلام
٢٩. لماذا ناديتني البارحة
٣٠. رأى الصبي طائرة تحلق في السماء

SEIRIE D

٣٠. علم الصياد بوجود الغزال
٣١. كان يحلم بمكان بعيد
٣٢. تمكن السائق من شق طريقه
٣٣. لدى الحيوانات حاسة غامضة
٣٥. اجرى المعلم امتحانا لتلاميذه
٣٦. كان الجد نائما في الفراش
٣٧. لم اكتب حرفا واحدا في الرواية
٣٨. اتصلت بدائر شؤون الموظفين
٣٩. الاحساس بضيق الوقت يرهق الحياة

SERIE G

٦١. واحد وخمسون الف
٦٢. عشرون دينارا
٦٣. ثلاثة مائة مليون
٦٤. سبعة عشرة درهما
٦٥. ثمان و اربعون
٦٦. مائة جنييه
٦٧. ستة و تسعون ريال
٦٨. مليون واحد
٦٩. ألف درهم
٧٠. تسعة و ثمانون

SERIE H

٧١. واحد وعشرون — ثلاثة عشرة
٧٢. خمسة الاف — احدى عشرة
٧٣. الف وست مائة و اثنان و سبعون
٧٤. اربعة و ثلاثون الفا و اربعة عشرة
٧٥. خمس مائة و ثمان و ستون

SERIE E

٤٠. جثم الصياد في وضعه المختار
٤١. باتت السماء صافية زرقاء
٤٢. احط بالمكان رعب كبير
٤٣. احتضنت المدينة معرضا دوليا
٤٤. اكتشف الحيوان قرب حدوث الخطر
٤٥. كثر الضجيج في شوارع المدينة
٤٦. كان يحفظ حكاية مسلية
٤٧. لا يزال المغامر يجول في العلم
٤٨. ها قد وصل الطبيب الشرعي
٤٩. اعاد الولد السنة الدراسية

SERIE F

٥٠. تتردد في تنفيذ مشاريعك
٥١. اسكت يا نبلي
٥٢. الويل من تمرنا
٥٣. كم بعث ماء اليوم
٥٤. و سأل هل ضغط
٥٥. آذاه ولم يشفق عنه
٥٦. اذا ظننته زحف جنبنا
٥٧. تمنم رمله فوق برتقالك
٥٨. لا تنس اكل النمر
٥٩. صعد الغلام و خردب
٦٠. الويل للأحمد