REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Ecole Nationale Polytechnique

Département d'Electronique

Laboratoire Signal & Communications

Thèse de Doctorat en Electronique

Système d'Aide Orthophonique à la Substitution Phonémique Infantile Basé sur les HMM/GMM

Ahcéne ABED, Magister en Systèmes Electroniques, EMP

Sous la direction de Mdm.
Mhania GUERTI Professeur à l'ENP

Présenté et soutenu publiquement le 30/04/2017

Composition du Jury

Présidente	: Mme Latifa HAMAMI	Professeur	ENP
Directeur de thèse	: Mme Mhania GUERTI	Professeur	ENP
Examinateurs	: Mme Nadjia BENBLIDIA	Professeur	USD Blida 1
	Mme Leila FALEK	Professeur	USTHB
	M. Chérif LARBES	Professeur	ENP
	M. Halim SAYOUD	Professeur	USTHB
Invitée	: Mme Lamia BENMOUSSA	Dr.Orthophoniste	U. d'Alger 2

ENP 2017

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Ecole Nationale Polytechnique

Département d'Electronique

Laboratoire Signal & Communications

Thèse de Doctorat en Electronique

Système d'Aide Orthophonique à la Substitution Phonémique Infantile Basé sur les HMM/GMM

Ahcéne ABED, Magister en Systèmes Electroniques, EMP

Sous la direction de Mdm.
Mhania GUERTI Professeur à l'ENP

Présenté et soutenu publiquement le 30/04/2017

Composition du Jury

Présidente	: Mme Latifa HAMAMI	Professeur	ENP
Directeur de thèse	: Mme Mhania GUERTI	Professeur	ENP
Examinateurs	: Mme Nadjia BENBLIDIA	Professeur	USD Blida 1
	Mme Leila FALEK	Professeur	USTHB
	M. Chérif LARBES	Professeur	ENP
	M. Halim SAYOUD	Professeur	USTHB
Invitée	: Mme Lamia BENMOUSSA	Dr.Orthophoniste	U. d'Alger 2

ENP 2017

à

La mémoire de ma mère

Remerciements

Je voudrais sincèrement exprimer mes plus vifs remerciements à ma Directrice de thèse GUERTI Mhania, Professeur au Département d'Electronique, Ecole Nationale Polytechnique d'Alger pour l'intéressant sujet qu'elle m'a proposé. Sa disponibilité et ses conseils ont été des éléments décisifs pour l'aboutissement de cette thèse. Qu'elle trouve ici ma profonde reconnaissance pour tout ce qu'elle a fait pour moi !

Je tiens à exprimer ma profonde gratitude et mes remerciements les plus vifs à Madame HAMAMI Latifa, Professeur au Département d'Electronique, Ecole Nationale Polytechnique d'Alger, pour avoir fait l'honneur d'accepter de présider le jury de cette soutenance.

Mes remerciements, tout aussi vifs, vont à Mesdames BENBLIDIA Nadjia, et FALEK Leila, Professeurs à l'Université Saâd Dahleb Blida 1 et à l'USTHB.

Je remercie également Mrs LARBES Chérif et Mr SAYOUD Halim, Professeurs au Département d'Electronique, Ecole Nationale Polytechnique d'Alger et à l'Université des Sciences et Technologies Houari Boumediène Alger, d'avoir bien voulu accepter d'examiner et d'évaluer ce travail.

Je remercie Mme BENMOUSSA Lamia, Dr Orthophoniste à l'Université d'Alger 2, pour avoir accepté l'invitation de participer à ma soutenance.

Mes remerciements vont aussi pour tous mes collègues chercheurs et administrateurs du Centre de Recherche Scientifique et Technique pour le Développement de la Langue Arabe CRSTDLA-Bouzaréah Alger.

Je souhaite aussi remercier tous mes amis témoins de mes joies, de mes fatigues, de mes enthousiasmes et de mes hauts et bas, qui m'ont soutenu durant cette période.

ملخص:
يتضمن العمل المقدم نظاما آليا لتصنيف أخطاء الإبدال الصوتي SCAESP لدى الأطفال الناطقين بالعربية، تتراوح أعمارهم من 5 إلى 6 سنوات. يمكن استعمال هذا النظام لتصحيح مجموعة من الأخطاء الناتجة عن التأخر في عملية اكتساب الكلام. مما يسمح للآباء بتوفير الوقت والمال اللازمين لهذا النوع من التأهيل.

لتحقيق الأهداف المرجوة ؛ أجرينا مجموعة من التسجيلات الصوتية لعشرين (20) طفلا في قسم تحضيري، على نصوص تحتوي كل الأصوات العربية الممكنة. 15 منهم تمكنوا من نطق كل الأصوات بشكل سليم أما الخمسة الباقين فكل واحد أبدل صوتا من الأصوات : [s]، [z]، [r]، [ʤ] و [k] بالأصوات : [θ]، [ð]، [ʁ]، [ʃ] و [t] على الترتيب.

يعتمد النظام المقترح على النموذج HMM/GMM الذي يتطلب تحليلا صوتيا لإشارات الكلام وذلك باستعمال المعاملات MFCC. ولإثبات فعالية هذا النظام، درسنا تجريبيا قدراته بدلالة عدد الحالات للنموذج HMM وعدد مكونات النموذج GMM حيث أنجزنا مدونة صوتية لحالات الإبدال الصوتي الخمس بمساعدة 50 طفلا من 5 إلى 6 سنوات بثلاث مدارس ابتدائية جزائرية. باعتبار النتائج المحصل عليها، فإن SCAESP أعطى قدرات محفزة فيما يخص تقابلات الصوامت : [θ]/[s] ؛ [ð]/[z] ؛ [ʁ]/[r] ؛ [ʃ]/[ʤ] و [t]/[k] والتي هي على التوالي : 84.15% ؛ 80.27% ؛ 85.57% ؛ 90.10% و 81.46%.

كلمات دالة : الكشف الآلي للإبدال الصوتي ، اللغة العربية ، أخطاء الإبدال الصوتي ، HMM/GMM، MFCC.

Abstract :
In our work, we present an Automatic Classification System of Phonemic Substitution Errors (SCAESP) among Arabic-speaking children aged between 5 and 6 years. This system can be used for correcting a set of errors related to the speech delays. It allows parents to save the time and the money required for this type of rehabilitation.

To achieve the desired goals, we have recorded a set of texts by 20 children in a preschool class. The texts contain all the possible sounds of the Arabic language. 15 children have pronounced all sounds correctly. However, for the others, each of them substitute one of the sounds [s], [z], [r], [ʤ] and [k] respectively by [θ], [ð], [ʁ], [ʃ] and [t].

Our system is based on HMM/GMM, which requires an acoustic analysis of the speech signals using MFCC. To evaluate SCAESP, we have studied its performances versus the number of states of HMM and GMM components. A speech corpus was built by 50 children of 5-6 years in three primary schools in Algeria. The SCAESP gives satisfactory performances for the consonant oppositions : [s]/[θ] ; [z]/[ð] ; [r]/[ʁ] ; [ʤ]/[ʃ] and [k]/[t], which are respectively : 84.15% ; 80.27% ; 85.57% ; 90.10% and 81.46%.

Key words: Automatic Classification of Phonemic Substitution, Arabic Language, Phonemic Substitution Errors, HMM/GMM, MFCC.

Résumé :
Dans le cadre de notre travail, nous présentons un Système de Classification Automatique des Erreurs de Substitution Phonémique (SCAESP) chez des enfants arabophones âgés entre 5 et 6 ans. Ce système peut être utilisé pour rééduquer un ensemble d'erreurs du retard de la parole. Il permet aux parents de gagner du temps et de l'argent, nécessaires pour ce type de rééducation.

Pour arriver aux objectifs désirés, nous avons enregistré un ensemble de textes par 20 enfants, dans une classe préscolaire. Les textes contiennent tous les sons possibles de la langue arabe. 15 enfants ont prononcé l'ensemble des sons correctement. Cependant, parmi les autres, nous avons trouvé des enfants qui substituent, l'un de ces sons : [s], [z], [r], [ʤ] et [k] respectivement par [θ], [ð], [ʁ], [ʃ] et [t].

Notre système repose sur les modèles HMM/GMM. Il nécessite une étape d'analyse acoustique des signaux de parole en utilisant les MFCC. Pour évaluer SCAESP, nous avons étudié ses performances en fonction des nombres d'états des HMM et des composantes gaussiennes. Un corpus de parole a été élaboré avec 50 enfants de 5-6 ans dans trois écoles primaires algériennes. Le SCAESP donne des performances satisfaisantes concernant les oppositions consonantiques [s]/[θ] ; [z]/[ð] ; [r]/[ʁ] ; [ʤ]/[ʃ] et [k]/[t], qui sont respectivement de : 84.15% ; 80.27% ; 85.57% ; 90.10% et 81.46%.

Mots clé : Classification Automatique de la Substitution Phonémique, Langue Arabe, Erreurs de Substitution Phonémique, MFCC, HMM/GMM.

Table des Matières

Table des Matières

Liste des Tableaux

Liste des Figures

Liste des Abréviations

Acronyme	Signification
AD	Arabe Dialectal
AS	Arabe Standard
dB	décibel
EAP	Erreurs d'Addition Phonémique
EDP	Erreurs de Distorsion Phonémique
EOP	Erreurs d'Omission Phonémique
ESP	Erreurs de Substitution Phonémique
EM	Expectation Maximisation
FFT	Fast Fourier Transform
F_0	Fréquence Fondamentale
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
MFCC	Mel Frequency Cepstral Coefficients
RAL	Reconnaissance Automatique du Locuteur
RAP	Reconnaissance Automatique de la Parole
SCAESP	Système de Classification Automatique des Erreurs de Substitution Phonémique
TCC	Taux de Classification Correct
TOP	Transcription Orthographique-Phonétique
PLP	Perceptual Linear Prediction
CG	Composantes Gaussiennes
TCD	Transformée en Cosinus Discrète

Introduction Générale

1. Introduction Générale

L'objectif principal dans cette thèse est d'élaborer un système d'aide orthophonique basé sur l'une des techniques de la Reconnaissance Automatique de la Parole (RAP). Il est conçu pour aider les enfants algériens âgés entre 5 et 6 ans, ayant des Erreurs de Substitution Phonémique (ESP).

Pour communiquer efficacement, l'enfant doit maîtriser parfaitement son langage parlé. Pendant le développement de son langage, il peut rencontrer un problème de retard de la parole. Ce dernier influe directement sur la communication de l'enfant en montrant quelques erreurs de prononciation, telles que les ESP. Celles-ci sont fréquemment observées chez l'enfant durant les périodes d'apprentissage du langage oral.

Ce problème minimise les compétences linguistiques de l'enfant et rend sa communication avec les individus très compliquée. Généralement, l'enfant apprend la parole d'une manière progressive, en passant par deux périodes principales : pré-linguistique et linguistique. Un enfant de 5 ans est capable de prononcer correctement tous les sons de la langue maternelle. Cependant, dans le cas contraire, il peut avoir un retard de la parole [1].

Les enfants souffrant de troubles d'articulation sont pris en charge par un orthophoniste. Celui-ci fixe un bilan thérapeutique pour corriger les erreurs de prononciation. Cette procédure occupe beaucoup de temps et implique un effort énorme par les enfants et leurs parents. En plus, si la période entre deux séances successives de rééducation est très grande, l'évaluation du niveau de prononciation devient une tâche très difficile.

En Algérie, les enfants ayant ces erreurs rencontrent d'autres problèmes. A titre d'exemple, nous citons le nombre insuffisant des orthophonistes. En sachant que des spécialistes s'installent au niveau des grandes villes et particulièrement dans le secteur privé. Les prises en charge d'un enfant sont coûteuses pour les parents (frais des déplacements et honoraires des spécialistes).

Tous ces problèmes nous ont motivés pour préparer un système d'aide pour la rééducation orthophonique de ce type d'erreurs. Nous avons élaboré ce Système en vue de la Classification Automatique des Erreurs de Substitution Phonémique (SCAESP) en utilisant les Modèles de Markov Cachés ou HMM (Hidden Markov Model) et les Modèles de

Mélange de Gaussiennes ou GMM (Gaussian Mixture Models) [2], [3], [4]. Pour la représentation des signaux de parole, nous avons employé les MFCC (Mel Frequency Cepstral Coefficients).

Pour arriver à nos objectifs, nous avons élaboré un corpus de parole composé de plusieurs mots. Ces derniers sont choisis pour couvrir cinq cas différents de la substitution phonémique. Un ensemble de 50 enfants, âgés entre 5 et 6 ans, ont participé à la procédure d'enregistrement. Cette opération a été réalisée au niveau de trois écoles primaires en Algérie. Le choix de cette tranche d'âge est interprété par l'hypothèse qu'une meilleure rééducation orthophonique doit s'effectuer à un âge précoce.

Le corpus construit est exploité pour analyser ces différents cas, selon deux aspects principaux : le lieu et le mode d'articulation des phonèmes. Il est utilisé aussi pour mesurer les performances du SCAESP en fonction du nombre d'états du HMM et de l'ordre du modèle du GMM.

Finalement, pour simplifier l'utilisation et montrer l'efficacité du système proposé, nous l'avons implémenté sous forme d'une application graphique testée initialement avec un enfant ayant une bonne prononciation puis sur cinq enfants de 5 ans. Après trois mois de rééducation, pour une séance de 30 minutes/semaine, ces enfants ont montré des résultats satisfaisants dans leurs prononciations.

Cette thèse est organisée en quatre chapitres :

- le premier expose les notions de base de la langue Arabe et les problèmes liés au retard simple de la parole. Nous avons donné un aperçu sur les modes et lieux d'articulation, la production de la parole ainsi que sur les Erreurs de prononciation en particulier les Erreurs de la Substitution Phonémique ;
- le deuxième est consacré à la Reconnaissance Automatique de la Parole et les différentes techniques utilisées dans ce domaine. Nous avons donné une attention particulière aux Modèles de Markov Cachés (HMM) et aux Modèles de Mélange de Gaussiennes.
- le troisième porte sur la procédure suivie pour la sélection des cibles dans les ESP et les approches proposées. Pour ce faire, nous avons exploité l'outil d'analyse PRAAT pour extraire les paramètres pertinents : la fréquence fondamentale, la

durée et l'énergie. Nous avons également implémenté l'algorithme de Baum-Welch pour l'apprentissage des modèles et l'algorithme Forward pour la classification des ESP.

- le dernier chapitre concerne les résultats obtenus en évaluant les performances du système proposé pour la classification automatique des ESP en fonction du nombre d'états du HMM et des composantes gaussiennes du GMM. De plus, nous employons l'application développée pour le traitement de ce type de problèmes.

Chapitre 1

Arabe Standard, Acquisition et retard simple de parole

1.1 Introduction

Ce chapitre s'intéresse essentiellement à la langue arabe et au retard simple de parole durant son acquisition. Nous introduisons les notions de base sur l'arabe standard concernant la transcription phonétique, la description et l'articulation des phonèmes arabes. Ensuite, nous décrivons les différentes erreurs de prononciation liées au retard simple de la parole qui regroupent à la fois l'addition, l'omission, la distorsion et la substitution phonémique. Cette dernière est le cadre de notre travail.

1.2 Rappel sur la langue Arabe

La langue arabe est une langue sémitique, rassemblant un large groupe de langues, nommées "Afro-asiatique" et plus lointainement reliées aux familles des langues indigènes de l'Afrique du Nord. Elle possède un héritage littéraire très riche remontant jusqu'à l'ère préislamique. Elle est devenue la langue administrative officielle de l'empire islamique pendant son rayonnement et son expansion. Actuellement, elle est la langue maternelle de plus de 377 millions de personnes dans vingt-deux pays, et la langue liturgique pour plus d'un milliard de musulmans dans le monde entier [5].

La langue Arabe constitue le rameau méridional de plusieurs langues sémitiques associant l'akkadien, l'amorite, l'ougaritique, le cananéen et l'araméen. Cet ensemble inclut, plus que les langues sémitiques, quatre sous-familles : Tamazight; la langue du Tchad; l'ancienne langue de l'Egypte et la langue couchitique [5]. Une hypothèse donne pour berceau à ce sémitique méridional la péninsule arabique d'où serait parti le processus de sémitisation de l'Ethiopie, jusqu'alors domaine des langues couchitiques [6].

La langue Arabe se révèle sous deux aspects principaux : l'Arabe Dialectal (AD) et l'Arabe Standard (AS). L'AD est représentée par les différents idiomes dans les pays arabes. Par contre, l'AS est fixé par l'écrit et offre une situation linguistique très caractérisée. Elle est la langue officielle de plus de 392 millions de personnes. Cependant, du point de vue pratique, les arabophones utilisent, dans la vie quotidienne, l'AD plus que l'AS.

L'Arabe Standard est la langue officielle en Algérie. Elle est, d'une part, la langue de la littérature, de l'éducation, des journaux, des revues et aussi des manifestations scientifiques

et d'autre part, elle est la langue de communication parlée des chaînes télévisées et des stations radios. Cependant, l'AS est essentiellement écrit et lu mais secondairement parlé.

De point de vue pratique, les algériens utilisent l'AD, qui est essentiellement parlé. Il présente quelques variantes régionales aux niveaux phonologiques et lexicaux mais ses variantes ne posent aucun problème dans l'intercompréhension. Généralement, l'AD se distingue de l'AS par une syntaxe simplifiée et un lexique très riche des mots étrangers.

L'Arabe Standard est constitué par 40 phonèmes ; 28 consonnes, 6 voyelles (3 longues et 3 courtes voyelles) et 6 variantes vocaliques en contexte emphatique. Rappelons que le phonème est le plus petit élément des unités de la parole. Il indique la différence dans le sens, le mot et la phrase. Les phonèmes arabes contiennent deux classes distinctives appelées pharyngale et emphatique. Ces deux classes caractérisent les langues sémitiques comme l'arabe et l'hébreu [7], [8].

Nous distinguons les syllabes permises dans la langue arabe qui sont : [CV], [CV:], [CVC], [CV:C], [CVVC] et [CVCC], cette dernière apparaît dans certains cas de coordination entre deux mots successifs, dont [V] indique une voyelle courte, [V:] une voyelle longue et [C] une consonne. Toutes les syllabes arabes commencent par une consonne [7].

1.2.1 Transcription et description phonétique

La transcription phonétique a pour but d'aider les lecteurs étrangers à prononcer les différents mots arabes utilisés dans ce document. Pour l'épellation des phonèmes, on emploie souvent une seule ou une combinaison des lettres pour les représenter. Certaines consonnes sont représentées par des consonnes similaires en prononciation avec les langues étrangères (tableau 1). Cependant, les voyelles présentent quelques différences par ce qu'elles sont ajoutées au-dessus ou au-dessous des consonnes. L'absence de voyelles génère une certaine ambiguïté à deux niveaux : le sens du mot et sa fonction dans la phrase.

Le système vocalique de l'Arabe Standard a six voyelles, dont trois sont courtes et les autres sont longues. La distinction entre courte et longue est semblable à la différence dans la longueur des notes musicales, qui sont jugées deux fois plus longues que les demi-notes.

Les voyelles longues sont représentées par les lettres /ا/ [a:], /و/ [u:] et /ي/ [i:]. Elles sont écrites dans le mot en tant qu'éléments de l'épellation. Cependant, les voyelles courtes ne sont pas des lettres indépendantes, mais elles sont écrites seulement en tant que marques diacritiques au-dessus ou au-dessous du corps du mot. Elles sont représentées par les trois timbres: /بَ/ [ba], /بُ/ [bu] et /بِ/ [bi]. Pratiquement, les voyelles courtes ne sont pas indiquées en texte arabe écrit ; elles sont invisibles.

Tableau 1 : Transcription Orthographique-Phonétique (TOP)

Graphème Arabe	Transcription phonétique	Exemple en Arabe	Exemple en Français
ء	[ʔ]	أسد	Lion
ب	[b]	برتقالة	Orange
ت	[t]	تفاحة	Pomme
ث	[θ]	ثلاجة	Réfrigérateur
ج	[ʤ]	جرار	Tracteur
ح	[ħ]	حديد	Fer
خ	[χ]	خروف	Mouton
د	[d]	دراجة	Bicyclette
ذ	[ð]	ذئب	Loup
ر	[r]	رمان	Grenadier
ز	[z]	زرافة	Girafe
س	[s]	سرير	Lit
ش	[ʃ]	شمعة	Bougie
ص	[sˤ]	صورة	Photo
ض	[dˤ]	ضفدعة	Grenouille
ط	[tˤ]	طائرة	Avion
ظ	[ðˤ]	ظهر	Dos
ع	[ʕ]	علبة	Boîte
غ	[ʁ]	غراب	Corbeau
ف	[f]	فلفل	Piment
ق	[q]	قهوة	Café
ك	[k]	كتاب	Livre
ل	[l]	لغة	Langue
م	[m]	مدرسة	Ecole
ن	[n]	نهر	Fleuve
ه	[h]	هلال	Croissant
و	[w]	وردة	Rose
ي	[j]	يمين	Droite

La prononciation des phonèmes vocaliques change selon la structure du mot et l'influence des consonnes adjacentes. En outre, la voyelle [a:] a plusieurs variantes d'épellation et les

deux phonèmes [w] et [j] sont utilisés parfois comme des voyelles et parfois comme des consonnes.

1.2.2 Articulation des phonèmes arabes

L'Arabe Standard est caractérisée par la présence de deux classes de phonèmes : quatre consonnes emphatiques [sˤ], [dˤ], [tˤ] et [ðˤ] qui sont les versions emphatiques des consonnes dentales [s], [d], [t] et [ð] et cinq pharyngales. Cette dernière classe comprend : deux fricatives [ħ] et [ʕ], caractérisées par la rétraction entre la langue et la partie inférieure du pharynx, et trois uvulaires [χ], [ʁ] et [q] dont les deux premières sont caractérisées par la rétraction formée entre la partie supérieure du pharynx et la langue et l'autre est caractérisée par la fermeture totale au même niveau [9].

Généralement, la prononciation des différents phonèmes arabes se révèle sur leurs points d'articulation, l'état des cordes vocales et l'état de la langue (tableau 2). Selon ces critères, les consonnes peuvent être classifiées comme : labiales, dentales, palatales, vélaires ou laryngales [10].

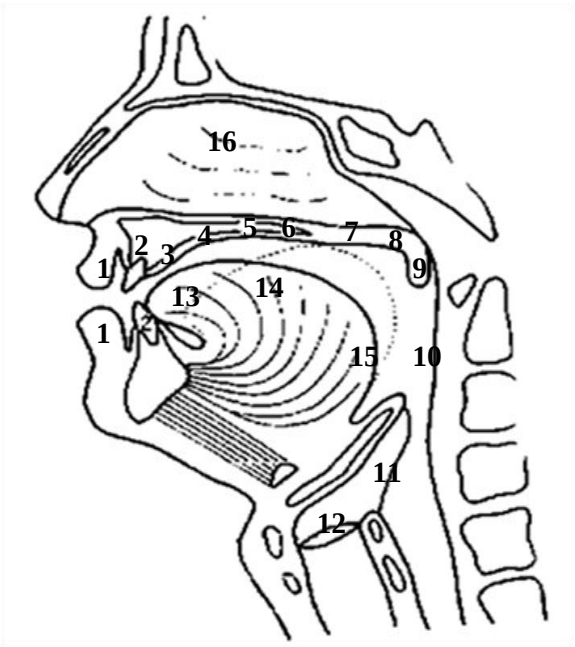
Tableau 2 : Lieux et modes d'articulation des phonèmes arabes [11]
V : Voisé ; NV : Non Voisé

	Plosives		**Nasales**		**Fricatives**		**Liquides**		**Semi-voyelle**	
	V	NV	V	NV	V	NV	V	NV	V	NV
Bilabiale	ب [b]		م [m]						و [w]	
Labiodentale						ف [f]				
Interdentale					ذ [ð]	ث [θ]				
					ظ [ðˤ]					
Alvéolaire	د [d]	ت [t]	ن [n]		ز [z]	س [s]	ر [r]			
		ط [tˤ]			ض [dˤ]	ص [sˤ]	ل [l]			
Alvéopalatale						ش [ʃ]			ي [j]	
Palatale					ج [dʒ]					
Vélaire		ك [k]								
Uvulaire		ق [q]			غ [ʁ]	خ [χ]				
Pharyngale					ع [ʕ]	ح [ħ]				
Glottale		ء [ʔ]			ه [h]					

En outre, on peut les classifier selon leurs tenues comme occlusives ou spirantes. Elles sont occlusives quand leur articulation est le résultat d'une détente et spirantes quand leur articulation peut être prolongée. De plus, si la prononciation d'une consonne s'accompagne

d'une résonance du larynx, elle est considérée comme consonne sonore, ou sourde dans le cas contraire [12].

Chaque phonème se diffère de l'autre par le mode et le lieu d'articulation. Suivant [sibawayh], on distingue 16 lieux d'articulation (figure 1) [13].



1	lèvres	labiales
2	dents	dentales
3	alvéoles	alvéolaires
4	palais dur	pré-palatales
5		médio-palatales
6		post-palatales
7	voile du palais	pré-vélaires
8		post-vélaires
9	luette	uvulaires
10	pharynx	pharyngales
11	larynx	laryngales
12	glotte	glottales
13	langue : Apex	apicales
14	langue : dos	médiodorsales
15	langue : racine	radicales
16	Nasalisation	

Figure 1 : Lieux d'articulation des phonèmes arabes [13]

1.3 Signal de parole

Le signal de parole est une onde acoustique qui apparaît physiquement comme une variation de la pression d'air causée et émise par le système phonatoire. La parole est essentiellement générée comme une onde acoustique, elle est rayonnée par les cavités nasale et buccale lorsque l'air est expulsé des poumons. Le flux résultant de l'air est perturbé par les constrictions à l'intérieur du conduit vocal. Il est utile d'interpréter la production de la parole en termes de 'filtrage acoustique'. Les trois cavités principales du système de production de la parole sont les cavités nasale, buccale et pharyngale, qui forment le filtre acoustique principal. Le filtre est excité par l'air des poumons et est chargé à sa sortie principale par une impédance de rayonnement associée aux lèvres [14].

1.3.1 Production de la parole

La production de la parole est un mécanisme très complexe qui repose sur une interaction entre les systèmes neurologique et physiologique. Nous présentons ici les organes essentiels qui caractérisent l'appareil phonatoire. Généralement, le processus de phonation comporte trois étapes essentielles [15] :

- génération d'une énergie phonatoire qui va servir à mettre en mouvement oscillatoire les cordes vocales ou à les écarter afin de générer un son ;
- vibration des cordes vocales qui intervient lors de la production de tous les sons voisés ;
- réalisation dans un dispositif articulatoire dans ce qu'il est commode de désigner sous le nom de cavités supra-glottiques.

1.3.1.1 Génération d'une énergie phonatoire

L'énergie phonatoire est nécessaire pour actionner les organes du mécanisme de production de la parole. Cette énergie est produite par le flux d'air à partir des poumons et transmise par la trachée à travers les cordes vocales (figure 2).

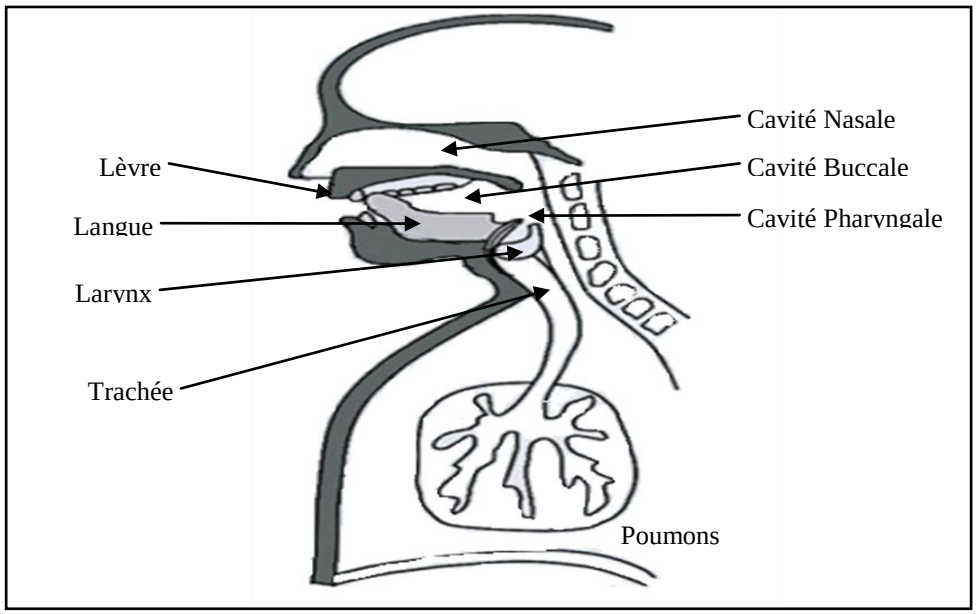


Figure 2 : Appareil phonatoire humain [16]

Pendant l'inspiration, l'air est versé dans les poumons, et pendant l'expiration l'énergie sera spontanément libérée. La trachée transporte l'air résultant au larynx. Ce dernier se réfère comme un fournisseur d'énergie aux entrées des cavités vocales, et le volume d'air

détermine l'amplitude du son. Les poumons sont la source principale de l'aire phonatoire pour tous les sons consonantiques et les voyelles.

1.3.1.2 Cordes vocales et voisement

Les cordes vocales et l'espace triangulaire glottique sont les parties critiques dans la production de la parole (figure 3). Ils séparent la trachée de la base du conduit vocal. L'action des cordes vocales (appelées 'excitation') détermine la nature du son produit. On distingue les sons voisés et les sons non-voisés.

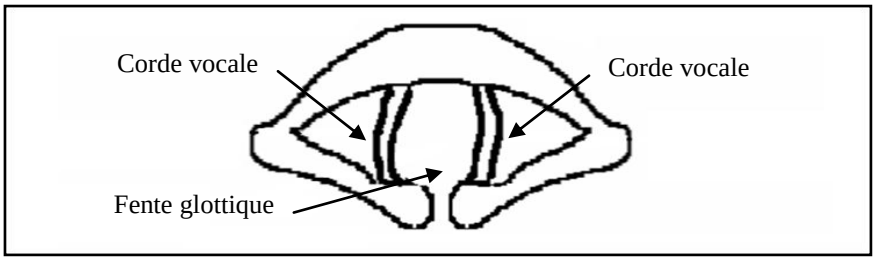


Figure 3 : Cordes vocales

1.3.1.2.1 Sons voisés

Ils sont produits en modulant l'air phonatoire par l'ouverture et la fermeture périodiques des cordes vocales. Sa fréquence de vibration est appelée la fréquence fondamentale F_0. Par conséquent, celle-ci est un facteur de distinction physique important pour les sons voisés et non voisés.

1.3.1.2.2 Sons non-voisés

Ils sont produits par une constriction des tractus vocaux étroits pour faire une turbulence du flux d'air, qui est produit par un bruit ou par une voix respiratoire [17]. Le son non-voisé est souvent considéré comme un bruit blanc. Lors de la production des sons non-voisés, les cordes vocales ne vibrent pas.

1.3.1.3 Articulation du conduit vocal

Le conduit vocal est généralement considéré comme un organe de production de la parole au-dessus des cordes vocales. Il inclut les organes d'excitation et les articulateurs. Les poumons, la trachée et les cordes vocales sont considérés comme des organes responsables de la production d'excitation. Les articulations incluses dans le tractus sont groupées dans :

- le pharynx ;

- le larynx ;
- la cavité buccale ;
- la cavité nasale.

Lorsque l'onde acoustique traverse le conduit vocal, elle sera transformée selon sa forme. Le processus de production de la parole peut être représenté par un modèle source/filtre. Le signal de parole est modélisé comme étant la sortie d'un filtre linéaire variant dans le temps, qui simule les caractéristiques spectrales de la fonction de transfert du conduit vocal, excité par un signal source qui reflète l'activité des cordes vocales dans les zones voisées et le bruit de friction dans les zones non voisées [18].

1.3.2 Perception de la parole

La compréhension d'un message oral peut être décomposée en deux niveaux : transformation de l'information contenue dans le signal acoustique par l'oreille au cerveau et reconstitution du message linguistique dans le cerveau [18].

L'appareil auditif se divise en deux parties : le système auditif périphérique correspondant à ce que l'on nomme communément l'oreille (qui se décompose en oreille externe, moyenne et interne) et le système auditif central. Cependant, tout ce qui peut être acoustiquement mesuré ou observé par la phonétique articulatoire n'est pas nécessairement perçu. Les psycho-acousticiens tentent de comprendre comment l'information auditive est traitée par le cerveau. En effet, au-delà des caractéristiques mesurables (comme l'intensité et la fréquence), le son a deux qualités subjectives, la force articulatoire et la hauteur, qui s'apprécient différemment.

Les qualités subjectives, relèvent des sensations éprouvées par un sujet qui écoute, et ne peuvent pas se mesurer sans lui. Ainsi, l'intensité d'un son est égale à l'intensité physique (mesurée en décibels). Quant à sa hauteur, elle dépend de l'intensité avec laquelle ce son est transmis à l'auditeur [19], [20]. Par ailleurs, la prosodie assure la fonction de segmentation syntaxique et sémantique de l'énoncé. La hauteur d'un son pur est liée à la fréquence de l'onde sonore.

L'échelle de tonie est graduée en Mels. Un écart constant en Mels est perçu comme un écart constant en hauteur. Dans le domaine de l'audition, l'oreille est capable de discerner 1400 hauteurs distinctes. La figure 4 représente le système auditif humain.

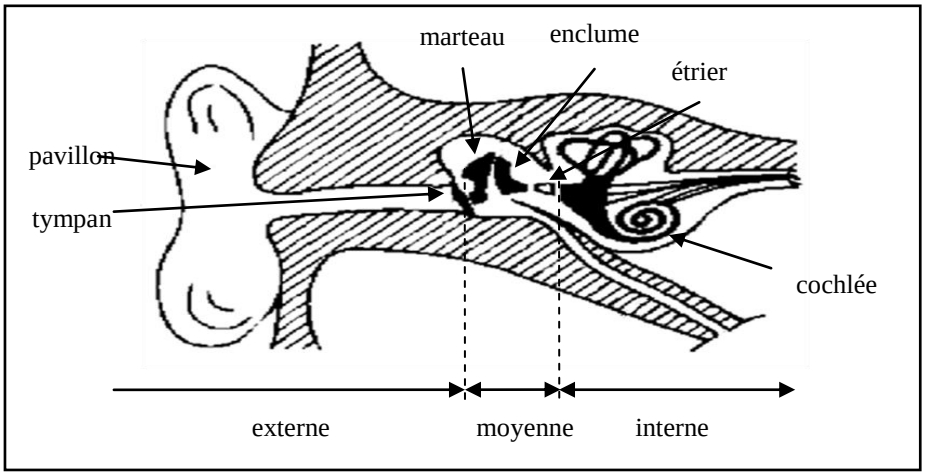


Figure 4 : Système auditif humain [20]

1.4 Apprentissage du système phonétique

L'enfant en petit âge commence à apprendre le système phonétique de la langue maternelle d'une façon progressive. Généralement, cette période est caractérisée par la présence des erreurs de prononciation incluant l'omission et la substitution des phonèmes. Ces erreurs sont considérées comme une partie de l'apprentissage de la parole et la majorité d'enfants n'ont pas l'articulation parfaite jusqu'à l'âge de 6 ans. Cependant, l'enfant, au cours de son progrès vers l'articulation parfaite, doit suivre quelques étapes importantes. Les enfants qui ne rencontrent pas ces étapes importantes, généralement, risquent de perdre le développement correct du langage parlé. Ils peuvent également subir l'anéantissement ou la perte totale de leur langage parlé.

Le développement du langage oral diffère d'un enfant à un autre, suivant plusieurs critères: la capacité neurocognitive; les propriétés génétiques et l'environnement social. Mais, en générale, le déroulement peut être considéré assez fixe, avec quelques variations concernant le temps nécessaire pour sa maîtrise. Toutefois, l'évolution du langage oral chez l'enfant passe par deux périodes essentielles : la période pré-linguistique et la période linguistique proprement dite.

1.4.1 Période pré-linguistique

Le nourrisson de quelques jours commence à différencier entre les diverses événements phonétiques de la langue. Ceci a été observé chez les enfants de 36 à 40 semaines. Vers l'âge de six mois environ, le nourrisson est capable de contrôler les ajustements phonatoires et commence à produire quelques syllabes simples de type consonne-voyelle. Ces syllabes sont très souvent [ba:], [da:] et [ma:]. Dans cette période, la préférence est remarquée pour les consonnes occlusives [b], [t] et [d], la consonne nasale [m] et la voyelle la plus ouverte [a:] [21].

Généralement, l'activité linguistique est organisée en mots qui sont eux-mêmes organisés en morphèmes. On distingue deux sortes de morphèmes : les morphèmes lexicaux qui constituent une marque d'appartenance à une famille de mots et les morphèmes grammaticaux qui constituent des marques des variations de la forme des mots selon les catégories du nombre, du genre, de la personne et du temps.

Pour la langue arabe, l'enfant ne peut apprendre que quelques dizaines de phonèmes, alors que le nombre de mots de cette langue est très grand. Par conséquence, l'apprentissage des mots apparaît difficile que l'apprentissage des phonèmes. Heureusement, le nourrisson présente des compétences intellectuelles nécessaires pour cette période d'apprentissage. Le problème majeur pour l'enfant est de reconnaitre les mots de la langue dans un flot continu de parole qui ne contient pas de marque évidente de frontières entre les mots.

Généralement, l'enfant s'appuie sur trois caractéristiques du langage oral pour progresser dans son apprentissage :

- les contraintes phono-tactiques où certaines séquences de phonèmes marquent une frontière entre les mots;
- les régularités distributionnelles, qui peuvent être caractérisées par l'utilisation de suites de sons fréquemment ; ces suites ayant beaucoup de chance de constituer un mot ;
- enfin, la prosodie de la parole (l'intonation et le rythme) qu'il s'agit pour le bébé d'exploiter pour découvrir les contours des mots.

Lorsque nous parlons, nous ne prononçons pas toutes les syllabes et tous les mots au même rythme et sur le même ton. Le bébé dispose de l'ensemble de ces informations au plus tard vers l'âge de onze mois en moyenne (l'âge du début de l'acquisition des mots). Le tableau 3 présente les différentes étapes de cette période, ainsi que les compétences linguistiques acquises durant cette période.

Tableau 3 : Période pré-linguistique

Naissance	Emploie différents cris pour différents besoins. Sourit quand il voit les personnes familières.
4 à 8 mois	Commence à produire les sons de voyelle clairement. Imite les mouvements de bouche de l'adulte et les voyelles. Emploie la voix pour communiquer l'excitation et le déplaisir. Glougloute quand il joue avec l'adulte.
7 à 14 mois	Emploie la voix pour attirer l'attention. Emploie un ou deux mots corrects comme mama et byby.

1.4.2 Période linguistique

L'apparition des premiers mots et quelques énoncés rudimentaires qui libèrent l'enfant des contraintes du geste et de la mimique, caractérise le début de la période linguistique. Cette période continuera jusqu'à ce que l'enfant puisse communiquer facilement avec les autres. A partir de l'âge de trois ans environ, l'enfant commence à abandonner les structures rudimentaires en les remplaçants par des constructions linguistiques plus conformes au langage de l'adulte. En outre, le vocabulaire de l'enfant s'enrichisse par un nombre important de nouveaux mots.

Dans cet âge, l'enfant utilise les mots qui véhiculent un sens global et généralise les objets ou les situations qui présentent des caractéristiques communes. Les mots utilisés pour ce stade de communication emploient la signification du mot dépendant du contexte (gestes, environnement). Les gestes et les actions accompagnent toujours la langue mais ils ne la substituent pas encore. L'utilisation des gestes et des actions est exprimée par le nombre des mots acquis qui reste réduit au cours de cette période et l'augmentation du capital linguistique ayant pour conséquence un gain de précision dans le sens des mots.

Le vocabulaire de l'enfant augmente lentement à douze mois où l'enfant a acquis cinq à dix mots. Cependant, il augmente très rapidement à partir de deux ans où le nombre des

mots acquis peut atteindre deux cents mots. Dans cette période, l'enfant commence à construire des énoncés en combinant deux mots de sens différents et en même temps la négation commence à apparaitre. La langue employée par l'enfant s'accompagne toujours de simplifications phonématiques. Celles-ci sont caractérisées par une utilisation réduite et imprécise de la gamme des sons de la parole. Généralement, ces simplifications sont causées par des omissions, des substitutions et des assimilations articulatoires.

Ces diverses simplifications sont liées à une progression relativement lente dans l'acquisition du système phonologique. L'ensemble des phonèmes de la langue peut s'acquérir à partir des premières années, jusqu'à l'âge de quatre ans, où il arrive à la maîtrise articulatoire. Toutefois, certaines simplifications phonématiques peuvent persister jusqu'à l'âge de 6-7 ans. L'enrichissement spectaculaire du vocabulaire se passe de 200 mots à l'âge de trois ans à 1500 mots vers l'âge de cinq ans.

Sur le plan qualitatif, on observe l'apparition de réalisations concrètes de phrases dans quelques situations de communication. Cependant, l'enfant doit réaliser des expériences linguistiques principales pour garantir l'acquisition de phrases grammaticales. Ces dernières ne semblent pas résulter d'un simple processus de répétition ou d'imitation. En outre, pour progresser sur le plan du langage, l'enfant effectue une comparaison entre ses propres productions et celles que lui adresse son entourage.

Finalement, l'enfant utilise une langue bien conforme à la langue de l'adulte vers l'âge de six ans. Il peut également utiliser les pronoms personnels et les différentes formes de la langue: négation, interrogation et conjugaison (tableau 4). Généralement, les enfants qui n'acquièrent pas ces compétences à la fin de cette période risquent de rencontrer des troubles concernant leurs langages parlés.

Tableau 4 : Période linguistique

1 à 2 ans	Emploie au moins 10 mots (18 mois). Emploie différentes consonnes aux débuts des mots. Simplifie la parole d'adulte en supprimant une ou plusieurs syllabes. Enrichissement remarquable du vocabulaire, en particulier de 18 à 24 mois. Commence à construire des expressions en concaténant deux mots.
2 à 3 ans	Emploie un seul mot pour plusieurs situations. Utilisation d'expressions et de questions de trois mots.

	Emploie la voix pour demander une chose.
3 à 4 ans	Ses parents peuvent comprendre la majorité de sa parole. Peut décrire les événements qu'il voit dans la maison. Utilisation d'expressions et de questions plus longues. Parle habituellement et couramment sans répéter des syllabes ou des mots.
4 à 5 ans	La voix devient claire. Emploie des phrases avec beaucoup de détails. Raconte quelques histoires. Communique facilement et clairement avec d'autres enfants et adultes. Prononce la majorité des sons correctement. Emploie le dialecte régional et familial.

1.5 Retard simple de parole

La parole est le fruit de combinaisons de différents éléments signifiants qui forment le mot. Au cours de son apprentissage, on trouve des altérations qui vont dans le sens d'une simplification. On note des omissions de phonèmes à la fin ou au milieu du mot et des substitutions résultant d'une économie articulatoire. Le mot ne peut être reproduit dans son ensemble des phonèmes. Ces erreurs sont normales chez l'enfant qui apprend à parler. Leur persistance au-delà de 5-6 ans peut être considérée comme un trouble du langage oral. Parmi ces troubles le retard simple de parole qui nécessite un traitement avant l'entrée en préscolaire [22].

Les troubles du langage oral affectent la façon dont une personne parle. Une personne ayant ce problème sait généralement exactement ce qu'elle veut dire, mais elle a de la difficulté à produire les sons pour communiquer efficacement. Ces troubles du langage oral comprennent une variété de conditions qui affectent les enfants et les adultes. Ils peuvent aller de problèmes pour prononcer un son spécifique à l'incapacité de produire un discours compréhensible.

Ces troubles du langage oral peuvent être les conséquences d'une anomalie organique qui affectent l'un des organes de l'appareil phonatoire (larynx, lèvres, dents, langue et palais) [23], [24]. Dans ce cas, une intervention complète de spécialistes est nécessaire.

Ces troubles Cependant, la majorité de ces erreurs résultent d'un apprentissage incorrect du langage oral. Elles peuvent être classées en quatre grandes familles: la substitution, l'addition, l'omission et la distorsion phonémique [25].

1.5.1 Erreurs de Substitution Phonémique (ESP)

La substitution phonémique est définie comme une mauvaise articulation phonétique où le locuteur remplace un phonème par un autre, lors de la prononciation d'un mot. Les erreurs de substitution sont relativement fréquentes et typiques dans la parole de l'enfant. Elles sont les plus fréquentes parmi les erreurs d'articulation dans l'âge scolaire, bien que la fréquence d'apparition tende à diminuer avec le temps. Ces erreurs sont généralement le résultat d'un changement dans une caractéristique distinctive, à savoir le lieu et le mode d'articulation. La majorité des substitutions phonémiques sont reliées au changement des lieux d'articulation [26], [27].

Ainsi, le son [r] est habituellement remplacé par le son [l]. Par exemple, le mot ([mirwaħa] /ventilateur/ devient [milwaħa]). La substitution peut être remarquée pour le [ʤ] et le [k] qui sont remplacés respectivement par le [ʃ] et le [t] (Exemples : [ʃamal] pour [ʤamal] /chameau/ et [tita:b] pour [kita:b] /livre/) [28].

Généralement, la substitution est produite lors du déplacement du point d'articulation vers l'avant ou vers l'arrière, quand, par exemple, l'enfant substitue le [k] par le [t] donc le point d'articulation avançant de la zone vélaire vers la zone alvéodentale ou lorsque l'enfant substitue le son [r] par le son [ʁ] donc le point d'articulation se déplaçant en arrière (de la zone alvéodentale vers la zone uvulaire).

Notons que, dans certains cas, l'enfant substitue un phonème par différents sons suivant son emplacement dans le mot. Par exemple, il remplace le [s] par le [θ] au début du mot ([θajja:ra] pour [sajja:ra] /voiture/), par le [ʃ] au milieu du mot ([ʃamʃija] pour [ʃamsija] /parasol/) et par le [t] à la fin du mot ([mu:t] pour [mu:s] /rasoir/).

1.5.2 Erreurs d'Omission Phonémique (EOP)

Pour ce type d'erreurs, l'enfant omet l'un des sons et prononce le reste du mot. L'omission phonémique est plus fréquente chez les enfants en bas âge. En outre, les consonnes qui se trouvent à la fin du mot sont omises plus que les sons au début ou au milieu du mot. Le mot résultant est, généralement, incompréhensible sauf s'il est utilisé dans une phrase ou dans un contenu linguistique connu par l'auditeur [25].

Les erreurs d'omission peuvent produire des difficultés pour comprendre ce que l'enfant veut dire (par exemple [ma:m] pour [ħamma:m] /bain/ et [mak] pour [samaka] /poisson/). Elles peuvent être produites lorsque deux consonnes se suivent dans un seul mot ([marsa] ou bien [madsa] pour le mot [madrasa] /école/).

Généralement les enfants, qui souffrent de ce trouble, sont caractérisés par :

- leur parole plus proche de la parole enfantine ;
- une omission de quelques phonèmes plus que les autres. En outre, l'enfant omet des phonèmes au début ou à la fin du mot plus qu'au milieu du mot ;
- une omission diminuant avec le temps. Cependant, ceci apparait fréquemment pour les adultes qui rencontrent quelques problèmes organiques et les enfants qui parlent d'une manière très rapide.

1.5.3 Erreurs de Distorsion Phonémique (EDP)

Pour la distorsion, le phonème émis est proche de sa cible, mais cette dernière n'est pas encore produite de manière acceptable. L'enfant prononce tous les phonèmes possibles mais d'une façon irrégulière pour certains phonèmes [25].

La distorsion peut être produite lorsque l'aire phonatoire circule d'une manière fausse, ou lorsque la position de la langue, pendant la prononciation du son, est incorrecte. Elle est largement propagée chez les adultes plus que chez les enfants. Elle touche des sons particuliers plus que les autres sons, tels que le son [s] qui est prononcé avec un sifflement très long et le son [ʃ] qui est prononcé du côté de la langue [25]. Par exemple, pour le mot [risa:la] /lettre/, on entend un mot proche de [riʃa;la].

1.5.4 Erreurs d'Addition Phonémique (EAP)

Dans ce cas, l'enfant ajoute un son ou une syllabe au mot. En d'autres termes, il prononce le mot d'une manière correcte mais avec un son ou une syllabe supplémentaire (par exemple, [ssamaka] et [mmmirwaħa]). Parfois, on entend une répétition d'une ou plusieurs syllabes du mot ([wa:wa:] et [da:da:]). L'enfant peut également utiliser un son vocalique qui l'aide à prononcer quelques mots difficiles ([ku:u:ra] au lieu de [kura]) [25].

1.6 Conclusion

Dans ce chapitre, nous avons présenté quelques généralités concernant la langue arabe. Ces dernières incluent les différents phonèmes de cette langue, ainsi que leurs transcriptions, descriptions et leurs prononciations phonétiques. Ensuite, nous avons introduit les étapes de l'apprentissage de la langue parlée et nous avons terminé par une présentation des différents troubles existants au cours du développement linguistique. Dans le chapitre qui suit, nous allons introduire une des techniques utilisée pour le traitement et le diagnostic de ces différentes erreurs de prononciation, à savoir, la reconnaissance automatique de la parole.

Chapitre 2

Principales Techniques de la Reconnaissance Automatique de la Parole

2.1 Introduction

Dans ce chapitre, nous allons présenter les différentes techniques utilisées dans un système de Reconnaissance Automatique de la Parole (RAP). Ce système est basé sur les Modèles de Markov Cachés connue sous le nom de HMM (Hidden Markov Models). Nous exposons les méthodes d'extraction des paramètres MFCC et les techniques d'apprentissage et de reconnaissance du système.

2.2 Reconnaissance Automatique de la Parole (RAP)

La Reconnaissance Automatique de la Parole (RAP) est interprétée comme une tâche particulière de la reconnaissance des formes. Elle est considérée parmi les 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. Le champ de ses applications est très large, allant des applications éducatives, des applications financières aux applications médicales, etc.

De point de vue historique, plusieurs travaux ont été réalisés dans le domaine de la pathologie du langage oral. Le développement d'une nouvelle méthode d'extraction des paramètres acoustiques est introduit pour la RAP. Ce système est orienté vers la détection des voix pathologiques. Cette nouvelle méthode est basée sur l'incorporation entre les distributions des parties voisées et non voisées et les caractéristiques Voice Onset et Offset dans le domaine temps-fréquence [29]. L'influence des donnés d'apprentissage dans les systèmes de RAP est étudiée pour le problème de la substitution phonémique dans l'Arabe parlé [28]. Les HMM sont utilisés pour faire une classification automatique des consonnes arrière arabes. Cette étude est réalisée en vue de la correction orthophonique de la substitution phonémique [27].

Pour la classification automatique des voix pathologiques, une nouvelle méthode basée sur les réseaux de neurones (Neural Networks, NN) a été développée. L'algorithme repose sur une technique hybride qui utilise l'énergie des coefficients des ondelettes comme entrée pour les NN [30]. Un système de classification automatique des voix pathologiques, utilisant les chiffres Arabes, est introduit. Ce système exploite les deux premiers formants des deux voyelles Arabes [Fatha] et [Kasra] dans la procédure d'extraction des paramètres, en utilisant deux systèmes de classification: la quantification vectorielle (QV) et les

Réseaux de Neurones Artificiels (Artifical Neural Networks, ANN). Ces derniers montrent un meilleur taux de reconnaissance de 67.86 % pour les voix féminines et de 52.5 % pour les vois masculines [31].

Finalement, pour aider les patients dans la correction orthophonique, une nouvelle méthode de classification basée sur les Réseaux de Neurones a été développée. Cette méthode augmente les performances de la détection, la classification de voix pathologiques et la reconnaissance des chiffres Arabes [32].

2.2.1 Architecture du système

Supposant qu'une personne veuille exprimer une certaine pensée à une autre personne ou à une machine. Pour exprimer cette pensée, elle doit composer une phrase significative sous une forme d'ordre de mots. Une fois que les mots sont choisis, la personne envoie les signaux de commande appropriés aux organes de production de la parole qui forment une expression de la parole et prononce la phrase désirée. Cette phrase est représentée par un signal acoustique $s[n]$.

De point de vue fonctionnement, un système de RAP se décompose en deux processus différents : le premier concerne un processus acoustique qui analyse le son articulé et le transforme en une suite de vecteurs acoustiques, X, qui caractérisent le signal de parole. L'autre est linguistique et trouve le message correspondant $\widetilde{\Phi}$ selon le critère de maximum de vraisemblance (figure 5).

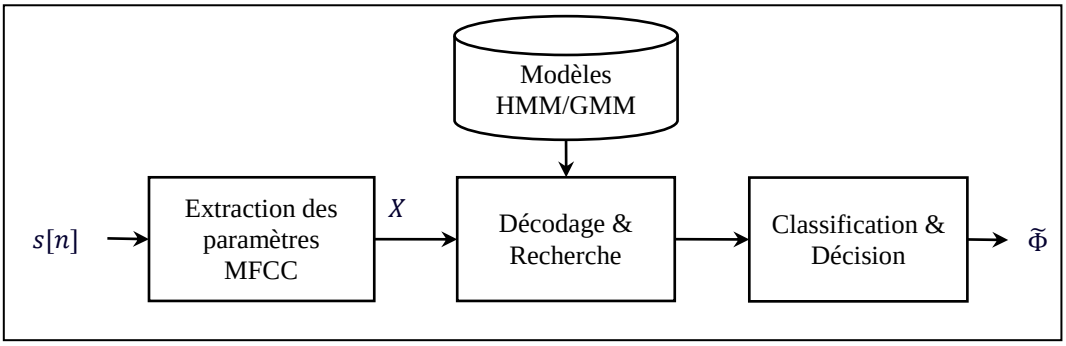


Figure 5 : Schéma fonctionnel d'un système de RAP

Le signal acoustique d'entrée, $s[n]$ est transformé en une suite de vecteurs de coefficients acoustiques. Généralement, ces vecteurs sont donnés par une représentation cepstral et sont calculés sur chaque trame. En particulier, les MFCC (Mel Frequency Cepstral

Coefficients) [33] et les PLP (Perceptuel Linear Prediction) [34] sont souvent utilisés pour représenter les caractéristiques spectrales à court terme. Le bloc de classification décode les vecteurs acoustiques dans une représentation symbolique, selon le sens de maximum de vraisemblance pour estimer le modèle $\widetilde{\Phi}$qui produit l'ordre des vecteurs acoustiques d'entrée. Le bloc final de ce système est un processus de vérification et de décision employé pour mesurer la confiance pour chaque mot identifié. Chacune de ces opérations implique beaucoup de détails et de calcul numérique étendu.

Généralement, les différentes étapes pour construire un système de RAP sont les suivantes:

- choisir l'ensemble des paramètres acoustiques et les traitements nécessaires pour mieux représenter les propriétés du signal de la parole ;
- choisir la tâche de reconnaissance, l'unité acoustique, le modèle de langage ou de syntaxe, et la tâche sémantique ;
- apprentissage des modèles acoustiques et linguistiques ;
- calculer et évaluer les performances du système résultant de la RAP.

2.2.2 Implémentation du système RAP

Le problème de RAP est représenté comme un problème de décision statistique. Il est formulé comme procédé de décision selon le critère de Maximum A Posteriori (MAP) où on cherche à trouver le mot qui maximise la probabilité a posteriori $P(X/\Phi)$ employant les paramètres des coefficients acoustiques [35], [36].

$$\widetilde{\Phi} = \arg\max_{\Phi} P(\Phi/X) \tag{01}$$

Suivant la règle de Bayes, on peut écrire l'équation 1 sous la forme :

$$\widetilde{\Phi} = \arg\max_{\Phi} \frac{P(X/\Phi)\mathrm{P}(\Phi)}{P(X)} \tag{02}$$

L'équation (02) montre que le calcul de la probabilité a posteriori est décomposé en deux termes, l'un définie la probabilité a priori de la séquence des mots, Φ, connaissant $P(\Phi)$, et l'autre définie la vraisemblance du score des mots qui produit l'ensemble des vecteurs acoustiques X connaissant $P(X/\Phi)$. Ces deux probabilités sont estimées d'un ensemble de données d'apprentissage choisies soigneusement par des experts. Le terme $P(X)$ est indépendant de la séquence des mots à optimiser, donc négligeable dans tous les calculs futurs.

L'équation (02) est souvent écrite sous une forme d'étapes comme suite :

$$\widetilde{\Phi} = \underbrace{\arg\max_{\Phi}}_{étape\ 3} \underbrace{P(X/\Phi)}_{étape\ 1} \underbrace{P(\Phi)}_{étape\ 2} \quad (03)$$

Avec :

- étape 1 : calcul de la probabilité associée au modèle acoustique des signaux de parole dans la phrase Φ ;
- étape 2 : calcul de la probabilité associée au modèle linguistique des mots de l'expression ;
- étape 3 : calcul des scores associés à toutes les phrases valides dans la langue, selon le critère du maximum de vraisemblance.

Afin d'être plus explicite au traitement des signaux de parole et des calculs associés à chacune des trois étapes, nous avons besoin de mieux expliciter la relation entre les vecteurs de coefficients acoustiques X et la séquence des mots Φ. Le vecteur des paramètres acoustiques X peut être considéré comme des observations acoustiques correspondant à chacune de T trames du signal de parole :

$$X = \{x_1, x_2, \dots, x_T\} \quad (04)$$

T : nombre de trames

2.2.2.1 Représentation du signal vocal

Le signal de parole est le résultat d'une opération complexe rassemblant tous les organes de l'appareil phonatoire. L'excitation de la cavité orale ou nasale par une source acoustique produit une onde acoustique qui véhicule le signal de parole [37]. Ce signal peut être considéré comme une concaténation des réalisations acoustiques, qui sont produites par des actions et des mouvements de l'appareil phonatoire. Chaque réalisation élémentaire, dans le signal de parole, est vue comme un phonème.

Le phonème est la plus petite unité dans un signal de parole. Il permet la distinction entre deux mots différents [37]. Par exemple, en arabe, le phonème [**d**] dans le mot [ʔaħma**d**] est différent du phonème [**r**] dans le mot [ʔaħma**r**] par ce que les deux mots ont des sens totalement différents. Les organes de l'appareil phonatoire sont soumis à des contraintes mécaniques qui limitent les variations rapides des éléments mobiles (la langue et les lèvres). L'influence d'un phonème sur les phonèmes voisins est connue sous le nom de 'coarticulation'.

La réalisation acoustique d'un même phonème, pour le même locuteur, peut varier en durée et en même temps dans la forme du conduit vocale utilisée pour la production. En outre, cette variabilité est étendue lorsque la réalisation d'un seul phonème est produite par plusieurs locuteurs d'âge, de sexe et de morphologie différents. La forme du conduit vocale est propre pour chaque locuteur, d'où pour un même phonème on peut trouver beaucoup de réalisations acoustiques [38].

Généralement, le signal de parole est considéré comme une association de plusieurs entités élémentaires stationnaires. Cependant, un seul mot peut être prononcé de différentes façons. Les origines de ces variabilités peuvent être vues comme des variabilités inter-locuteur et intra-locuteur. En outre, la transmission du signal acoustique, l'aire, le microphone et le câblage influent également sur le signal de parole. Toutes ces variabilités et difficultés rendent la tâche de reconnaître le mot désiré très complexe.

2.2.2.2 Prétraitement des signaux de parole

La figure 6 montre un schéma général du processus de traitement appliqué. Le signal est prétraité et segmenté dans une suite de trames de 30 ms. Généralement, ces trames sont recouvertes entre eux. Si T_s est le décalage entre deux trames consécutives, alors $1/T_s$ est la fréquence des trames (en Hertz). Dans cet intervalle, le signal de parole peut être considéré quasi stationnaire. Afin d'obtenir le spectre à court terme de chaque trame, les paramètres spectraux obtenus sont habituellement transformés pour fournir aux trames une représentation faiblement corrélé et de dimensions réduites. Chaque trame est représentée par un vecteur *x* de coefficients acoustiques contenant les paramètres d'analyse. Le vecteur de paramètres acoustiques est habituellement augmenté en ajoutant d'autres termes tels que l'énergie ou les coefficients dynamiques.

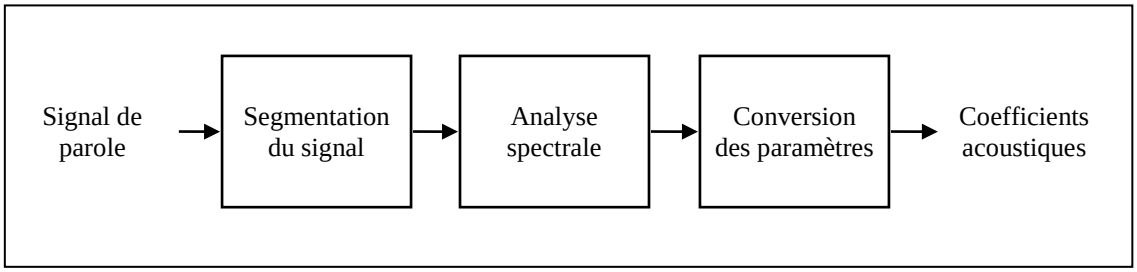


Figure 6 : Schéma générale de l'étape d'analyse d'un signal de parole

Le signal de parole est le résultat d'une convolution de la source par le conduit vocal. Ainsi, la déconvolution de ce signal exige de transporter le problème par homomorphisme dans un espace où la convolution est remplacée par une somme.

Afin de relever les hautes fréquences qui sont moins énergétiques que les basses fréquences, le signal de parole est en premier lieu pré-accentué. Cette étape consiste à faire passer le signal $x(t)$ dans un filtre numérique à réponse impulsionnelle finie de premier ordre donné comme suit [39] :

$$\mathrm{H(z)} = 1 - \mathrm{az}^{-1} \tag{05}$$

Le signal pré-accentué x_p est lié au signal d'origine *x* par la formule suivante :

$$x_p(t) = x(t) - ax(t-1) \tag{06}$$

Avec : $0.9 \leq a \leq 1$.

A ce niveau, deux principales méthodes existent pour l'analyse spectrale dans les systèmes de RAP : le banc de filtres et le codage prédictif (LPC). Cette dernière a été classiquement utilisée pour plusieurs raisons : elle est basée sur une méthode puissante de production de la parole, ce modèle est approprié aux sons voisés et en plus acceptable pour les sons non voisés. En outre, pour une parole de bonne qualité, la prédiction linéaire fournit de bons résultats que les méthodes basées sur le banc de filtres [40]. Cependant, ces derniers sont l'outil principal d'analyse lors de ces dernières années puisqu'ils montrent de meilleurs résultats en présence du bruit [41].

2.2.2.3 Extraction des paramètres MFCC

Le processus d'extraction des paramètres du signal de parole consiste à transformer ce signal en une suite des vecteurs acoustiques. Cette nouvelle représentation est plus compacte à la modélisation statistique et vectorielle. Pour le problème de la RAP, les paramètres les plus employés reposent sur une représentation cepstral du signal de parole (figure 7).

Le signal préaccentué est analysé par une fenêtre glissante de courte durée de l'ordre de 25 à 30 ms avec un recouvrement de 50 % où le signal de parole peut être considéré

quasi-stationnaire. La fenêtre souvent utilisée dans les systèmes RAP est la fenêtre de Hamming :

$$H(n) = \begin{cases} 0.54 - 0.46cos(2\pi n/N) & 0 \leq n \leq N-1 \\ 0 & ailleurs \end{cases} \quad (07)$$

Le processus de fenêtrage transforme le signal de parole en une suite des trames à court terme. La transformée de Fourier est calculée pour chacune de ces trames pour obtenir le spectre du signal. On peut trouver plusieurs algorithmes pour cette transformée. Ces algorithmes sont connus sous le nom de FFT (Fast Fourier Transform) [42].

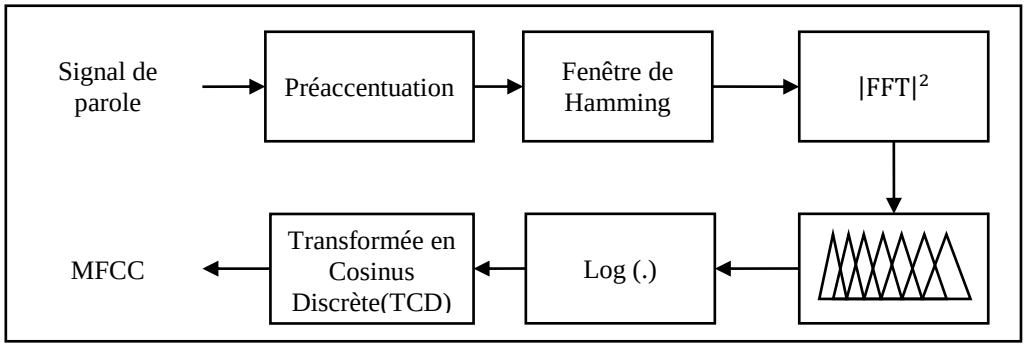


Figure 7 : Extraction des paramètres MFCC

Le spectre du signal présente beaucoup de fluctuations. Néanmoins, nous nous sommes intéressés qu'à l'enveloppe du spectre. En outre, la réduction de la taille des vecteurs spectraux exige une autre raison pour lisser le spectre du signal. Le lissage du spectre est obtenu en multipliant le spectre résultant par un banc de filtres, tenant compte de la réponse acoustique de l'oreille humaine. Ce banc de filtre est une série de filtres à bande passante équidistante dans l'échelle Mel [43]. Pour définir un banc de filtre, on doit définir la forme de chaque filtre ainsi que la localisation de ses fréquences gauche, centrale et droite. Généralement, ces filtres prennent une forme triangulaire et peuvent être différemment placés sur l'échelle de fréquences. La localisation des fréquences centrales des filtres est donnée par :

$$\mathrm{f_{mel}} = 2595 \log_{10}\left(1 + \frac{\mathrm{f}}{700}\right) \quad (08)$$

Avec f : fréquence en Hz

Finalement, nous prenons le logarithme de cette enveloppe spectrale et nous calculons l'enveloppe spectrale en dB. Ensuite, les coefficients cepstraux sont obtenus par une

transformée en cosinus discrète à partir des logarithmes des énergies issues du banc de filtres. L'expression de ces coefficients est donnée par :

$$MFCC(i) = \sqrt{\frac{2}{K}} \sum_{j=1}^{K} S_j cos\left[(j - 0.5)\frac{i.\pi}{K}\right] \quad (09)$$

Avec :

- $i = 1, 2, \ldots, L$;
- K : nombre de coefficients spectraux calculés précédemment ($K = 23$) ;
- S_j : coefficients spectraux ;
- L : nombre de coefficients cepstraux que nous voulons calculer ($L = 12$).

Jusqu'ici aucune information sur l'évolution de temps n'est incluse dans les MFCC. L'information dynamique dans le signal de parole est également différente d'un locuteur à l'autre. Cette information est souvent donnée par les dérivées cepstraux. Les dérivées premières sont les coefficients Δ. Elles montrent la vitesse de variation de ces vecteurs dans le temps. Les dérivées deuxièmes sont les coefficients ΔΔ. Elles donnent des informations sur l'accélération de la parole. Ces coefficients sont donnés par [32] :

$$\Delta MFCC_l(i) = 0.375 \sum_{p=-K}^{K} p\left(\Delta MFCC_{l-k}(i)\right) \quad (10)$$

$$\Delta\Delta \mathrm{MFCC_l(i)} = [\Delta \mathrm{MFCC_{l+1}(i)} - \Delta \mathrm{MFCC_{l-1}(i)}] \quad (11)$$

En général, le nombre de coefficients est pris égal à 13, et parfois réduit à 12, en considérant deux points essentiels :

- le premier coefficient $\mathrm{C_0}$ représentant l'énergie de la trame et ne pouvant réellement contribuer à la reconnaissance ;
- les 12 coefficients représentant l'enveloppe cepstral plus ou moins lissée, avec une suppression des hautes variations fréquentielles.

2.3 Modèles de Markov Cachés HMM

La majorité des systèmes de RAP reposent principalement sur les Modèles de Markov Cachés (HMM). Ils sont fondés sur l'hypothèse que la parole est une succession de plusieurs unités acoustiques (phonèmes). D'un point de vue pratique, un HMM est considéré comme une généralisation du Modèle de Markov.

2.3.1 Modèle de Markov

Supposant le processus représenté par un ensemble de N-états (figure 8), chaque état représente un événement ou une observation. Le système change d'un état vers un autre (transition) dans chaque intervalle de temps. On dénote par s_tl'état à l'instant$t$.

Le processus de Markov (chaîne de Markov) est caractérisé par la dépendance de l'état actuel avec les états précédents. En d'autres termes, le processus a une "mémoire". Dans le cas d'un processus de Markov discret de premier ordre, l'état actuel dépend uniquement de l'état précédent, indépendamment du temps considéré. Ce processus est décrit par les probabilités de transition d'un état à un autre :

$$a_{ij} = P(s_t = i/s_{t-1} = j) \tag{12}$$

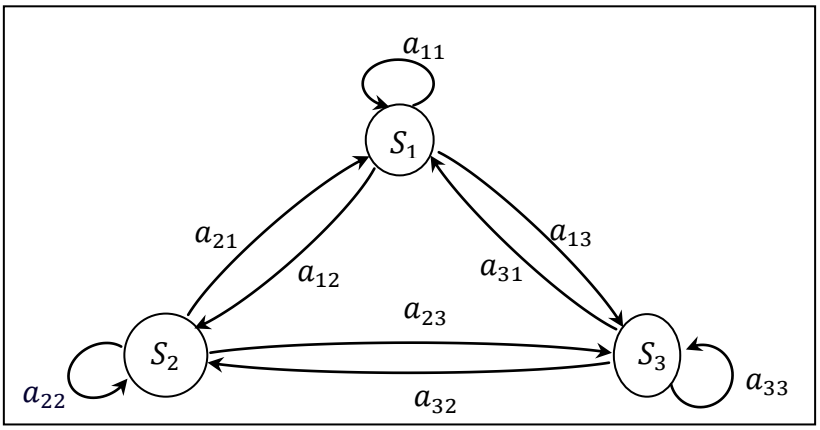


Figure 8 : Processus de Markov discret

2.3.2 Définition du Modèle de Markov Caché

Un HMM est constitué de deux processus superposés : l'un observable qui donne la séquence d'observations, et l'autre caché représentant la séquence d'états. Ceci est un processus non-déterministe qui génère les symboles d'observation de la sortie dans un état donné. Ainsi, l'observation est une fonction probabiliste de l'état. Généralement, un HMM est essentiellement une chaîne de Markov où l'observation de la sortie est une variable aléatoire générée selon une fonction probabiliste de sortie associée à chaque état.

La méthode la plus largement utilisée pour la construction des modèles acoustiques (modèle du mot ou sous-unité) est basée sur les HMM [44], [45], [4]. La figure 9 montre un simple HMM de 5-états pour modeler un mot entier. Chaque état est représenté par une distribution de densités multi-gaussiennes, qui caractérise le comportement statistique des

vecteurs de coefficients acoustiques dans les états du modèle [46]. En outre, un HMM est aussi caractérisé par un ensemble de transitions d'état, a_{ij}, qui indiquent la probabilité de faire une transition de l'état i vers l'état j à chaque trame.

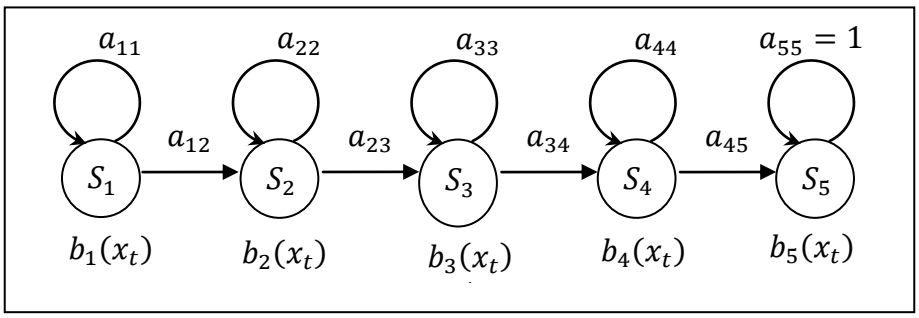


Figure 9 : 5-HMM gauche-droite 'modèle du mot'

Il est possible de basculer d'un HMM pour un modèle du mot (figure 11) vers un HMM pour le modèle de sous-unité (figure 10). Celui-ci est un HMM simple de 3-états basé sur un modèle d'unité secondaire, où le premier état représente les caractéristiques statistiques au début du son, l'état moyen représente le cœur du son et l'état final représente les caractéristiques spectrales à la fin du son. Un modèle de mot est obtenu en concaténant les modèles des phonèmes en regroupant un HMM de 3 états du son [f] avec un HMM de 3-états du son [i:], pour donner le modèle du mot (prononcé comme [fi:]).

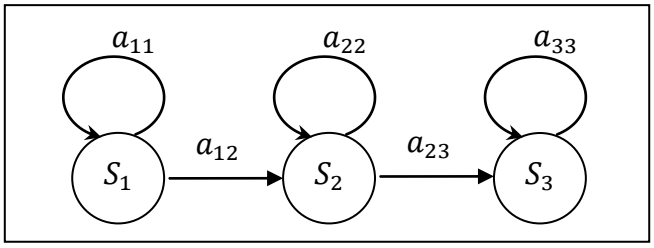


Figure 10 : 3-HMM 'modèle sous-unité'

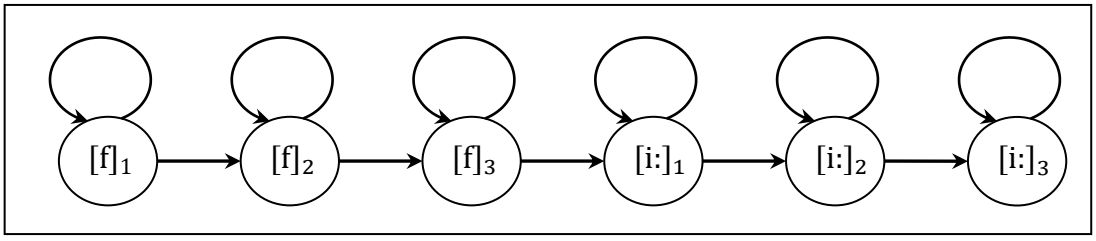


Figure 11 : Un HMM 'modèle du mot' pour le mot [fi:]

Généralement, la composition d'un mot à partir de sous-unités est indiquée dans le lexique ou le dictionnaire du mot. Cependant, si le modèle d'un mot a été établi, il peut être

employé comme un modèle du mot entier soit pour l'apprentissage ou bien pour le test du processus de reconnaissance de la parole.

Formellement, un HMM est défini par:

- $O = \{o_1, o_2, \dots, o_M\}$, alphabets d'observation de la sortie. Les symboles d'observation correspondent à la sortie du système en cours de modélisation ;
- $\Omega = \{1,2, \dots, N\}$, ensemble d'états représentant l'espace d'état. On note s_t comme l'état à l'instant t ;
- $A = \{a_{ij}\}$, matrice de probabilité de transition, où a_{ij} est la probabilité de faire une transition de l'état i à l'état j ;

$$a_{ij} = P(s_t = i / s_{t-1} = j) \quad (13)$$

- $B = \{b_i(k)\}$, matrice de probabilité de la sortie, où $b_i(k)$ est la probabilité de l'élément émis o_k lorsque l'état i est entré.
- soit $X = x_1, x_2, \dots, x_t, \dots$la sortie observée du HMM. La séquence d'états $S = s_1, s_2, \dots, s_t, \dots$ est cachée, et $b_i(k)$ peut-être réécrite comme suit:

$$b_i(k) = P(X_t = o_k / s_t = i) \quad (14)$$

- $\pi = \{\pi_i\}$, modèle initial du HMM :

$$\pi_i = P(s_0 = i), \quad 1 \leq i \leq N \quad (15)$$

Pour résumer, un HMM comprend deux paramètres constants, Net M, représentant le nombre total d'états et de la taille des alphabets d'observation. Il est aussi caractérisé par les observations alphabets O et les trois matrices de probabilité A, B et π. Généralement, on peut représenter un HMM par la notation suivante :

$$\Phi = (A, B, \pi) \quad (16)$$

L'utilisation des HMM dans les systèmes de RAP suppose de pouvoir résoudre le problème [40], [47] :

- d'évaluation : calculer $P(X/\Phi)$ à partir des observations $X = \{x_1, x_2, \dots, x_T\}$ connaissant le modèle Φ;
- de décodage : déterminer la séquence d'états $S = \{s_1, s_2, \dots, s_N\}$ la plus probable des observations $X = \{x_1, x_2, \dots, x_T\}$ connaissant le modèle Φ;

- d'apprentissage : déterminer les paramètres du modèle $\widetilde{\Phi}$ qui maximise la probabilité $P(X/\Phi)$.

2.3.2.1 Evaluation d'un HMM

La probabilité de la séquence d'observations X connaissant le modèle Φ est obtenue par la somme des probabilités de toutes les séquences d'états possibles :

$$P(X/\Phi) = \sum_{S} P(S/\Phi)P(X/S,\Phi) \tag{17}$$

Pour une séquence particulière d'état $S = (s_1,\ s_2, \dots ,\ s_T)$, dont s_1 est l'état initial, la probabilité de la séquence d'état dans l'équation (17) peut être réécrite en appliquant l'hypothèse de Markov:

$$P(S/\Phi) = \pi_{s_1} a_{s_1 s_2} \dots a_{s_{T-1} s_T} \tag{18}$$

Pour la même séquence d'état S, la probabilité conjointe de la sortie le long du chemin peut être réécrite en appliquant l'hypothèse d'indépendance des états :

$$\mathrm{P(X/S,\Phi)} = \prod_{\mathrm{t=1}}^{\mathrm{T}} \mathrm{P(x_t/s_t,\Phi)} \tag{19}$$

$$P(X/S,\Phi) = b_{s_1}(x_1) b_{s_2}(x_2) \dots b_{s_T}(x_T) \tag{20}$$

$$P(X/\Phi) = \sum_{S} \pi_{s_1} b_{s_1}(x_1) a_{s_1 s_2} \dots a_{s_{T-1} s_T} b_{s_T}(x_T) \tag{21}$$

Le calcul de cette probabilité prend beaucoup de temps. Ainsi, pour un modèle de N états et une séquence d'observations de durée T, il nécessite $(2T-1)N^T$ multiplications et $(N^T - 1)$ additions. En outre, pour calculer $P(S/\Phi)$, on numérote tous les états possibles S de taille T et on calcule la somme de toutes les probabilités. Ces dernières sont données pour chaque chemin par le produit de la probabilité de la séquence d'états et la probabilité conjointe de la sortie le long du chemin. Elle est connue comme l'algorithme Forward [38].

$$\alpha_t\ (i) = P(x_1, \dots, x_t, s_t = i\ /\ \Phi) \tag{22}$$

Cette probabilité peut être calculée récursivement [20] :

$$\alpha_t\ (j) = \left[\sum_{i=1}^{N} \alpha_{t-1}\ (i)\ a_{ij}\right] b_i(x_t) \tag{23}$$

Avec :

$$\alpha_1\ (i) = \pi_i b_i(x_1) \tag{24}$$

$$P(X/\Phi) = \sum_{i=1}^{N} \alpha_T\ (i) \tag{25}$$

2.3.2.2 Décodage d'un HMM

Pour résoudre le problème du décodage, un HMM utilise l'algorithme de Viterbi [48]. Cet algorithme permet de trouver la séquence optimale d'états qui maximise $P(X, S/\Phi)$. Il peut être considéré comme l'algorithme de programmation dynamique appliqué au HMM ou comme une version modifiée de l'algorithme avant. Au lieu de manipuler les probabilités de tous les chemins possibles, le choix de cet algorithme se souvient uniquement des meilleurs chemins. Cette probabilité est définit par :

$$V_t(i) = P(x_1, \dots, x_t, s_1, \dots, s_{t-1}, s_t = i\ /\ \Phi) \tag{26}$$

L'initialisation de ces deux quantités est donnée par :

$$V_1(i) = \pi_i b_i(x_1) \tag{27}$$

$$B_1(j) = 0 \tag{28}$$

On calcule par récurrence :

$$V_t\ (j) = \max_{1 \le i \le N} \left[V_{t-1}(i) a_{ij}\right] b_j(x_t) \tag{29}$$

$$B_t\ (j) = arg \max_{1 \le i \le N} \left[V_{t-1}(i) a_{ij}\right] \tag{30}$$

La probabilité finale du chemin optimale est alors :

$$P_{opt} = \max_{1 \le i \le N} [V_t(i)] \tag{31}$$

Et la séquence optimale des états $S^* = \{s_1^*, s_2^*, \dots, s_T^*\}$ est obtenue par :

$$s_T^* = arg \max_{1 \le i \le N} [B_T(i)] \tag{32}$$

2.3.2.3 Estimation d'un HMM

Il est important d'estimer les paramètres du modèle Φ pour décrire avec précision les séquences d'observation. Ceci est le plus difficile parmi les trois problèmes, parce qu'il n'y a aucune méthode analytique pour maximiser la probabilité conjointe des données

d'apprentissage. En outre, ce problème peut être résolu par l'algorithme itératif de Baum-Welch, connu sous le nom d'algorithme Forward-Backward [49].

Le problème d'apprentissage des HMM est un cas typique de l'apprentissage non supervisé, où les données sont incomplètes en raison de la séquence d'état caché. L'algorithme EM (Expectation Maximisation) est parfaitement adapté à ce problème. En fait, Baum-Welch ont utilisé le même principe que celui de l'algorithme EM. D'une manière similaire, on définit la probabilité Backward comme suit :

$$\beta_t(i) = P(x_{t+1}, \dots, x_T / s_t = i\,, \Phi) \tag{33}$$

Où $\beta_t(i)$ est la probabilité de générer les observations partielles $x_{t+1}, x_{t+2}, \dots, x_T$ connaissant le HMM pour l'état i dans le temps t. Elle peut être calculée récursivement :

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(x_{t+1}) \beta_{t+1}(j) \tag{34}$$

Avec :

$$t = T - 1, \dots, 1 \tag{35}$$

$$\beta_T(i) = \frac{1}{N} \tag{36}$$

Ensuite, on définit $\gamma_t(i,j)$, qui donne la probabilité de faire une transition de l'état i vers l'état j à l'instant t, connaissant le modèle Φ et la séquence d'observation :

$$\gamma_t(i,j) = P(s_{t-1} = i, s_t = j / x_1, \dots, x_T, \Phi) \tag{37}$$

$$= \frac{P(s_{t-1} = i, s_t = j, x_1, \dots, x_T / \Phi)}{P(x_1, \dots, x_T / \Phi)} \tag{38}$$

$$= \frac{\alpha_{t-1}(i) a_{ij} b_j(x_t) \beta_t(j)}{\sum_{k=1}^{N} \alpha_T(k)} \tag{39}$$

Nous pouvons réestimer itérativement le vecteur des paramètres du HMM, $\Phi = \{A, B, \pi\}$, en maximisant la vraisemblance $P(X/\Phi)$ pour chaque itération. Nous utilisons $\widetilde{\Phi}$ pour désigner le nouveau vecteur de paramètres dérivé du vecteur des paramètres Φ dans l'itération précédente. Selon l'algorithme EM, le processus revient à maximiser la fonction Q suivante :

$$Q(\Phi, \tilde{\Phi}) = \sum_{S} \frac{P(X, S/\Phi)}{P(X/\Phi)} \log P(X, S/\tilde{\Phi}) \tag{40}$$

Où $P(X, S/\Phi)$ et $\log P(X, S/\tilde{\Phi})$ peuvent être donnés par :

$$P(X, S/\Phi) = \prod_{t=1}^{T} a_{s_{t-1}s_t} b_{s_t}(x_t) \tag{41}$$

$$\log P(X, S/\tilde{\Phi}) = \sum_{t=1}^{T} \log a_{s_{t-1}s_t} + \sum_{t=1}^{T} \log b_{s_t}(x_t) \tag{42}$$

Finalement, les paramètres à estimer sont donnés par les deux relations suivantes :

$$\tilde{a}_{ij} = \frac{\sum_{t=1}^{T} \gamma_t(i,j)}{\sum_{t=1}^{T} \sum_{k=1}^{N} \gamma_t(i,k)} \tag{43}$$

$$\tilde{b}_j(k) = \frac{\sum_{t \in X_t = o_k} \sum_i \gamma_t(i,j)}{\sum_{t=1}^{T} \sum_i \gamma_t(i,j)} \tag{44}$$

Pour la RAP, le processus d'alignement de chaque trame du signal de parole à un modèle approprié du mot dans une phrase, est basé sur une procédure d'alignement optimal. Celle-ci est obtenue entre la séquence concaténée des modèles du mot et la séquence des vecteurs de coefficients acoustiques du signal de parole d'entrée. Le modèle du mot est encore décomposé en un ensemble d'états qui reflète le changement des propriétés statistiques des vecteurs de coefficients acoustiques dans le temps. Chacun de ces mots est représenté par un HMM de N-états. Dans chaque état, il y a une densité de probabilité qui caractérise les propriétés statistiques des vecteurs acoustiques.

2.4 Modèle de Mélange de Gaussiennes GMM

La densité de probabilité de chaque état, et pour chaque mot, est estimée pendant la phase d'apprentissage du système de reconnaissance. Pour caractériser la distribution des vecteurs des coefficients acoustiques dans chaque état j du modèle de mot, on utilise un mélange de densités Multi-Gaussiennes, notée $b_j(x)$, et données par la formule [50] :

$$b_j(x) = \sum_{m=1}^{M} c_{jm} \frac{\exp\left(-\frac{1}{2}(x - \mu_{jm})^{*} \Sigma_{jm}^{-1} (x - \mu_{jm})\right)}{(2\pi)^{\frac{D}{2}} \left|\Sigma_{jm}\right|^{\frac{1}{2}}} \tag{45}$$

Avec :

- K : nombre des composantes gaussiennes ;
- c_{jm} : poids du mélange m dans l'état j, avec $c_{jm} > 0$;
- μ_{jm} : vecteur de moyenne ;
- Σ_{jm} : matrice de covariance.

En vérifiant les deux contraintes suivantes :

$$\sum_{m=1}^{M} c_{jm} = 1, \qquad 1 \leq j \leq N \tag{46}$$

$$\int_{-\infty}^{+\infty} b_j(x_t)\, dx_t = 1, \qquad 1 \leq j \leq N \tag{47}$$

La probabilité qu'une trame x_t soit associée à l'état $j(t)$ avec le mot $i(t)$, P, est donnée par:

$$P\left(x_t / w_{j(t)}^{i(t)}\right) = b_{j(t)}^{i(t)}(x_t) \tag{48}$$

Le calcul de l'équation 48 est incomplet puisque on a ignoré le calcul de la probabilité associée aux liens entre les états des mots. On n'a pas aussi indiqué comment déterminer l'état à l'intérieur du mot pendant l'alignement entre un mot donné et l'ensemble des vecteurs de coefficients acoustiques correspondant à ce mot.

L'idée est qu'on assigne une probabilité à chaque réalisation acoustique de la séquence des mots utilisant les HMM de l'ensemble des vecteurs des coefficients acoustiques. Celle-ci implique un ensemble indépendant des données d'apprentissage, pour extraire les meilleurs paramètres des modèles acoustiques pour chaque mot. Les paramètres à estimer sont le vecteur de moyenne, la matrice de covariance et le vecteur de poids.

$$\bar{\mu}_{jm} = \frac{\sum_{t=1}^{T} \xi_t(j,m) x_t}{\sum_{t=1}^{T} \xi_t(j,m)} \tag{49}$$

$$\bar{\Sigma}_{jm} = \frac{\sum_{t=1}^{T} \xi_t(j,m)\left(x_t - \mu_{jm}\right)\left(x_t - \mu_{jm}\right)^*}{\sum_{t=1}^{T} \xi_t(j,m)} \tag{50}$$

$$\bar{c}_{jm} = \frac{\sum_{t=1}^{T} \xi_t(j,m)}{\sum_{k=1}^{M} \sum_{t=1}^{T} \xi_t(j,k)} \tag{51}$$

$$\xi_t(j,m) = \frac{\sum_i \alpha_{t-1}(i) a_{ij} c_{jm} b_{jm}(x_t) \beta_t(j)}{\sum_{i=1}^{N} \alpha_T(i)} \quad (52)$$

2.5 Application de la RAP

Les applications actuelles de la RAP sont nombreuses. Elles couvrent presque tous les secteurs dans lesquels il est souhaitable d'utiliser le signal de parole, des transactions ou toutes sortes d'interactions : La téléphonie, les systèmes de commandes vocales, les systèmes biométriques ou de sécurité sont les différents domaines de cette application. Nous retrouvons les systèmes RAP dans plusieurs axes et domaines, dont les principaux sont :

- l'automatisation de transactions téléphoniques pour l'accès aux services d'information ;
- le contrôle mains-libres des équipements tels que la radio, le conditionnement du système de navigation, le téléphone sans fil, les systèmes télématiques et le contrôle aérien automatique ;
- les logiciels de dictée vocale pour l'apprentissage des langues ;
- l'aide aux personnes handicapées, et la rééducation assistée ;
- le diagnostic assisté par ordinateur, le choix de médicaments, les comptes rendus, et la commande d'appareillages divers ;
- le repérage des indices physiologiques et psychologiques ;
- l'éducation de la voix des malentendants, et les commandes vocales pour les malades immobilisés ;
- le contrôle vocal de machines, l'application pour la gestion de stocks, et la consultation par entrée vocale ;
- la demande de renseignements, de réservation, et la consultation de bases de données ;
- la numérotation téléphonique automatique ;
- l'empreinte vocale pour l'accès en zones réglementées ;
- l'enseignement et la formation des pilotes, la programmation, et l'enseignement assisté par ordinateur.

2.6 Conclusion

Dans ce chapitre, nous avons présenté un système de RAP en employant les HMM. derniers sont incorporés avec le Modèle de Mélange de Gaussiennes (GMM) qui modélise la distribution des états dans la sortie du HMM. En outre, nous avons détaillé les techniques utilisées pour l'apprentissage, l'évaluation et le codage d'un HMM. Dans le chapitre suivant, nous allons montrer l'utilisation de ces techniques pour la construction de notre système SCAESP.

Chapitre 3

Sélection des Erreurs de la Substitution Phonémique et Approches Proposées

3.1 Introduction

Dans ce chapitre, nous présentons la sélection des Erreurs de la Substitution Phonémique (ESP), ainsi que les différentes étapes suivies dans le cadre de notre travail. Ces étapes concernent : l'analyse acoustique des signaux de parole ; l'implémentation des approches d'apprentissage des modèles et la procédure de classification.

3.2 Position du problème

Les enfants ayant un problème des Erreurs de la Substitution Phonémique doivent être pris en charge dans des programmes d'interventions. Ceux-ci peuvent prendre plusieurs séances de rééducation orthophonique avec un effort fourni par les enfants et leurs parents. Toutefois, si le temps ou le nombre d'enfants à rééduquer augmente, cette procédure sera une tâche très difficile. En outre, en Algérie, il n'y a pas suffisamment d'orthophonistes et la majorité d'entre eux s'installent au niveau des grandes villes. Pour prendre un rendez-vous, les parents doivent faire de longues distances et dépenser beaucoup d'argent.

Tous ces problèmes nous ont motivés à développer une nouvelle méthode de traitement pour les ESP. Celle-ci concerne un système d'aide basé sur les techniques de la RAP. Il repose sur le même principe d'un système d'apprentissage de l'arabe parlé pour les étrangers. Une hybridation HMM/GMM est employée pour construire le SCAESP.

L'application du SCAESP peut être utilisée à la maison par l'enfant en présence de son orthophoniste ou de ses parents. Elle est organisée sous forme d'une interface graphique avec plusieurs objets couvrant les ESP sélectionnées, dans plusieurs positions possibles.

3.3 Sélection des phonèmes cibles dans les ESP

Le choix des phonèmes de référence est basé sur l'ensemble des erreurs remarquées durant les différentes sessions d'enregistrement. Cette procédure a été effectuée au niveau d'une école primaire à l'aide d'un groupe de 20 enfants âgés entre 5 et 6 ans. Le tableau 5 donne le matériel utilisé pour sélectionner ces phonèmes.

Après quelques séances d'enregistrement, nous avons remarqué qu'un groupe de 15 enfants (parmi les 20) ont prononcé correctement tous les phonèmes du texte proposé. Cependant, parmi les autres, chaque enfant substitue l'un de ces sons : [s] ; [z] ; [r] ; [ʤ] et

[k] respectivement par [θ] ; [ð] ; [ʁ] ; [ʃ] et [t]. Ces différents cas peuvent être résumés par les oppositions suivantes :

- Alvéolaire / Interdentale [s]/[θ] ;
- Alvéolaire / Interdentale [z]/[ð] ;
- Alvéolaire / Uvulaire [r]/[ʁ] ;
- Palatale / Alvéopalatale [ʤ]/[ʃ] ;
- Vélaire / Alvéolaire [k]/[t].

Tableau 5 : Exemple d'un texte utilisé pour la sélection des phonèmes cibles

إن التفريق بين الأبناء في العطايا والهبات والتمييز بينهم، له مخاطر جسيمة تعود بالسلب عليهم، كالشعور بالظلم وعدم الاهتمام. مما يؤدي إلى العقوق، وقطع الأرحام، وزرع الشحناء والبغضاء، وكذلك نقص الثقة بينهم.
ʔinna ʔattafri:qa bajna ʔalʔabna:ʔi fi ʔalʕatˁa:ja: wa ʔalhiba:ti wa ʔattamji:zi bajnahum lahu maχa:tˁira ʤasi:ma taʕu:du bisilbi ʕalajhum kaʃʃuʕu:ri biðˁðˁulmi wa ʕadami ʔalʔihtima:mi mimma: juʔaddi: ʔila ʔalʕuqu:qi wa qatˁʕi ʔalʔarħa:mi wa zarʕi ʔaʃʃaħna:ʔi wa ʔalbaʁdˁa:ʔi wa kaða:lika naqsˁu ʔaθθiqati bajnahum.

3.3.1 Enregistrement du corpus

Nous avons élaboré un corpus de parole composé de plusieurs mots. Ces derniers comportent les phonèmes étudiés dans différents contextes possibles, suivant la position du phonème cible dans le mot : à la fin du mot et au début du mot suivi respectivement par les voyelles [a], [u] et [i] ([fatħa], [dˁamma] et [kasra]). Le tableau 6 résume les mots utilisés pour chaque opposition.

Tableau 6 : Mots utilisés pour l'élaboration du corpus

Phonèmes		Mot 1	Mot 2	Mot 3	Mot 4
س [s]		سيارة	سلّم	سروال	فانوس
	Correct	[sajja:ra]	[sullam]	[sirwa:l]	[fa:nu:s]
	Incorrect	[θajja:ra]	[θullam]	[θirwa:l]	[fa:nu:θ]
ز [z]		زهرة	زحل	زناد	همزة
	Correct	[zahra]	[zuħal]	[zina:d]	[hamza]
	Incorrect	[ðahra]	[ðuħal]	[ðina:d]	[hamða]
ر [r]		راجل	رمان	رمال	بدر
	Correct	[ra:ʤil]	[rumma:n]	[rima:l]	[badr]
	Incorrect	[ʁa:ʤil]	[ʁumma:n]	[ʁima:l]	[badʁ]
ج [ʤ]		جرة	جنود	جمال	مهرج
	Correct	[ʤarra]	[ʤunu:d]	[ʤima:l]	[muharriʤ]
	Incorrect	[ʃarra]	[ʃunu:d]	[ʃima:l]	[muharriʃ]
ك [k]		كعبة	كناش	كتاب	سمكة
	Correct	[kaʕba]	[kunnãʃ]	[kita:b]	[samaka]
	Incorrect	[taʕba]	[tunna:ʃ]	[tita:b]	[samata]

Le tableau 17 montre la procédure d'enregistrement du corpus. Nous avons segmenté manuellement les fichiers sons enregistrés en utilisant l'outil d'analyse PRAAT (5.1.25) [51]. Ces fichiers sont exploités pour l'extraction des paramètres acoustiques. Le corpus des enregistrements sonores est divisé par la suite en deux sous corpus :

- le premier comportant 1300 fichiers, répartis équitablement entre les différentes oppositions ;
- le second ayant 700 fichiers, concerne la phase de test et de classification.

Tableau 7 : Procédure d'enregistrement du corpus

Nombre d'enfants	50
Age	5-6 ans
Fréquence d'échantillonnage	16 KHz
Quantification	16 bits
N° signaux/opposition	400
N° totale des signaux	2000

Le nombre des signaux enregistrés pour chaque opposition est donné par :

$$4(mots) * 2(classes) * 50(enfants) = 400 \quad (53)$$

D'où le nombre total des signaux enregistrés est de.

$$400 * 5(\text{oppositions}) = 2000 \quad (54)$$

3.3.2 Analyse acoustique du corpus

Chaque phonème se diffère de l'autre par son lieu ou son mode d'articulation. Nous présentons ensuite la configuration de l'appareil phonatoire pendant la prononciation de chacun de ceux-ci. En outre, nous estimons ces paramètres pertinents : la durée, la fréquence fondamentale et l'intensité des phonèmes [52].

3.3.2.1 Opposition Alvéolaire / Interdentale [s]/[θ]

L'opposition alvéolaire / interdentale est connue comme la substitution du [s] par le son [θ]. Elle est connue aussi sous le nom de « stuttering » et apparaît généralement de 18 mois jusqu'à 9 ans. Elle est plus populaire durant l'âge de deux et 3 ans. La cause de cette substitution est l'irrégularité des dents en terme de composition volumétrique ou en terme

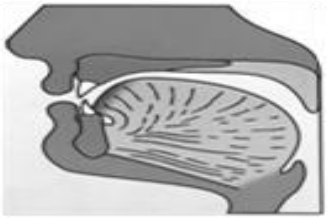
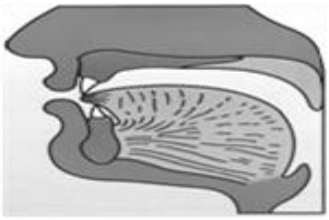
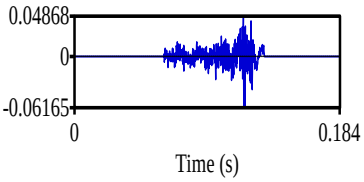
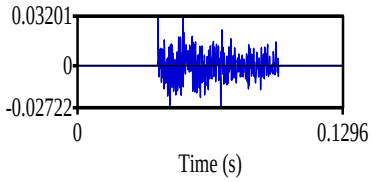
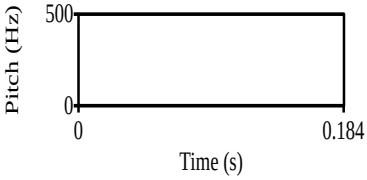
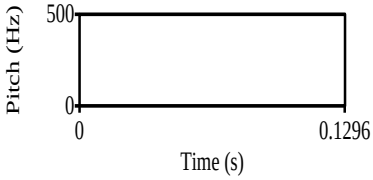
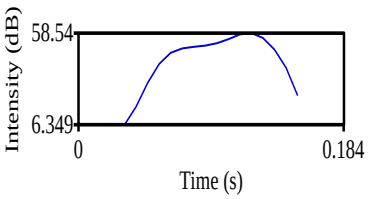
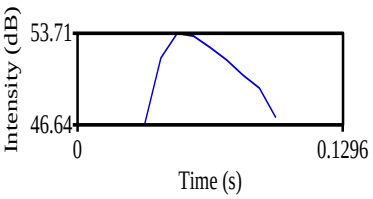
de proximité et de la distance. Les mots utilisés durant une session thérapeutique sont donnés au tableau 8.

Tableau 8 : Mots utilisés pour l'opposition [s]/[θ]

Mots en Arabe	Prononciation	
	Correcte	Incorrecte
سيارة	[sajja:ra]	[θajja:ra]
سلّم	[sullam]	[θullam]
سروال	[sirwa:l]	[θirwa:l]
فانوس	[fa:nu:s]	[fa:nu:θ]

Le phonème [s] est articulé au niveau de la zone alvéolaire. Il se produit, lorsque l'apex se rapproche de l'alvéole des dents d'en haut. Ce phonème est sourd, alvéolaire et fricatif. Cependant, si la langue avance et se place entre les dents supérieures et inférieures en quittant le lieu d'articulation d'origine, nous entendons généralement le [θ] au lieu de [s]. Celui-ci est sourd, inter-dental et fricatif (tableau 9).

Tableau 9 : Articulation et caractéristiques des phonèmes [s] et [θ]

	[s]	[θ]
Articulation		
Durée		
Pitch		
Intensité		

D'après la forme des signaux, les deux sons peuvent être considérés comme bruités. Ils ne présentent aucun pitch, ce qui confirme le non voisement des deux phonèmes. La distinction entre ces deux phonèmes est difficile. La correction d'articulation pour le phonème [s] peut être effectuée en suivant les étapes suivantes :

- mettre la main près de la bouche horizontalement ou verticalement pour sentir le passage de l'air phonatoire lors de la prononciation du phonème et ainsi pour différencier entre la prononciation du phonème [s] et le phonème [z] ;
- prononcer le phonème avec un simple sourire ;
- fermer les mâchoires en laissant un espace étroit pour permettre le passage de l'air entre les deux mâchoires.

3.3.2.2 Opposition Alvéolaire / Interdentale [z]/[ð]

Concernant cette opposition, l'enfant substitue le son [z] par le son [ð]. Le tableau 10 présente quelques exemples pour la prononciation de ce phonème dans différentes positions.

Tableau 10 : Mots utilisés pour l'opposition [z]/[ð]

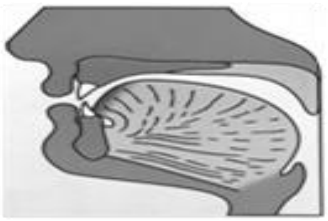
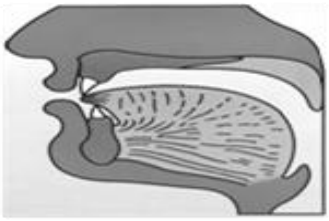
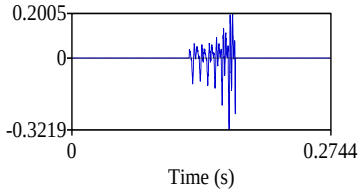
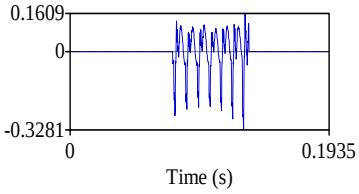
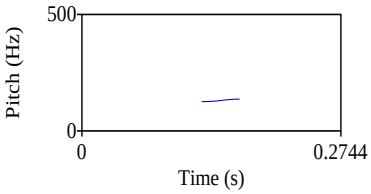
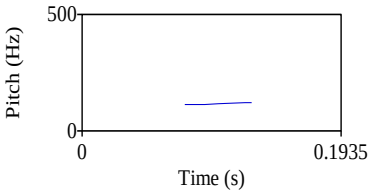
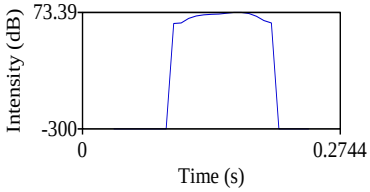
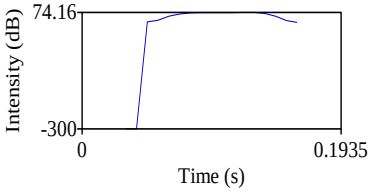
Mots en Arabe	**Prononciation**	
	Correcte	**Incorrecte**
زهرة	[zahra]	[ðahra]
زحل	[zuħal]	[ðuħal]
زناد	[zina:d]	[ðina:d]
همزة	[hamza]	[hamða]

Le son [z] est aussi articulé au niveau de la zone alvéolaire. La différence est basée sur la vibration des cordes vocales durant la prononciation du phonème [z]. Celui-ci est un phonème sonore, alvéolaire et fricatif.

Le déplacement de la langue vers la zone interdentale avec la vibration des cordes vocales fait que le [z] est remplacé par le [ð]. Ce dernier est un phonème sonore, inter-dental et fricatif (tableau 11).

Dans ce cas, la forme des signaux peut être vue comme quasi-périodique. La présence de pitch montre le voisement des deux sons. Par contre, l'énergie pour les deux sons est un peu élevée en la comparant avec le cas précédant.

Tableau 11 : Articulation et caractéristiques des phonèmes [z] et [ð]

	[z]	[ð]
Articulation		
Durée		
Pitch		
Intensité		

Pour corriger la prononciation de ce phonème, on demande à l'enfant de suivre les étapes suivantes :

- prononcer le son avec un simple sourire en essayant de cacher la langue derrière les dents ;
- mettre le doigt entre les dents en essayant de prononcer le [ð] pour entendre le [z];
- simplifier la prononciation avec des mots contenant le phonème [z] au début ;
- augmenter la complexité des exercices pour les différentes positions du [z].

3.3.2.3 Opposition Alvéolaire/Uvulaire [r]/[ʁ]

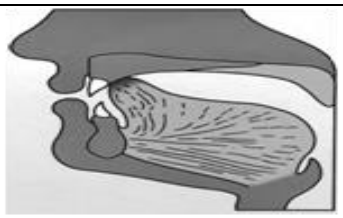
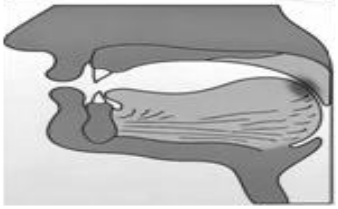
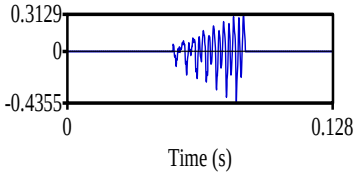
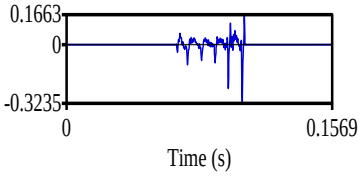
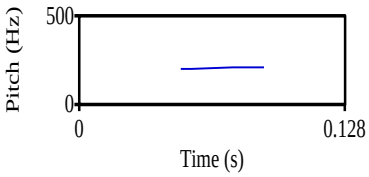
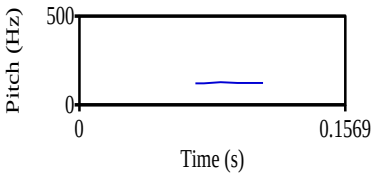
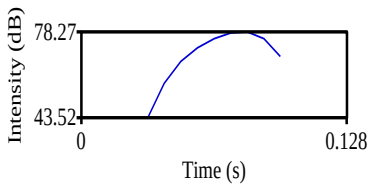
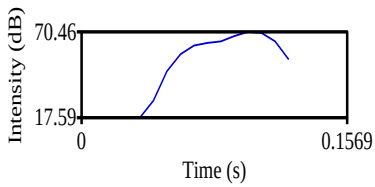
Cette opposition est largement connue durant la période de l'apprentissage de la langue parlée. Différents mots sont donnés (tableau 12), ceux-ci sont utilisés pour les deux phases d'apprentissage et de test.

Tableau 12 : Mots utilisés pour l'opposition [r]/[ʁ]

Mots en Arabe	Prononciation	
	Correcte	Incorrecte
راجل	[ra:ʤil]	[ʁa:ʤil]
رمان	[rumma:n]	[ʁumma:n]
رمال	[rima:l]	[ʁima:l]
بدر	[badr]	[badʁ]

Le phonème [r] est articulé au niveau de l'alvéole des dents supérieures. L'apex vibrant plusieurs fois d'une façon périodique au contact de l'alvéole. Ce phonème est alvéolaire, sonore et vibrant. Si le début de la langue s'abaisse avec un recul de son dos vers la zone vélaire, le [r] est substitué par le son uvulaire [ʁ] (tableau 13).

Tableau 13 : Articulation et caractéristiques des phonèmes [r] et [ʁ]

	[r]	[ʁ]
Articulation		
Durée		
Pitch		
Intensité		

Comme le cas précédant, nous remarquons que la forme des signaux présente un peu de quasi-périodicité. Aussi, la présence du pitch et les valeurs de l'énergie donnent une

information concernant le voisement des deux sons. La correction de la prononciation pour ce phonème peut être obtenue en demandant à l'enfant de suivre les étapes suivantes :

- des exercices de respiration ; une inspiration profonde puis une expulsion d'air de façon directe ;
- mettre la main gauche sur le thorax et la main droite sur la gorge pour sentir les vibrations sonores ;
- utiliser une règle pour déplacer la langue vers le haut et le bas plusieurs fois ;
- choisir un ensemble de mots faciles à prononcer.

3.3.2.4 Opposition Palatale/ Alvéopalatale [ʤ]/[ʃ]

Cette opposition concerne la substitution phonémique du son [ʤ] par le son [ʃ]. Celle-ci est moins fréquente par rapport aux autres oppositions. Les données et le matériel utilisés pour le traitement de celle-ci sont donnés au tableau 14, qui montre les prononciations possibles du son [ʤ] dans quelques positions.

Tableau 14 : Mots utilisés pour l'opposition [ʤ]/[ʃ]

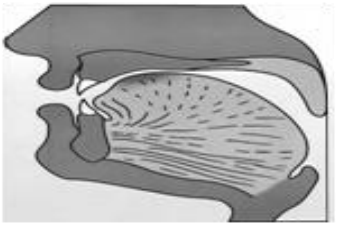
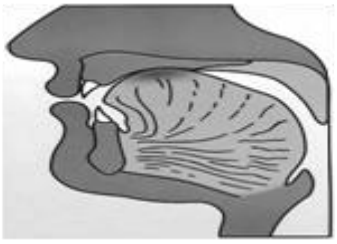
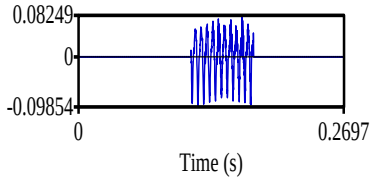
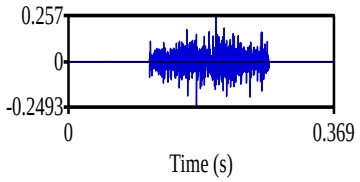
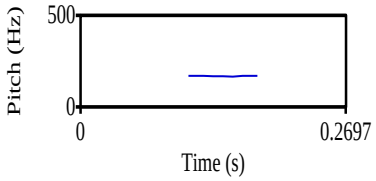
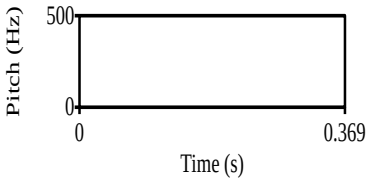
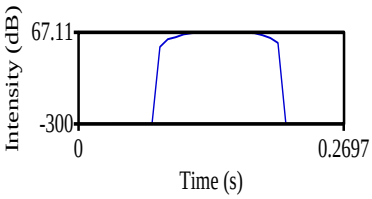
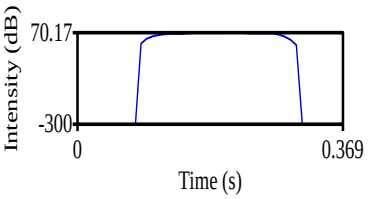
Mots en Arabe	**Prononciation**	
	Correcte	**Incorrecte**
جرة	[ʤarra]	[ʃarra]
جنود	[ʤunu:d]	[ʃunu:d]
جمال	[ʤima:l]	[ʃima:l]
مهرج	[muharriʤ]	[muharriʃ]

Le phonème [ʤ] est articulé au niveau de la zone palatale. Il résulte du passage de l'aire phonatoire dans un espace étroite entre l'apex et le palais, avec une vibration des cordes vocales. Cependant, si les cordes vocales ne vibrent pas, le son [ʤ] est prononcé comme le son [ʃ].

Les deux sons sont articulés au niveau de la zone palatale mais, en générale, le point d'articulation du son [ʤ] est avancé par rapport à celui du son [ʃ] (tableau 15).

La forme du signal pour le phonème [ʤ] montre un peu de quasi-périodicité en comparaison avec le phonème [ʃ] qui peut être vu comme un bruit. En outre, la présence du pitch pour le premier phonème et son absence pour le deuxième montre la possibilité de différencier entre ces deux sons.

Tableau 15 : Articulation et caractéristiques des phonèmes [ʤ] et [ʃ]

	[ʤ]	[ʃ]
Articulation		
Durée		
Pitch		
Intensité		

La correction d'articulation pour ce phonème peut être réalisée en suivant ces étapes :

- prononcer le phonème [ʃ] comme il le faisait habituellement ;
- mettre l'index près des dents supérieures pour bloquer l'air en dehors de la bouche;
- essayer de déplacer la langue un peu en avant ;
- fermer un peu la bouche au cours de la prononciation du phonème.

3.3.2.5 Opposition Vélaire/Alvéolaire [k]/[t]

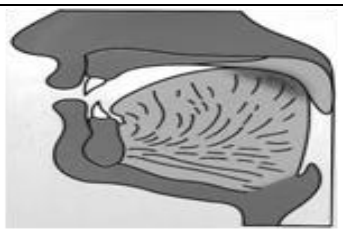
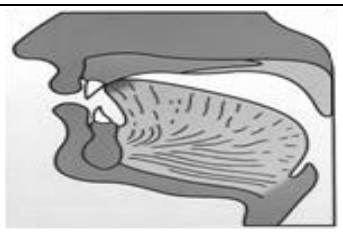
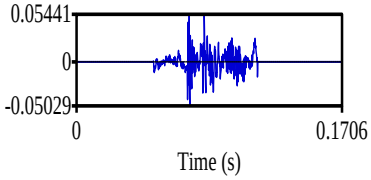
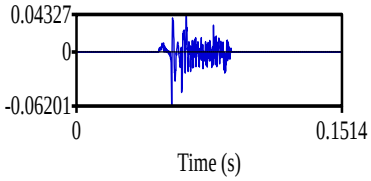
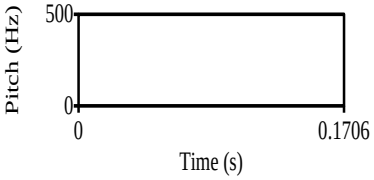
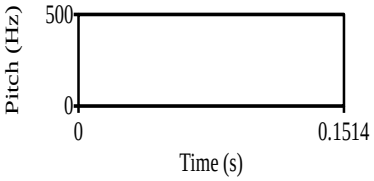
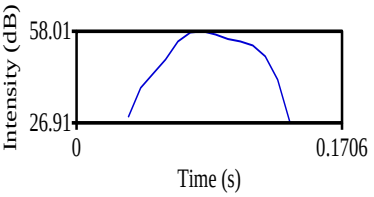
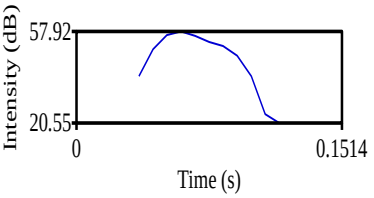
Cette dernière opposition représente la substitution phonémique du phonème [k] par [t]. Cette erreur peut être constatée même chez les adultes. Pour la prononciation de celui-ci, nous avons choisi un ensemble de mots suivant ses emplacements et pour différentes positions (tableau 16).

Tableau 16 : Mots utilisés pour l'opposition [k]/[t]

Mots en Arabe	Prononciation	
	Correcte	Incorrecte
كعبة	[**k**aʕba]	[**t**aʕba]
كناش	[**k**unna:ʃ]	[**t**unna:ʃ]
كتاب	[**k**ita:b]	[**t**ita:b]
سمكة	[samak**a**]	[samat**a**]

Le phonème [k] est un son sourd, vélaire et occlusif. Ce phonème est articulé au niveau de la zone vélaire, sans vibrations des cordes vocales. Si la langue avance et se place au niveau de la zone alvéolaire le son [k] est remplacé par le son alvéolaire [t] (tableau 17).

Tableau 17 : Articulation et caractéristiques des phonèmes [k] et [t]

	[k]	[t]
Articulation		
Durée		
Pitch		
Intensité		

D'après ce tableau, nous remarquons une similitude bien claire dans les caractéristiques des deux sons soit pour la forme des signaux pour le pitch ou pour l'énergie. Cependant, pour le lieu d'articulation, ceux-ci sont totalement différents.

L'enfant peut corriger l'articulation de ce phonème en s'appuyant sur les deux règles suivantes :

- il essaie de prononcer le phonème [t] comme il le faisait ;
- en pressant sur le début de la langue à l'aide d'une règle pour entendre le son [k].

3.4 Architecture du système de classification élaboré

Pour atteindre nos objectifs, nous avons implémenté notre système de classification sous Matlab 7.10 (R2010a) à l'aide du toolbox HMM [53]. L'architecture globale du système est donnée par la figure 12. Dans notre cas, la procédure de classification revient à une décision binaire en acceptant la prononciation comme étant correcte ou incorrecte.

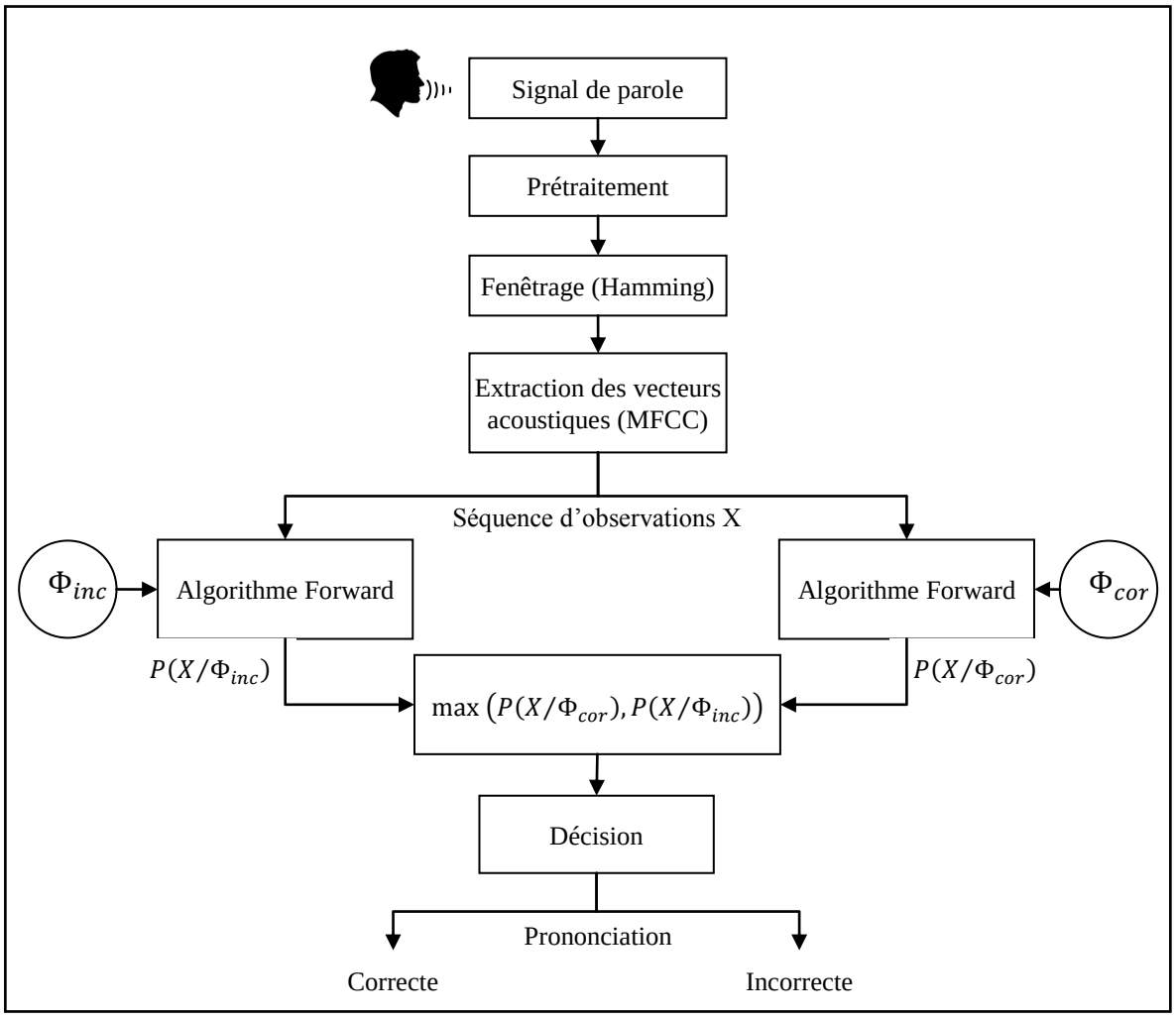


Figure 12 : Schéma fonctionnel de l'approche proposée
Φ_{cor} : Modèle de la prononciation correcte
Φ_{inc} : Modèle de la prononciation incorrecte

3.4.1 Prétraitement des signaux

Le signal de parole capté au moyen d'un microphone est échantillonné à une fréquence de 16 kHz, avec une précision de 16 bits. Nous avons rehaussé les hautes fréquences qui sont souvent atténuées par le module de production de la parole. Le signal de parole enregistré est passé par un filtre numérique de premier ordre de la forme :

$$\mathrm{H(z)} = 1 - 0.95\mathrm{z}^{-1} \tag{55}$$

Le signal préaccentué est segmenté en une succession de trames de $N = 400$ échantillons, chacune (un intervalle de temps de $400 \,/\, 16 = 25\, ms$), dans cet intervalle le segment de parole est considéré comme stationnaire. Ces segments sont extraits avec un recouvrement de 160 échantillons entre les trames, soit à un intervalle de temps de $10ms$. Pour cela, nous avons utilisé la fenêtre de Hamming :

$$H(n) = \begin{cases} 0.54 - 0.46cos(2\pi n/400) & 0 \leq n \leq 399 \\ 0 & ailleurs \end{cases} \tag{56}$$

3.4.2 Extraction des vecteurs acoustiques

L'extraction des vecteurs acoustiques est une étape très importante. Dans notre travail, nous avons utilisé les paramètres MFCC (Mel Frequency Cepstral Coefficients). Les étapes de l'extraction de ces coefficients sont résumées comme suit :

- calcul de la FFT pour chaque trame de 400 échantillons ;
- changement d'échelle pour rendre compte de la perception humaine. Pour cela, nous multiplions chaque trame par un banc de filtres équidistants en échelle Mel. Nous rappelons que la correspondance entre la fréquence en Hz et la fréquence en Mel est donnée par (cf. équation 08) :

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

- enfin, une application de la Transformée en Cosinus Discrète DCT (Discret Cosine Transform) nous permet de convertir le logarithme du spectre Mel dans le domaine temporel ;
- Nous avons pris en compte uniquement les 12 premiers coefficients, en rajoutant les coefficients relatifs à l'énergie, les dérivées premières ΔMFCC et les dérivées secondes ΔΔMFCC.

Cette opération transforme chaque signal de parole en une suite de vecteurs de coefficients acoustiques X. Ces vecteurs d'observations, sont utilisés pour le calcul de la vraisemblance par rapport à chaque classe (Φ_{cor} et Φ_{inc}) en exploitant l'algorithme Forward.

3.4.3 Calcul de la vraisemblance (Algorithme Forward)

Le calcul de la vraisemblance d'une séquence d'observations X par rapport à un HMM (Φ_{cor} et Φ_{inc}) consiste à évaluer la probabilité $P(X/\Phi)$.

Dans notre application, nous avons utilisé l'algorithme Forward (cf. équations 22 et 24).

$$\alpha_t\,(i) = P(x_1, \dots, x_t, s_t = i\ /\ \Phi)$$
$$\alpha_1\,(i) = P(x_1, s_1 = i\ /\ \Phi)$$

Ces deux équations montrent que la relation de récurrence suivante est vérifiée (cf. équations 23 et 25).

$$\alpha_t\,(j) = \left[\sum_{i=1}^{N} \alpha_{t-1}\,(i)\,a_{ij}\right] b_i(x_t)$$
$$P(X/\Phi) = \sum_{i=1}^{N} \alpha_T\,(i)$$

Nous avons programmé l'algorithme Forward sous l'environnement Matlab 7.10, pour calculer la vraisemblance d'une séquence donnée.

pour $i = 1{:}\,N$
  $\alpha_1\,(i) = \pi_i b_i(x_1)$
fin
pour $t = 1 : T - 1$
  pour $j = 1 : N$
    $\alpha_t\,(j) = \left[\sum_{i=1}^{N} \alpha_{t-1}\,(i)\,a_{ij}\right] b_i(x_t)$
  fin
fin
  $P(X/\Phi) = \sum_{i=1}^{N} \alpha_T\,(i)$

Pour prendre la décision concernant la prononciation si elle est correcte ou incorrecte, nous avons calculé deux vraisemblances à partir des observations X, qui correspondent aux prononciations correcte Φ_{cor} et incorrecte Φ_{inc}. Nous montrons par la suite la procédure d'estimation des paramètres de chaque modèle.

$$\text{Décision} = \arg \max_{(\Phi_{cor},\Phi_{inc})} \left(P(X/\Phi_{cor}), P(X/\Phi_{cor})\right) \quad (57)$$

3.4.4 Bases de références

Pour assurer le fonctionnement de notre classifieur, une étape d'apprentissage a été envisagée. Cette opération permet d'ajuster les paramètres du HMM de manière à maximiser la vraisemblance.

$$\Phi^* = \arg \max_{\Phi} P(X/\Phi) \quad (58)$$

Pour cela, Nous avons implémenté l'algorithme de Baum-Welch dans le même environnement (Matlab 7.10). La figure 13 montre l'organigramme d'apprentissage de cet algorithme.

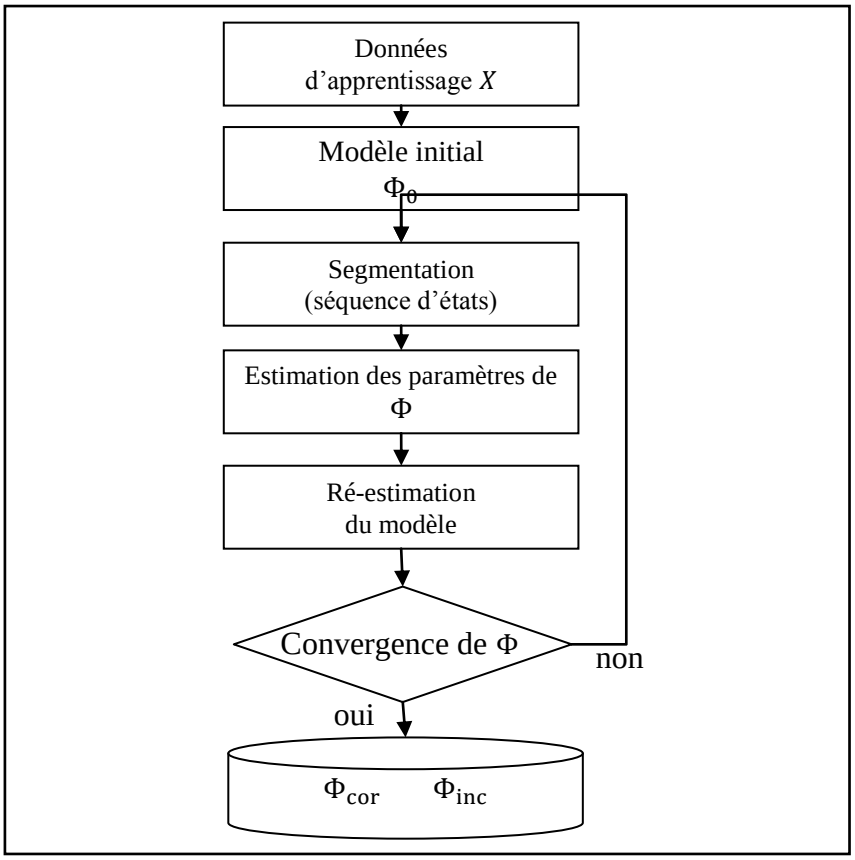


Figure 13 : Apprentissage de Baum-Welch des modèles Φ_{cor} et Φ_{inc}

3.4.4.1 Algorithme de Baum-Welch

Pour chaque mot d'une opposition, nous avons estimé deux HMM correspondant aux prononciations : correcte Φ_{cor} et incorrecte Φ_{inc}. Nous avons appliqué l'algorithme EM

(Expectation Maximisation) à la maximisation des deux probabilités $P(X/\Phi_{\text{cor}})$ et $P(X/\Phi_{\text{inc}})$. Ce processus revient à maximiser $Q(\Phi, \widetilde{\Phi})$ donné par l'équation (40), avec :

- $\Phi = \{A, B, \pi\}$: le nouveau modèle ;
- $\widetilde{\Phi}$: le modèle actuel.

A partir des équations 43 et 44 (page 33), nous avons :

$$a_{ij} = \frac{\sum_{t=1}^{T} \gamma_t(i,j)}{\sum_{t=1}^{T} \sum_{k=1}^{N} \gamma_t(i,k)}$$

$$b_j(k) = \frac{\sum_{t \in X_t = o_k} \sum_i \gamma_t(i,j)}{\sum_{t=1}^{T} \sum_i \gamma_t(i,j)}$$

$$\pi_i = P(s_1 = i/X, \Phi) \qquad (59)$$

Choisir un modèle initial Φ_0
  $t = 0$
Tant que $t < t_{max}$ *et* $P(X/\Phi_t) > P(X/\Phi_{t-1})$
  calculer $\beta_{t-1}(j)$ *et* $\alpha_{t-1}(i)$ *pour le modèle* Φ_{t-1}
  calculer π *de* Φ_t
  calculer A de Φ_t
  calculer B de Φ_t
fin

L'apprentissage des deux modèles de prononciations (correcte, incorrecte) par l'algorithme de Baum-Welch est effectué sous les conditions suivantes :

- le nombre d'états varie de 3, 4, 5 et 6.
- la structure du HMM est de Gauche-Droite.

Nous avons modélisé la sortie dans chaque état du HMM par une distribution Multi Gaussiennes ou GMM. Nous montrons les différentes étapes pour l'estimation de ses paramètres.

3.4.4.2 Apprentissage des modèles GMM

Une densité Multi Gaussienne est caractérisée par trois paramètres principaux, le nombre des composantes gaussiennes, le type de la matrice de covariance et le vecteur de paramètres (cf. équation 45). Ce dernier regroupe le vecteur des poids de mélange "c_{jm}", le vecteur de moyennes "μ_{jm}" et la matrice de covariance "Σ_{jm}". Ces paramètres sont regroupés dans un seul vecteur :

$$\lambda = \{c_{jm}, \mu_{jm}, \Sigma_{jm}\} \quad (60)$$

L'apprentissage du GMM revient à estimer les paramètres du vecteur λ qui donnent la meilleure distribution possible des vecteurs acoustiques dans chaque état. Nous avons implémenté l'algorithme EM (Expectation Maximisation) sous Matlab 7.10., qui nous permet d'estimer les paramètres du vecteur λ suivant le critère de maximum de vraisemblance.

Choisir un modèle initial
λ_0
Tant que $t < t_{max}$ *et* $P(X/\lambda_t) > P(X/\lambda_{t-1})$
calcul de $\xi_t(j,m)$ *% étape d'expectation*
calcul de $\bar{\mu}_{jm}$ *de* λ_t *% étape de maximisation*
calcul de $\bar{\Sigma}_{jm}$ *de* λ_t
calcul de $\bar{c}_{jm}$ *de* λ_t
fin

Les équations 49, 50, 51 et 52, p. 35, donnent, respectivement : $\bar{\mu}_{jm}$, $\bar{\Sigma}_{jm}$, $\bar{c}_{jm}$ et $\xi_t(j,m)$ Nous avons modélisé chaque état par 2, 4, 8 et 16 GMM avec une matrice de covariance de structure complète "$full$". La figure 14 montre l'organigramme d'apprentissage du GMM par l'algorithme EM.

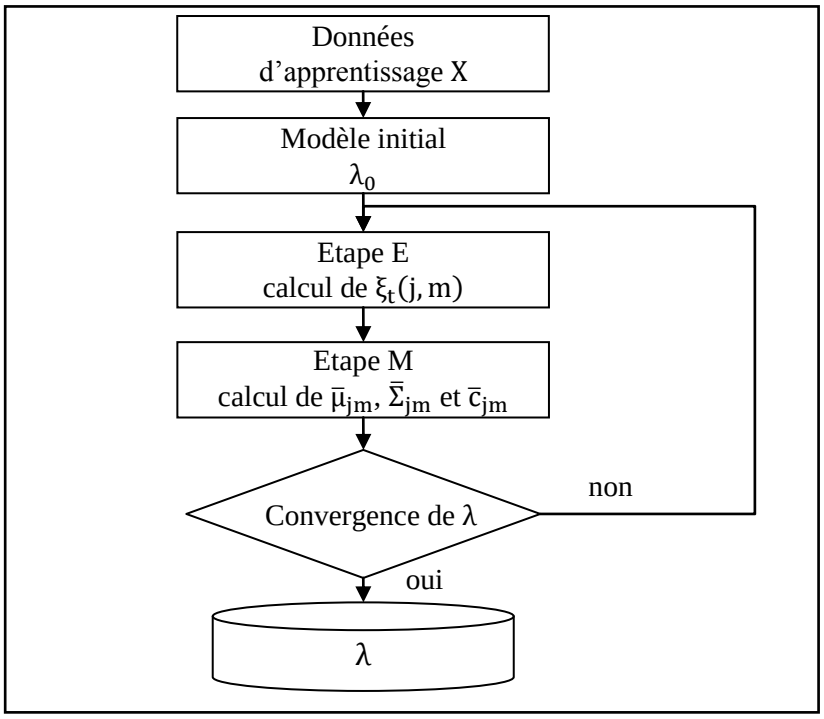


Figure 14 : Apprentissage du GMM par l'algorithme EM

3.4.5 Critères de performance du système

La mesure de l'efficacité du système proposé est effectuée en estimant les erreurs de prononciation. Nous mesurons la Sensibilité, la Spécificité et le Taux de Classification Correcte (TCC), en fonction du nombre des composantes gaussiennes du GMM et le nombre d'états du HMM [54].

Le phonème cible est considéré comme une classe Positive. Par contre, le son substitué est considéré comme une classe Négative.

La sensibilité représente la vraisemblance quand un enfant prononce le phonème cible et le système le classifie comme correct.

La spécificité montre la vraisemblance lorsqu'un enfant substitue le son et le système le classifie comme un son substitué.

$$Sensibilité\ \% = \frac{VP}{VP + FN} * 100 \quad (61)$$

$$spécificité\ \% = \frac{VN}{VN + FP} * 100 \quad (62)$$

$$TCC\ \% = \frac{VP + VN}{VP + FP + VN + FN} * 100 \quad (63)$$

Avec :

- VP : Vraie Positive.
- FP : Fausse Positive.
- VN : Vraie Négative.
- FN : Fausse Négative.

3.5 Conclusion

Dans ce chapitre, nous avons décrit la procédure de sélection des différentes oppositions concernant les Erreurs de la Substitution Phonémique. Nous avons également montré la contribution des HMM à la classification automatique de ces erreurs. Celle-ci est effectuée en implémentant le module d'extraction des MFCC ; l'apprentissage des modèles et la procédure de classification. Dans le dernier chapitre, nous allons montrer les résultats obtenus pour la classification des ESP pour toutes les oppositions étudiées.

Chapitre 4

Application des HMM/GMM à la Classification Automatique des ESP

4.1 Introduction

Dans ce chapitre, nous allons implémenter les différentes techniques décrites aux chapitres précédents et évaluer les performances du système proposé pour la classification automatique des ESP. Nous allons étudier les performances du SCAESP en fonction des nombres d'états des HMM et des composantes gaussiennes GMM. De plus, nous employons le système proposé pour le traitement des différentes oppositions étudiées précédemment.

4.2 Evaluation des performances du SCAESP

La conception du SCAESP est effectuée sous l'environnement Matlab 7.10. en implémentant l'algorithme de Baum-Welch. La représentation du signal vocal se repose sur les MFCC et les Delta-MFCC.

4.2.1 Participants

Cette étude a été réalisée au niveau de trois écoles primaires en Algérie. Un groupe de 50 enfants arabophones, âgés entre 5 et 6 ans, ont participé à la procédure d'enregistrement (tableau 18). Les participants normaux, qui ne montrent pas des ESP, sont utilisés afin de construire un corpus de parole pour la phase d'apprentissage et celle de test. En outre, les patients ont participé aux sessions de rééducation orthophonique à l'aide du SCAESP.

Tableau 18 : Sélection des cibles des ESP

Nombre d'enfants	Age	Sexe	Remarque
34	6 ans	17 M	1 avec ESP
		17 F	2 avec ESP
16	5 ans	8 M	1 avec ESP
		8 F	1 avec ESP

La segmentation des signaux de parole est effectuée en employant l'outil d'analyse PRAAT (5.1.25). Cet outil est un logiciel très souple pour faire l'analyse acoustique et la reconstruction des signaux de parole. Il présente un éventail très vaste de fonctionnalités standards et non-standards, parmi lesquelles l'analyse spectrale, la synthèse articulatoire, la segmentation et les réseaux neuronaux.

4.2.2 Évaluation et résultats expérimentaux

Nous allons étudier les cinq oppositions mentionnées précédemment pour couvrir une large gamme des ESP chez les enfants arabophones. Nous commençons par une segmentation manuelle des corpus de parole à l'aide du logiciel PRAAT. Ensuite, nous faisons une étude comparative visuelle entre les signaux des prononciations possibles suivie par une autre étude basée sur les spectrogrammes. Finalement, nous mesurons les performances de SCAESP pour toutes les oppositions étudiées.

4.2.2.1 Modélisation des mots de l'opposition [s]/[θ]

La figure 15 montre une étape de segmentation des mots [sajja:ra] et [θajja:ra]. Celle-ci est basée sur des paramètres et des connaissances à priori sur les différentes phonèmes de la langue Arabe. Le phonème [s] apparait comme un bruit avec une durée très courte. Ces remarques sont valables pour le phonème [θ].

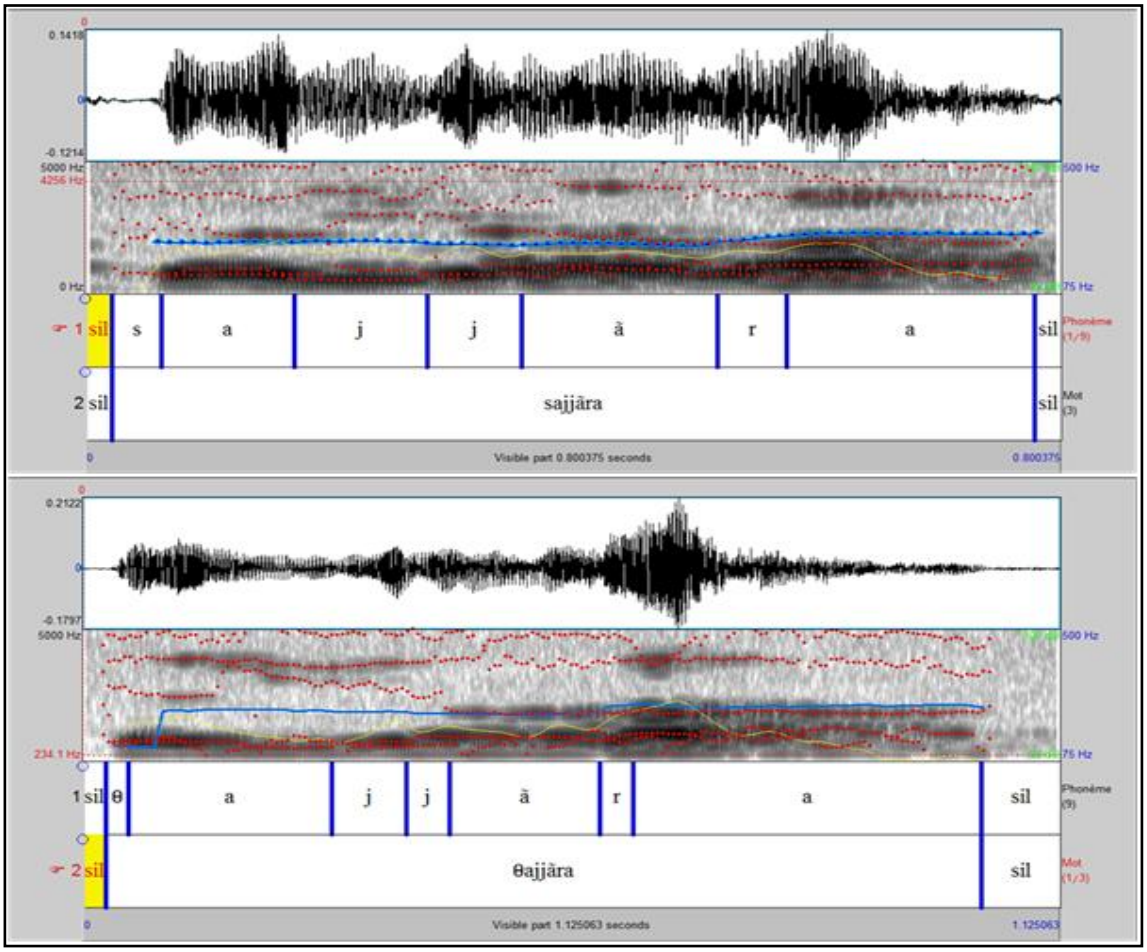


Figure 15 : Segmentation manuelle du mot [sajja:ra]

A titre de comparaison visuelle, nous avons choisi des signaux de parole prononcés par des locuteurs masculins pour les quatre mots de l'opposition [s]/[θ]. Nous remarquons la difficulté de décider visuellement sur la substitution phonémique (figure 16). Pour cela, l'utilisation de méthodes se basant sur les techniques d'analyse et de classification plus avancées est impérative.

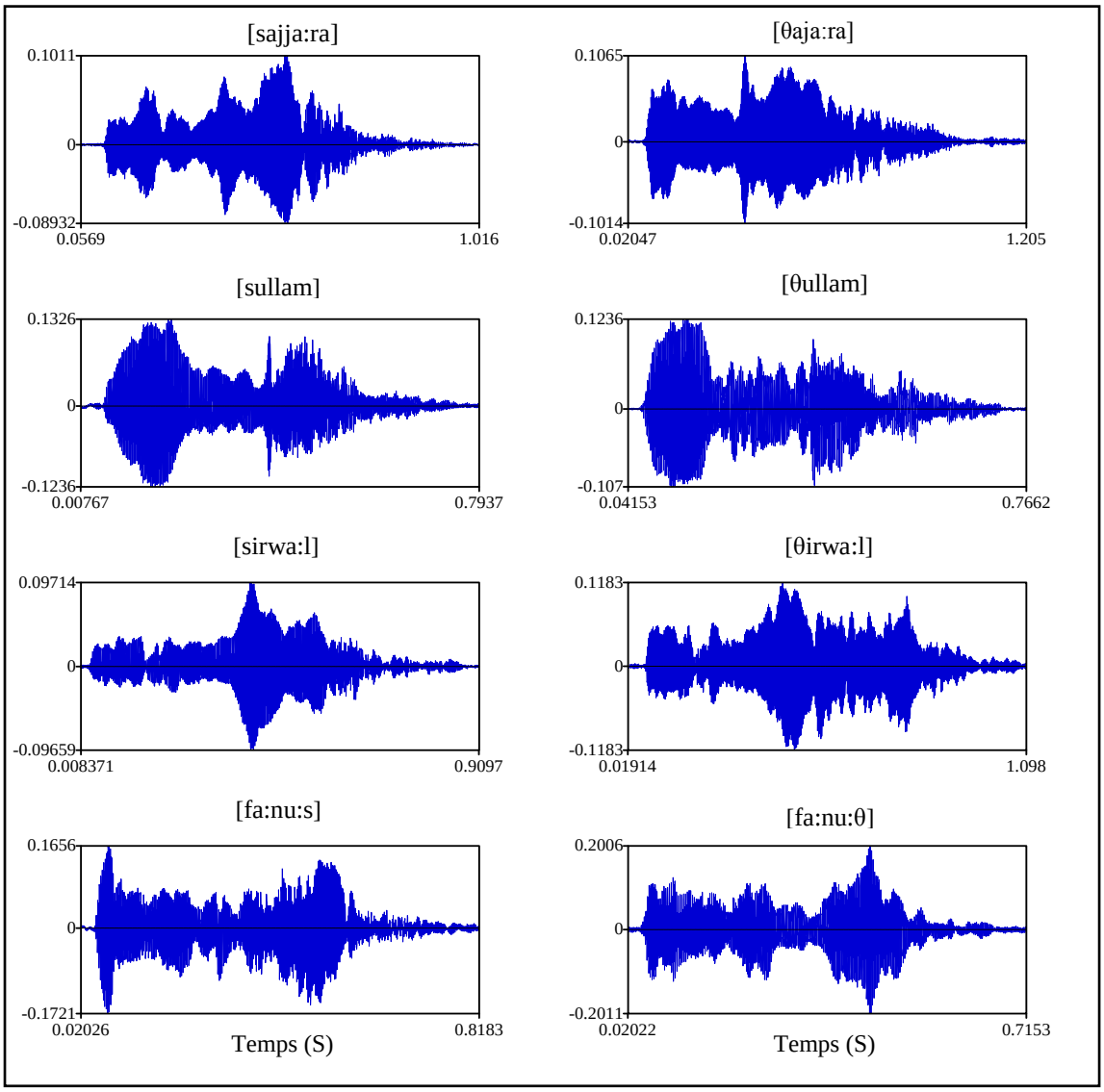


Figure 16 : Comparaisons visuelles des prononciations (opposition [s]/[θ])

Le spectrogramme permet de visualiser les paramètres pertinents de la parole ainsi que l'évolution formantique et donne une information concernant la prononciation. Ceci peut être remarqué par une analyse visuelle comparative des fréquences propres à chaque phonème. Généralement, nous ne pouvons pas garantir la présence d'une substitution

phonémique lors de la prononciation du mot demandé en exploitant seulement le spectrogramme (figure 17).

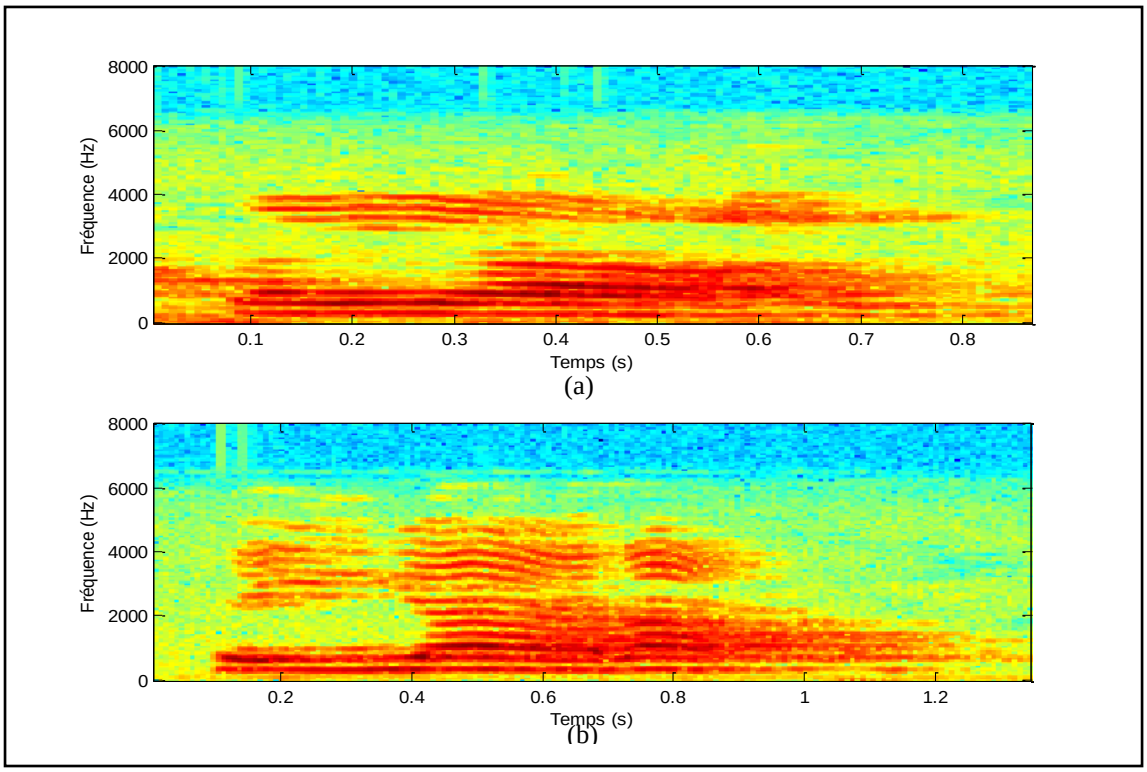


Figure 17 : Spectrogramme du mot [sajja:ra] (a) prononciation correcte (b) prononciation incorrecte

Dans cette expérience, nous mesurons les performances du système proposé en fonction de l'ordre du modèle pour l'opposition [s]/[θ]. L'ordre du modèle varie entre 2, 4, 8 et 16. Ces performances sont étudiées en employant les MFCC et les Delta-MFCC (tableau 19).

Tableau 19 : Performances du système en fonction de l'ordre du modèle (opposition [s]/[θ])

Paramètres	**Performances**	**Ordre du modèle**			
		2	**4**	**8**	**16**
MFCC	Sensibilité	78.04%	78.50%	76.74%	81.54%
	Spécificité	79.18%	81.59%	81.64%	80.12%
	TCC	78.01%	79.55%	78.79%	80.66%
Delta-MFCC	Sensibilité	79.25%	78.27%	84.69%	81.16%
	Spécificité	78.01%	87.24%	84.24%	84.20%
	TCC	78.41%	82.19%	84.15%	82.58%

Pour cette opposition, nous mesurons aussi les performances du système en fonction du nombre d'états pour 3, 4, 5 et 6 états. L'ordre du modèle est fixé à 16 en employant les MFCC et les Delta-MFCC (tableau 20).

Tableau 20 : Performances du système en fonction du nombre d'états (opposition [s]/[θ])

Paramètres	Performances	Nombre d'états 3	4	5	6
MFCC	Sensibilité	80.09%	80.64%	79.35%	81.98%
	Spécificité	80.77%	83.24%	80.97%	82.13%
	TCC	80.33%	81.85%	79.98%	81.86%
Delta-MFCC	Sensibilité	79.82%	81.03%	80.08%	82.02%
	Spécificité	84.83%	86.25%	86.66%	84.58%
	TCC	81.43%	83.36%	83.01%	82.96%

Pour l'opposition [s]/[θ], nous remarquons que le meilleur taux de 84.15%, est obtenu pour un modèle à 8 Composantes Gaussiennes (CG). La spécificité et la sensibilité sont respectivement 87.24% pour 4 CG et 84.69% pour 8 CG. Le nombre d'état optimal pour cette opposition est de 4 avec un TCC de 83.36%. Généralement, pour cette opposition un HMM de 4 états et de 8 CG avec les Delta-MFCC est nécessaire.

4.2.2.2 Modélisation des mots de l'opposition [z]/[ð]

La figure 18 montre une étape de segmentation des mots [zuħal] et [ðuħal]. Les deux phonèmes [z] et [ð] sont sonores (voisés), donc leurs extractions se reposent sur la présence du pitch, et de l'énergie qui peut être un peu élevée par rapport à un son sourd.

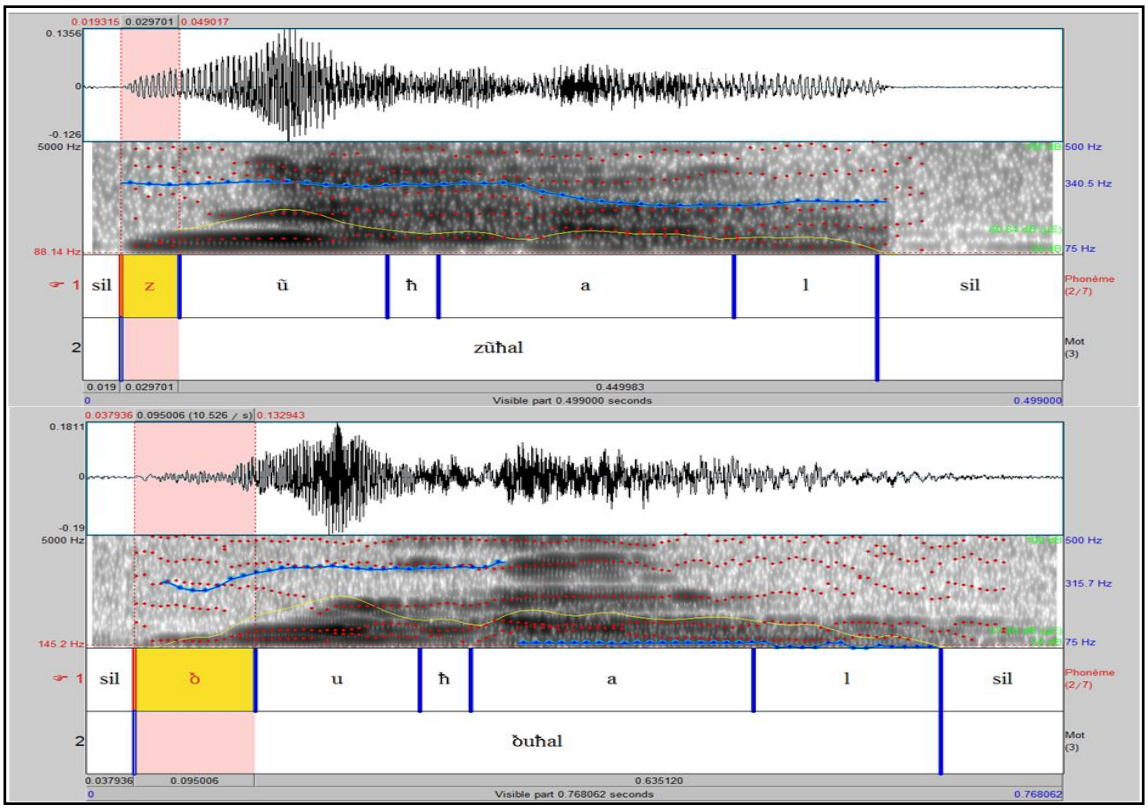


Figure 18 : Segmentation du mot [zuħal]

Nous avons illustré les signaux de parole des quatre mots pour l'opposition [z]/[ð]. Généralement, la forme des différents signaux ne donne aucune information concernant une substitution phonémique du [z] par le [ð]. En outre, la présence du bruit dans certains signaux rend cette opération très compliquée (figure 19).

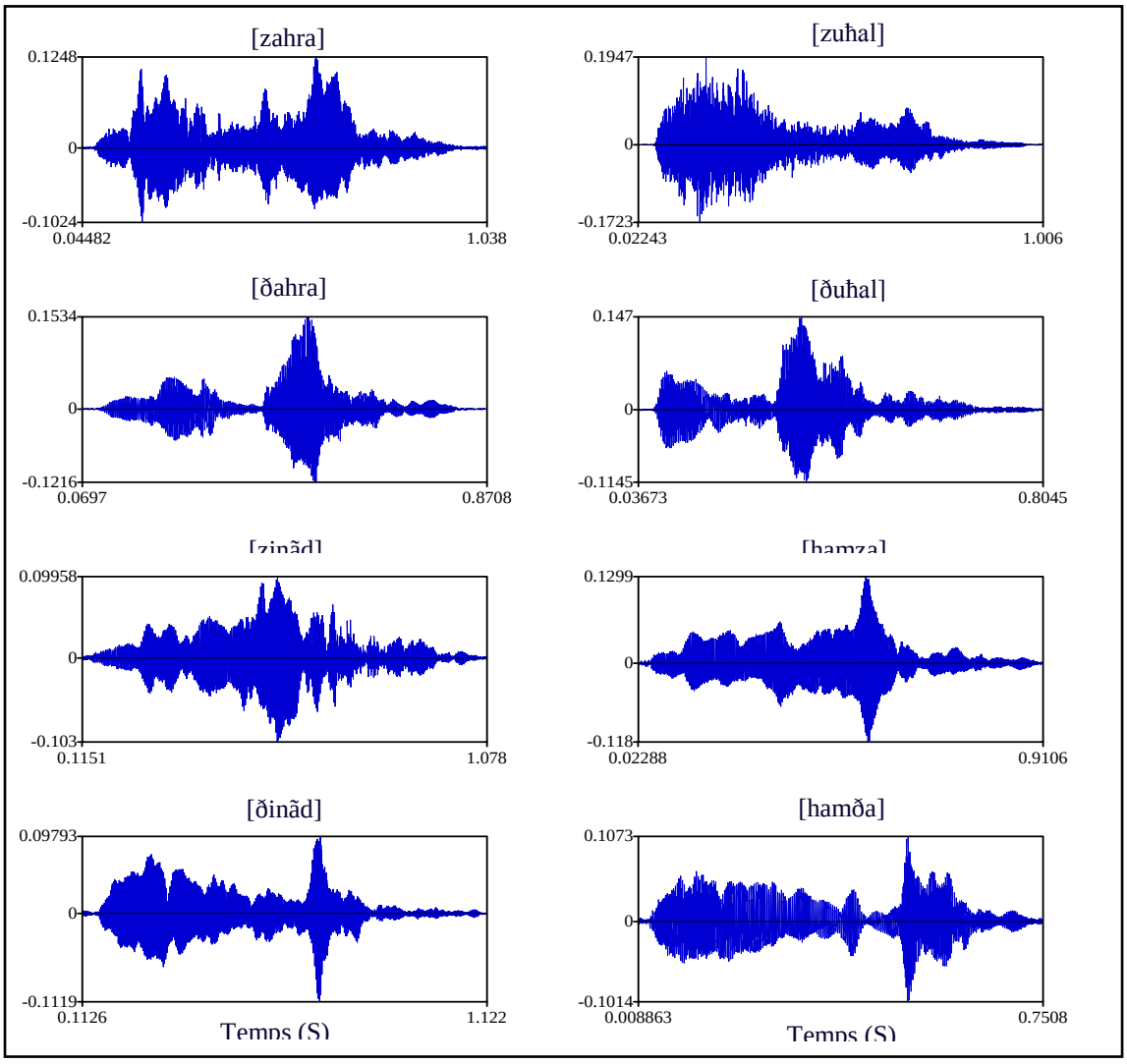


Figure 19 : Comparaisons visuelles des prononciations (opposition [z]/[ð])

Les spectrogrammes du mot [zahra] (figure 20), pour les deux prononciations, ne présentent pas un changement remarquable qui nous aide à juger la présence d'une substitution phonémique. En outre, même si nous voyons l'évolution formantique pour les deux énoncés, les deux mots sont bien prononcés. Ainsi, les informations données par une comparaison visuelle des fréquences propres à chaque phonème, restent incomplètes pour détecter la présence d'une substitution phonémique.

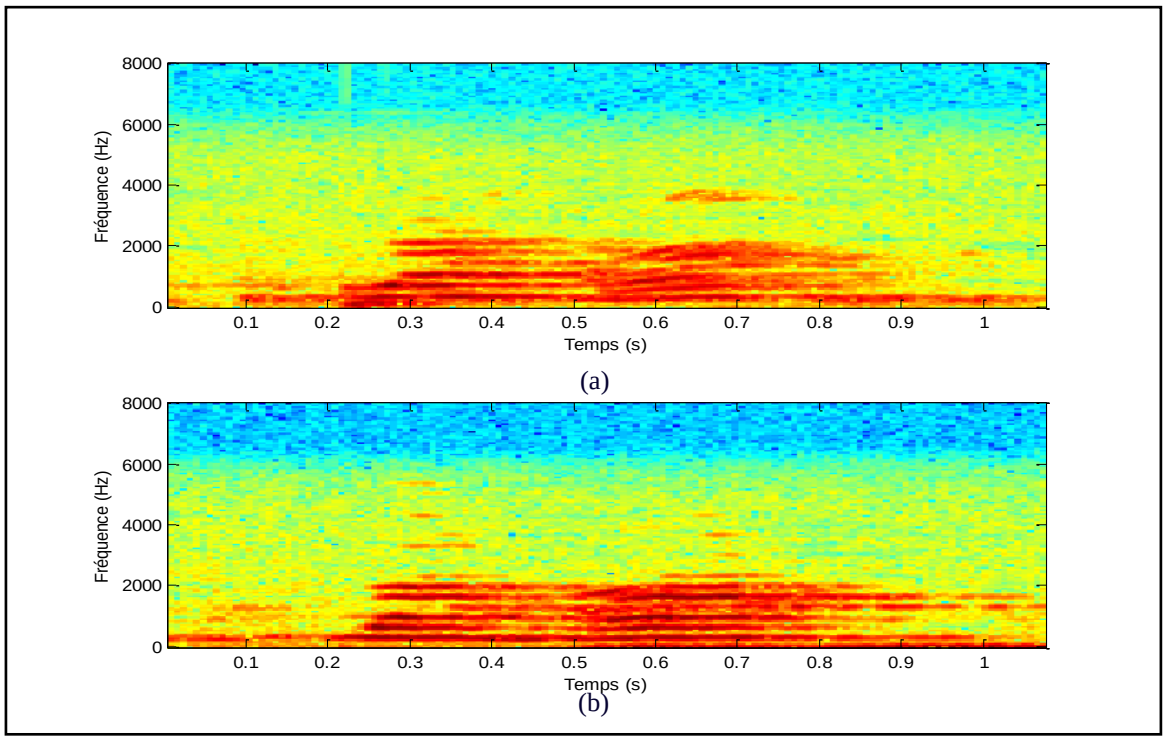


Figure 20 : Spectrogramme du mot [zahra] (a) prononciation correcte (b) prononciation incorrecte

Pour estimer les meilleurs paramètres du système HMM/GMM pour la deuxième opposition, nous mesurons la sensitivité, la spécificité et le TCC pour tous les mots de cette opposition en fonction de l'ordre du modèle (tableau 21) et en fonction du nombre d'états (tableau 22)

Tableau 21 : Performances du système en fonction de l'ordre du modèle (opposition [z]/[ð])

		Ordre du modèle			
Paramètres	**Performances**	**2**	**4**	**8**	**16**
MFCC	Sensibilité	72.04%	77.56%	79.87%	78.25%
	Spécificité	77.15%	79.70%	75.86%	78.77%
	TCC	74.34%	77.69%	77.33%	77.62%
Delta-MFCC	Sensibilité	75.78%	79.03%	78.31%	81.73%
	Spécificité	78.40%	82.73%	80.73%	80.66%
	TCC	76.94%	79.94%	79.17%	79.92%

Tableau 22 : Performances du système en fonction du nombre d'états (opposition [z]/[ð])

		Nombre d'états			
Paramètres	**Performances**	**3**	**4**	**5**	**6**
MFCC	Sensibilité	74.97%	79.24%	78.18%	75.12%
	Spécificité	78.00%	78.06%	77.46%	77.75%
	TCC	76.08%	78.37%	77.26%	76.14%
Delta-MFCC	Sensibilité	79.71%	81.66%	79.40%	81.41%
	Spécificité	81.30%	80.16%	80.10%	79.02%
	TCC	80.23%	80.27%	79.14%	79.51%

Nous remarquons qu'un modèle de 4 CG donne un TCC maximal de 79.94%, avec les Delta-MFCC. Ce modèle montre aussi une meilleure spécificité de 82.73%. Par contre, pour la sensibilité, le modèle nécessite 16 CG pour une valeur optimale de 81.73%.

Pour le nombre d'états, un HMM de 4 états montre un TCC et une sensibilité maximale, respectivement, de 80.27% et 81.66%. Cependant, la meilleure spécificité de 81.30% est donnée par un HMM de 3 états. L'utilisation des Delta-MFCC améliore les performances du système aussi bien pour l'ordre du modèle que pour le nombre d'états.

4.2.2.3 Modélisation des mots de l'opposition [r]/[ʁ]

La segmentation des mots [ra:ʤil] et [ʁa:ʤil] est illustrée à la figure 21. Les sons [r] et [ʁ] sont sonores et cela peut être montré par la présence du pitch au cours des deux prononciations. La durée du son [r] et celle du son [ʁ] sont très courtes par rapport à la voyelle [a]. Cependant, les énergies des deux sons sont un peu élevées.

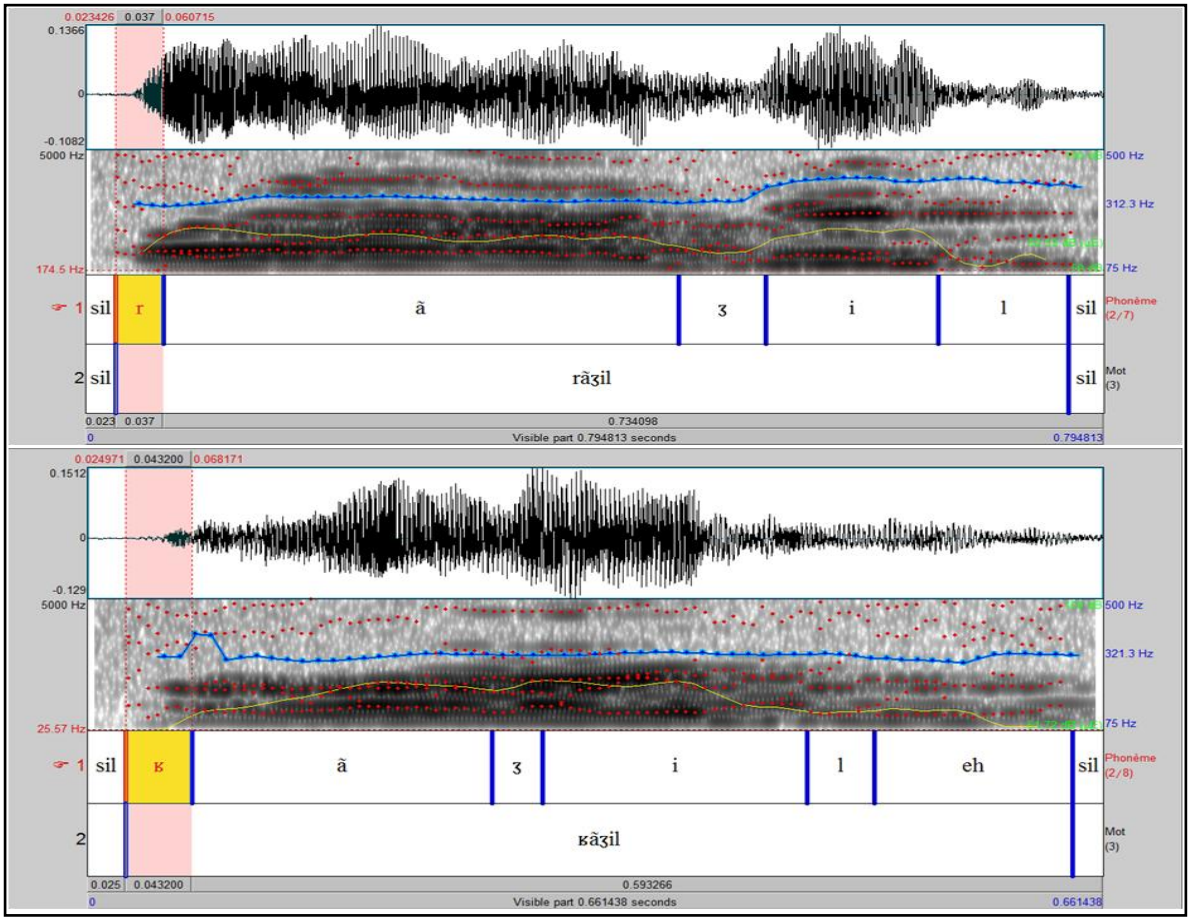


Figure 21 : Segmentation du mot [ra:ʤil]

Pour tous les cas, nous remarquons une différence totale entre les deux prononciations (correcte et incorrecte). Par conséquent, cette différence est observée pour tous les phonèmes et non pas pour les sons cibles seulement, ce qui rend impossible la décision sur la substitution phonémique visuellement (figure 22).

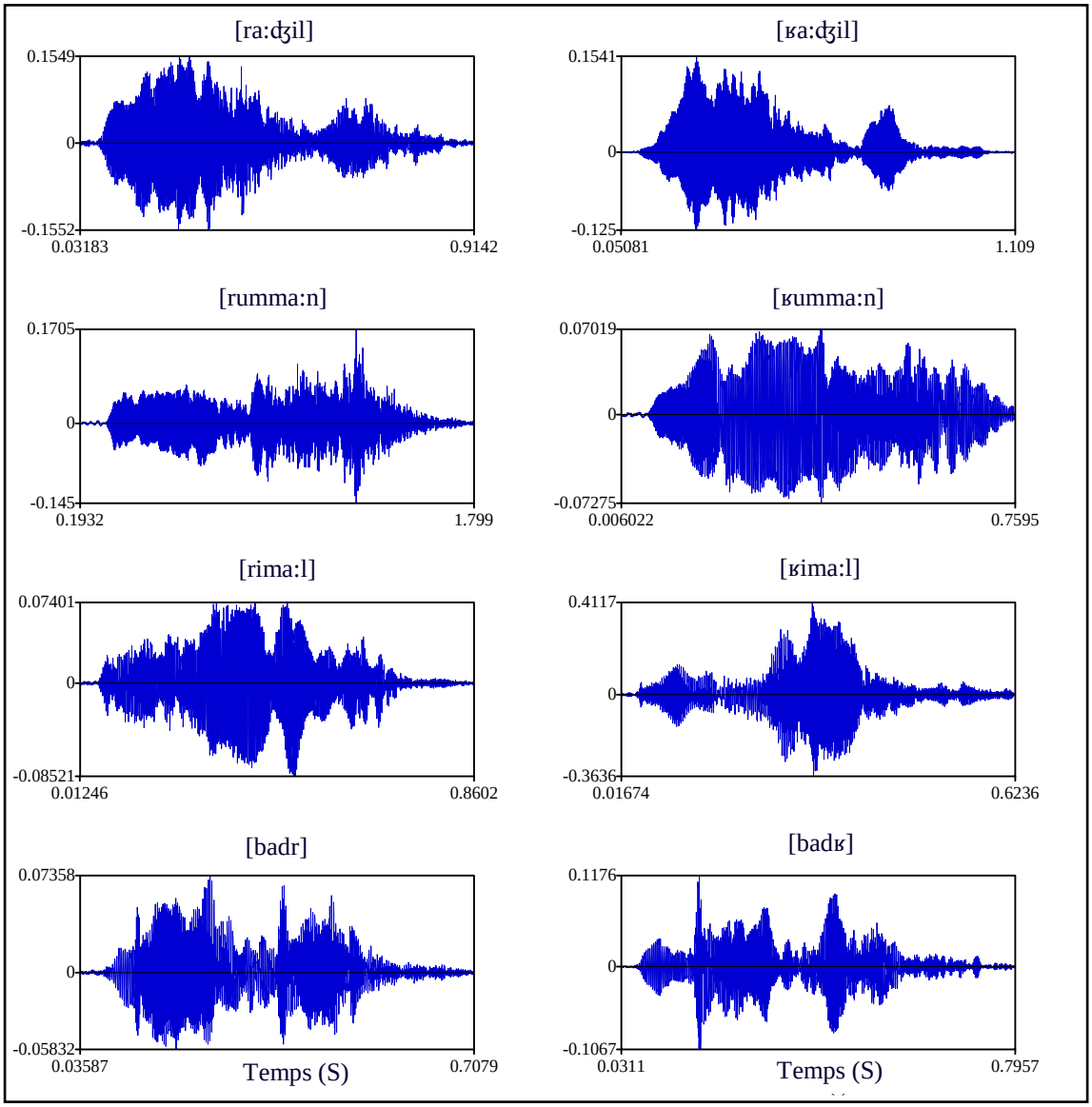


Figure 22 : Comparaisons visuelles des prononciations (opposition [r]/[ʁ])

La figure 23 illustre les spectrogrammes des deux signaux de parole (prononciations correcte et incorrecte) du mot [ra:ʤil]. D'après celles-ci, nous ne pouvons pas assurer la présence d'une substitution phonémique. En outre, nous ne pouvons pas décider à quel niveau un son a été substitué par un autre.

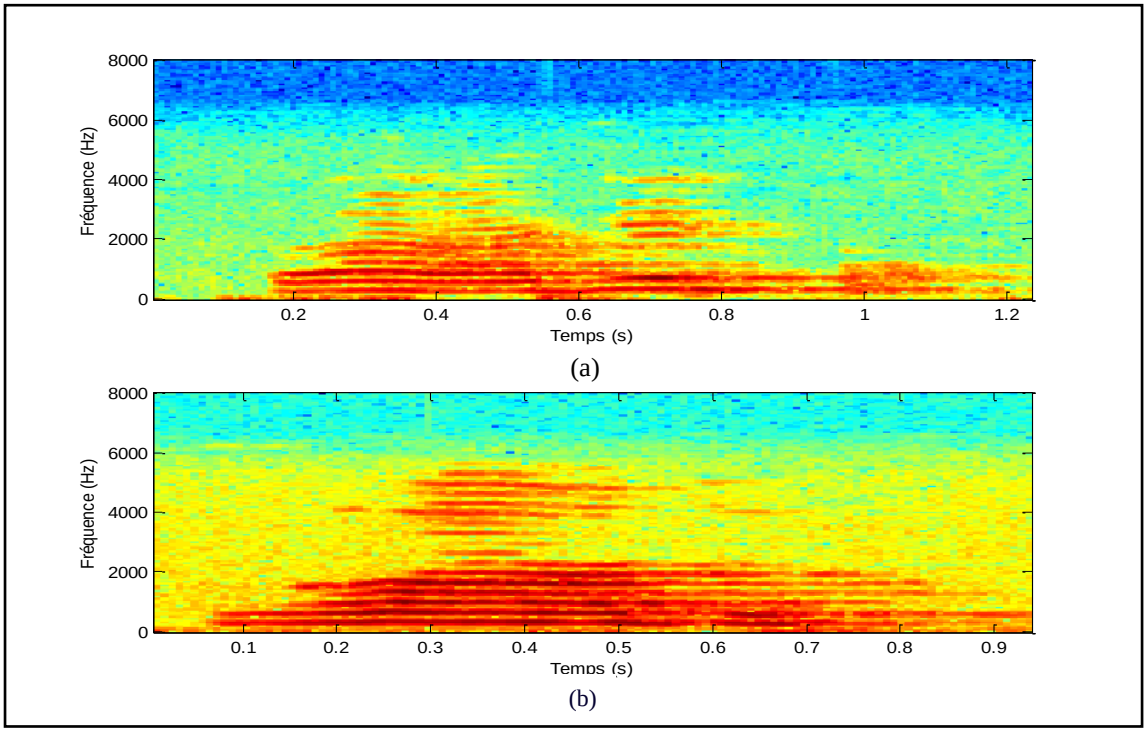


Figure 23 : Spectrogramme du mot [ra:ʤil] (a) prononciation correcte (b) prononciation incorrecte

Pour chaque position du phonème pour cette opposition, nous calculons les trois paramètres des performances en fonction de l'ordre du modèle, en utilisant 2, 4, 8 et 16 CG (tableau 23). Aussi, nous calculons également ces performances en fonction du nombre d'états, en utilisant 3, 4, 5 et 6 états (tableau 24).

Tableau 23 : Performances du système en fonction de l'ordre du modèle (opposition [r]/[ʁ])

		Ordre du modèle			
Paramètres	**Performances**	**2**	**4**	**8**	**16**
MFCC	Sensibilité	79.99%	80.83%	82.32%	82.26%
	Spécificité	79.59%	78.58%	83.99%	81.59%
	TCC	79.60%	79.67%	83.05%	81.65%
DDMFCC	Sensibilité	82.89%	80.57%	86.53%	84.88%
	Spécificité	80.03%	84.42%	83.98%	82.62%
	TCC	81.09%	82.02%	85.09%	83.51%

Tableau 24 : Performances du système en fonction du nombre d'états (opposition [r]/[ʁ])

		Nombre d'états			
Paramètres	**Performances**	**3**	**4**	**5**	**6**
MFCC	Sensibilité	81.67%	85.74%	82.32%	84.41%
	Spécificité	81.49%	82.88%	81.90%	79.84%
	TCC	81.20%	84.20%	82.04%	82.00%
Delta-MFCC	Sensibilité	86.53%	82.92%	82.07%	85.48%
	Spécificité	80.17%	84.75%	87.22%	85.80%
	TCC	83.15%	83.62%	84.33%	85.57%

D'après le tableau 23, nous remarquons que 8 Composantes Gaussiennes sont nécessaires pour obtenir des performances adéquates avec cette opposition. Cela est observé pour les deux types de paramétrisation, MFCC et Delta-MFCC. Le taux, la spécificité et la sensibilité optimaux sont respectivement, 85.09%, 84.42% et 86.53%.

Ces résultats montrent qu'un nombre d'états inférieur à 6 est insuffisant pour modéliser les mots de cette opposition. Pour un HMM de 6 états, le TCC atteint un pourcentage de 85.57%. Cependant, la meilleure spécificité est donnée seulement avec un HMM de 3 états et la meilleur sensibilité avec un HMM de 5 états.

4.2.2.4 Modélisation des mots de l'opposition [ʤ]/[ʃ]

Dans cette section, nous étudions la substitution phonémique du son [ʤ] par le son [ʃ]. L'extraction de ces deux sons est obtenue grâce à une opération de segmentation des mots de corpus. La figure 24 montre l'une de cette étape pour le mot [ʤima:l]. La distinction entre les phonèmes [ʤ] et [ʃ] est un peu facile en comparant avec les cas précédents. Ceci est traduit par la présence du pitch pour le [ʤ] et son absence pour le [ʃ].

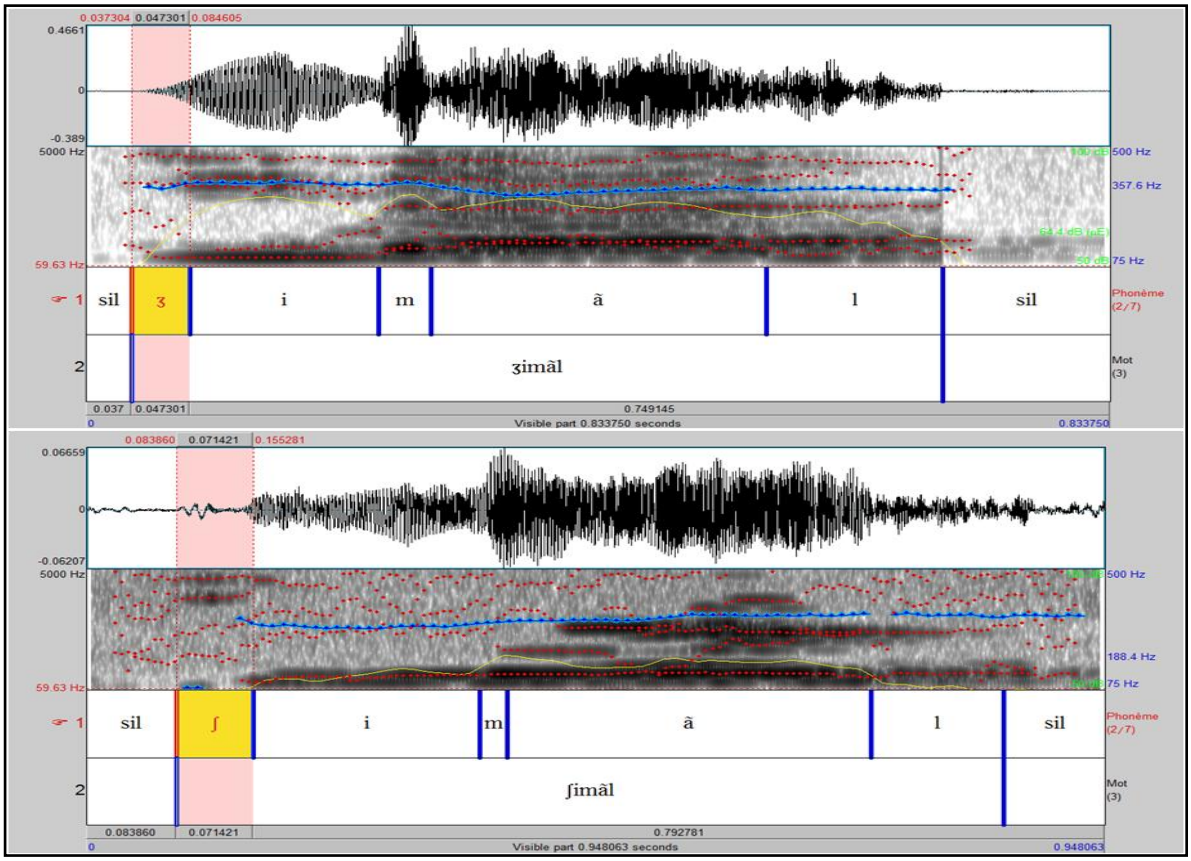


Figure 24 : Segmentation du mot [ʤima:l]

Nous avons aussi fait une comparaison visuelle pour prendre une décision concernant la présence d'une substitution phonémique ou non. Le son [ʤ] se diffère du son [ʃ] par la présence du voisement. Cependant, et d'après la figure 25, la différence n'est pas visible. Tous les signaux sont différents deux à deux, mais cette comparaison ne donne aucune information sur la nature de cette différence.

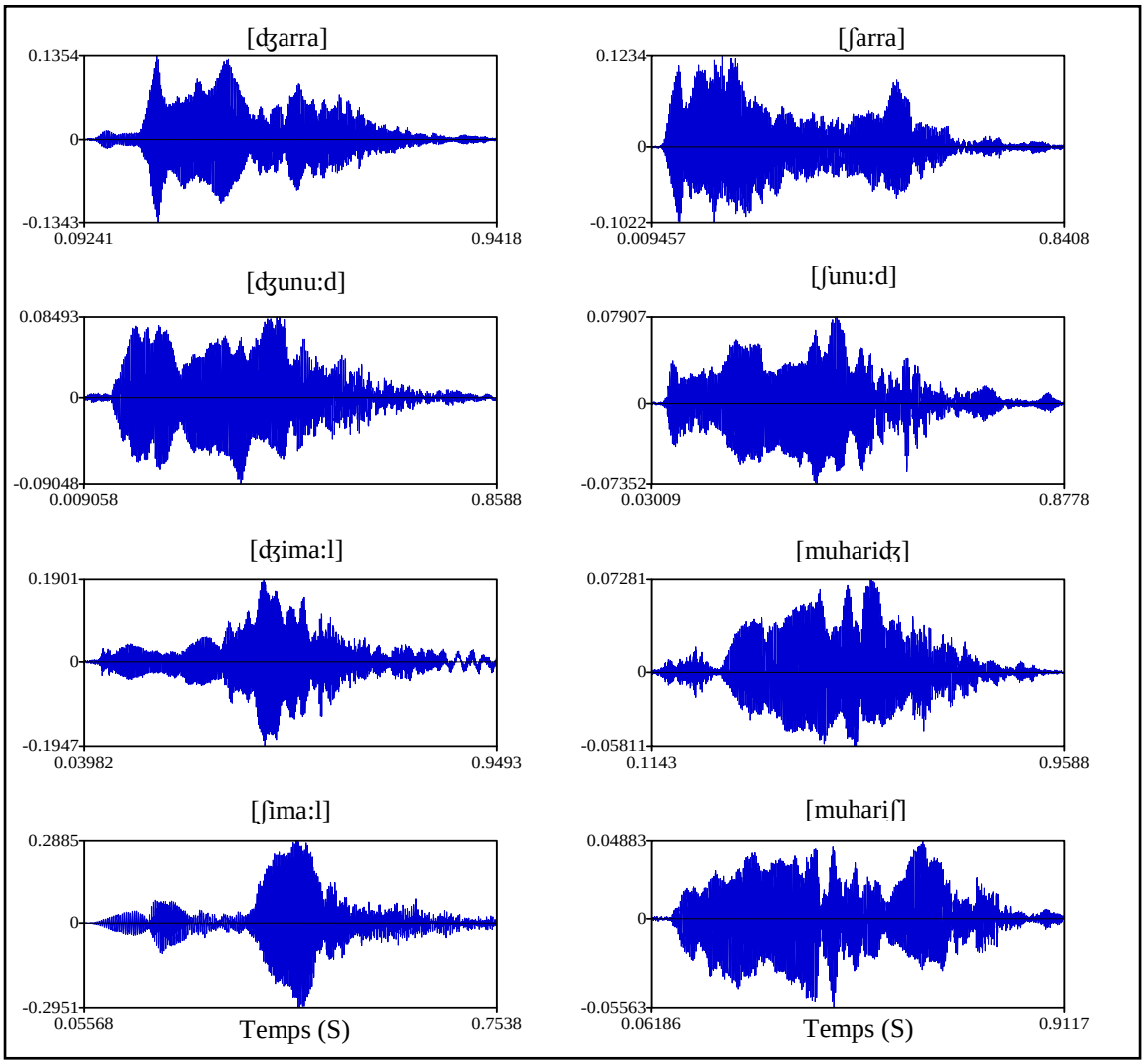


Figure 25 : Comparaisons visuelles des prononciations (opposition [ʤ]/[ʃ])

Après la comparaison visuelle entre ces différents signaux, nous avons un autre type de comparaison qui repose sur les spectrogrammes du mot [ʤarra] pour les deux prononciations possibles. La figure 26 montre l'apparition de certains formants supplémentaires pour la prononciation incorrecte. Ce qui donne une information concernant la présence d'une substitution phonémique. Cependant, on ne peut pas s'assurer quel est le phonème qui a été substitué lors de la prononciation de ce mot.

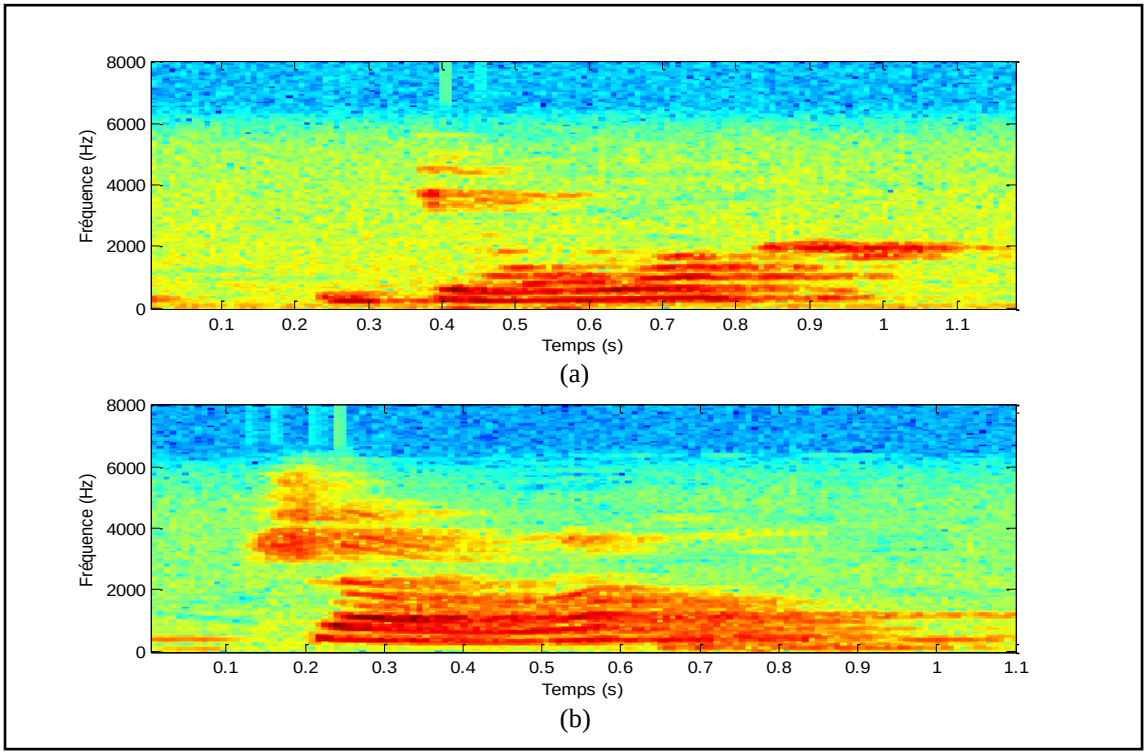


Figure 26 : Spectrogramme du mot [ʤarra] (a) prononciation correcte (b) prononciation incorrecte

Les performances du système proposé pour l'opposition [ʤ]/[ʃ] sont étudiées en fonction de l'ordre du modèle et le nombre d'états. Les résultats obtenus sont résumés, respectivement, dans les tableaux 25 et 26.

Tableau 25 : Performances du système en fonction de l'ordre du modèle (opposition [ʤ]/[ʃ])

		Ordre du modèle			
Paramètres	**Performances**	**2**	**4**	**8**	**16**
MFCC	Sensibilité	84.62%	87.18%	85.43%	86.44%
	Spécificité	81.97%	81.15%	88.86%	84.90%
	TCC	82.60%	84.00%	86.34%	85.18%
Delta-MFCC	Sensibilité	87.07%	88.31%	89.21%	90.04%
	Spécificité	85.11%	84.61%	86.40%	91.46%
	TCC	86.00%	85.99%	87.44%	90.10%

Tableau 26 : Performances du système en fonction du nombre d'états (opposition [ʤ]/[ʃ])

		Nombre d'états			
Paramètres	**Performances**	**3**	**4**	**5**	**6**
MFCC	Sensibilité	85.38%	84.67%	88.46%	87.30%
	Spécificité	84.48%	86.61%	83.67%	83.96%
	TCC	84.86%	85.23%	85.69%	85.25%
Delta-MFCC	Sensibilité	86.27%	88.62%	85.31%	89.36%
	Spécificité	86.21%	89.51%	87.17%	89.44%
	TCC	85.96%	88.17%	85.99%	89.03%

Le système montre un bon TCC avec un modèle de 16 Composantes Gaussiennes et les paramètres Delta-MFCC, avec un taux de 90.10%. En outre, le même modèle donne de meilleures performances pour la spécificité (91.46%) et la sensibilité (90.04%).

Ce dernier tableau présente les performances du système proposé en fonction du nombre d'états. D'après ces résultats, nous remarquons que le taux optimal est de 89.03%. Ce taux est obtenu par un HMM de 6 états où la spécificité est de 89.44%. Par contre, la sensibilité maximale (88.62%) est donnée par un HMM de 4 états. Généralement, les Delta-MFCC améliorent ces performances par rapport aux MFCC pris seuls.

4.2.2.5 Modélisation des mots de l'opposition [k]/[t]

Nous avons fait une segmentation manuelle des différents mots de cette opposition à l'aide du logiciel PRAAT. La figure 27 montre cette opération pour le mot [samaka]. Les deux sons sont sourds, donc la présence du pitch lors de la prononciation du son [t] est causée par sa coarticulation avec la voyelle [a], et sa durée qui est très courte par rapport à celle de cette voyelle.

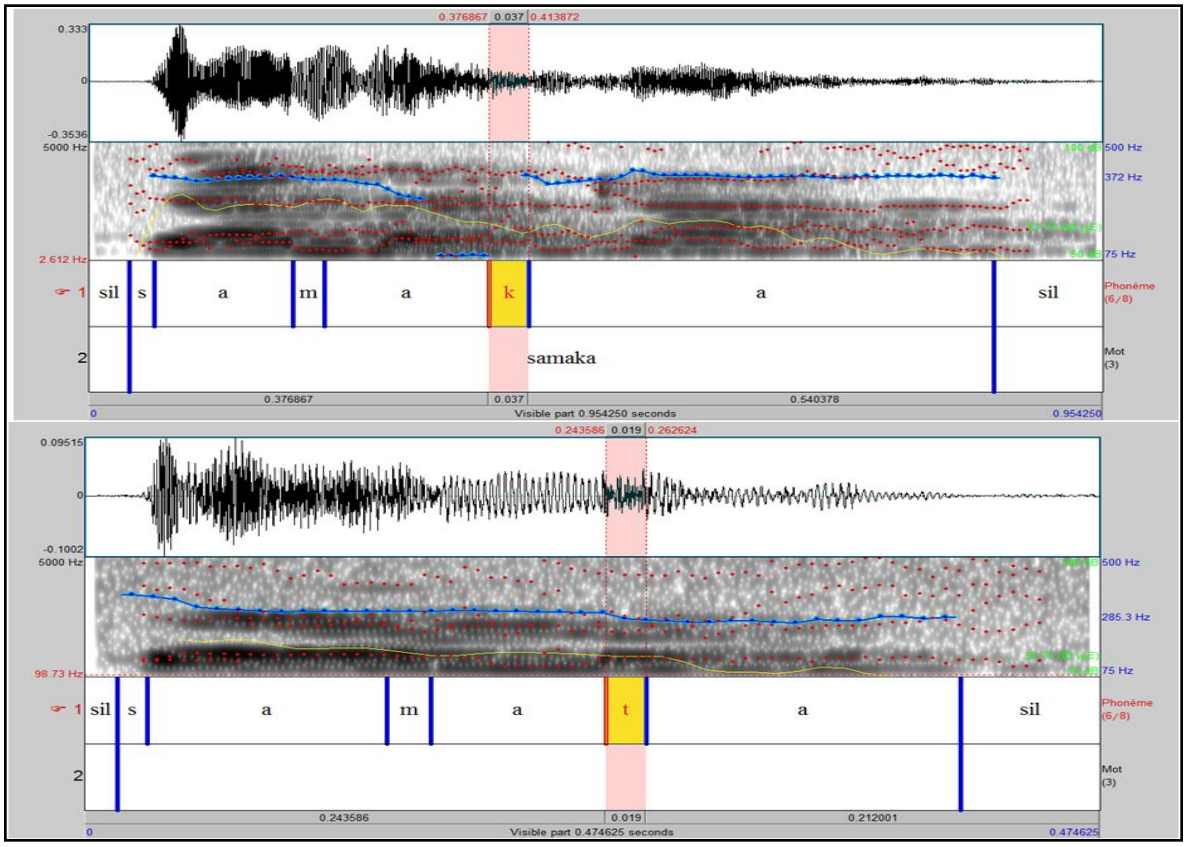


Figure 27 : Segmentation du mot [samaka]

Nous illustrons, dans la figure 28, la forme du signal pour chaque mot de cette opposition. Ces signaux sont produits par des locuteurs masculins. La comparaison visuelle ne montre aucune décision sur la présence d'une substitution phonémique pour le son [k]. La difficulté est observée même pour ces deux phonèmes qui sont trop éloignés dans la zone d'articulation (le vélaire [k] et l'alvéolaire [t]).

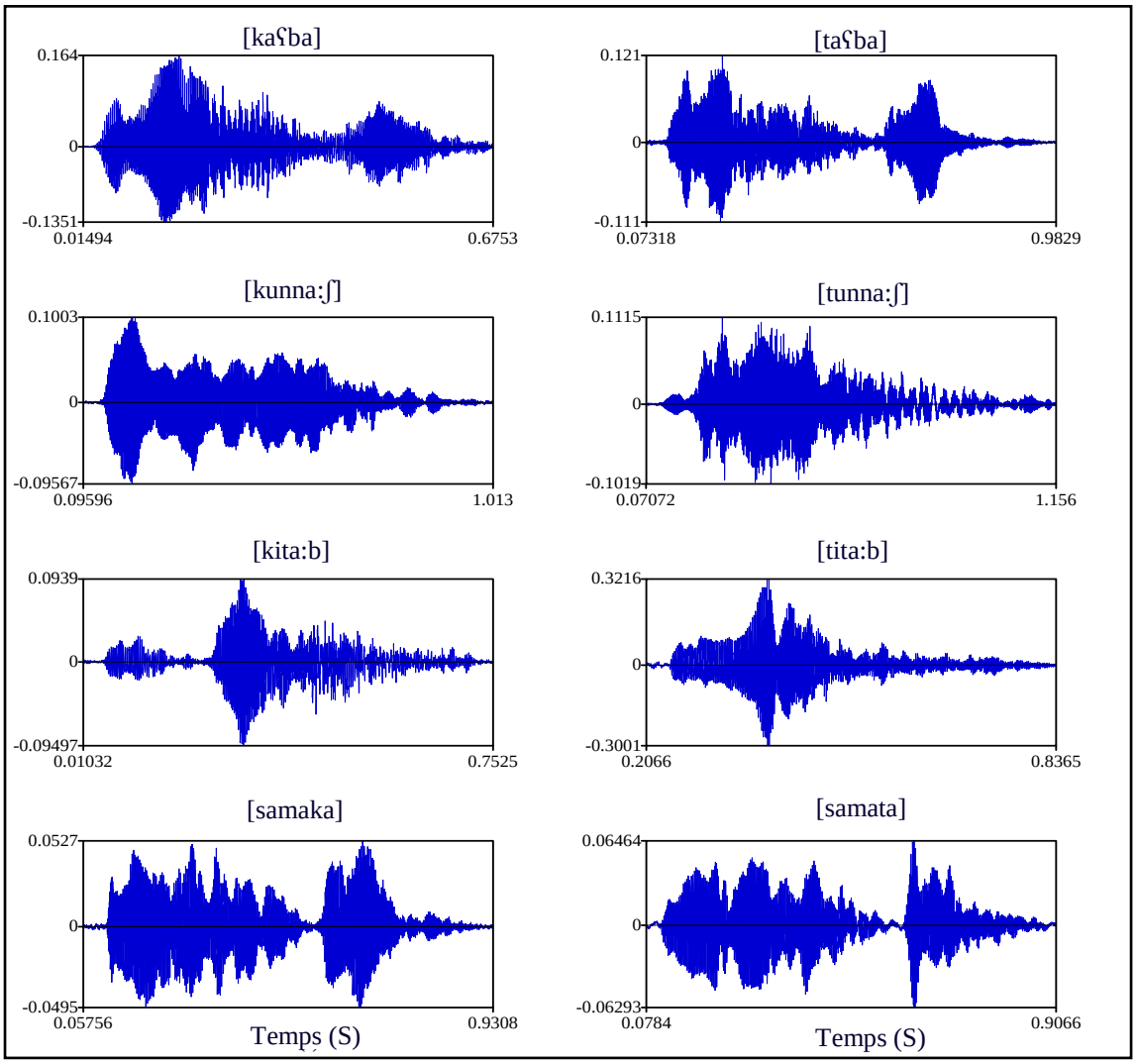


Figure 28 : Comparaisons visuelles des prononciations (opposition [k]/[t])

Aussi, une étude sur spectrogramme est réalisée pour voir la possibilité de décider sur la présence du ESP pour cette opposition (figure 29). Ce spectrogramme est montré pour deux prononciations du mot [kaʕba]. Nous remarquons un changement total entre les deux cas, mais nous ne pouvons pas garantir sur quel phonème ce changement apparaît. Par conséquent, la nécessité d'utiliser d'autres techniques basées sur les statistiques est impérative.

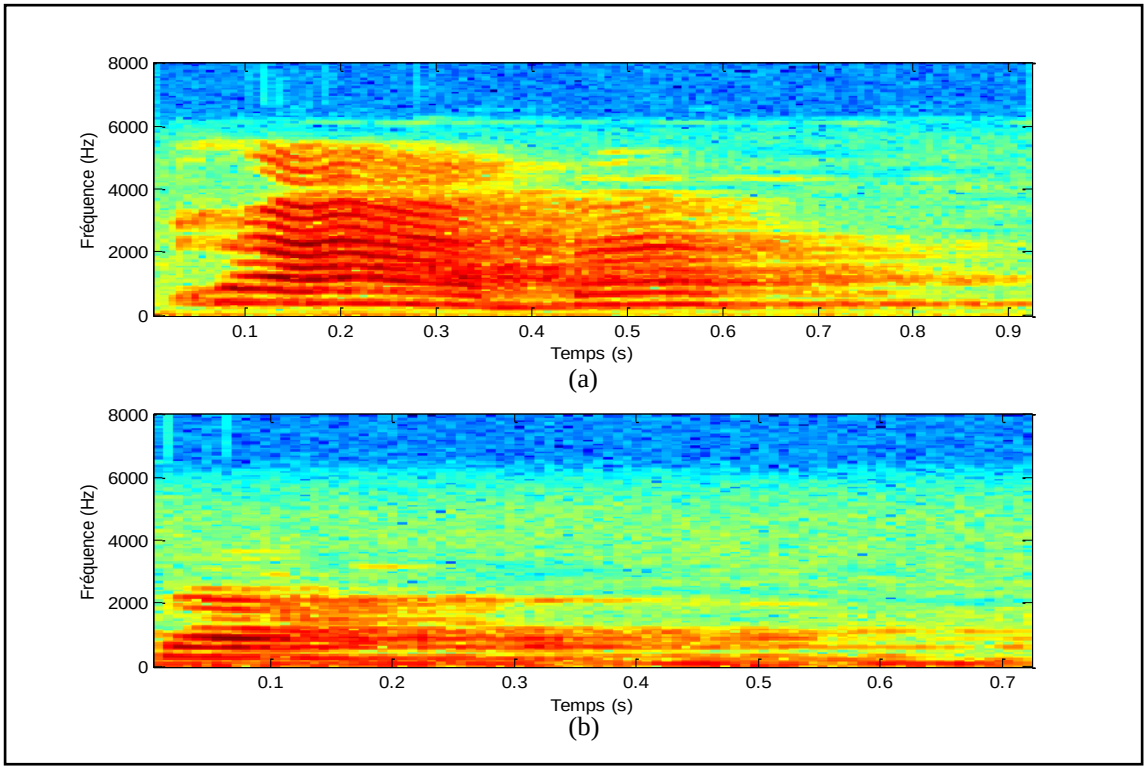


Figure 29 : Spectrogramme du mot [kaʕba] (a) prononciation correcte (b) prononciation incorrecte

Finalement, pour cette opposition, nous montrons les performances du système proposé toujours en fonction du nombre des Composantes Gaussiennes (tableau 27) et en fonction du nombre d'états (tableau 28). Les MFCC et les Delta-MFCC sont employés pour la représentation des signaux de parole.

Tableau 27 : Performances du système en fonction de l'ordre du modèle (opposition [k]/[t]))

		Ordre du modèle			
Paramètres	**Performances**	**2**	**4**	**8**	**16**
MFCC	Sensibilité	74.98%	77.55%	80.91%	82.92%
	Spécificité	73.64%	79.96%	78.22%	77.88%
	TCC	73.83%	78.56%	79.25%	80.03%
Delta-MFCC	Sensibilité	75.48%	79.83%	80.06%	79.77%
	Spécificité	75.83%	77.06%	79.68%	83.50%
	TCC	74.94%	78.21%	79.64%	81.46%

Pour ce dernier cas, le TCC maximal est obtenu par 16 composantes gaussiennes, ce taux est de 81.46%. Les Delta-MFCC montrent toujours de meilleures performances en comparaison avec les MFCC seuls. La spécificité et la sensibilité sont respectivement de 83.50% et de 80.06%.

Tableau 28 : Performances du système en fonction du nombre d'états (opposition [k]/[t]))

Paramètres	Performances	Nombre d'états 3	4	5	6
MFCC	Sensibilité	75.78%	80.67%	78.82%	78.47%
	Spécificité	77.89%	78.50%	77.87%	78.99%
	TCC	76.41%	78.95%	78.23%	78.53%
Delta-MFCC	Sensibilité	79.53%	81.14%	78.94%	79.88%
	Spécificité	80.26%	79.82%	84.07%	81.78%
	TCC	79.62%	79.63%	81.11%	80.71%

Un HMM de 5 états montre un TCC de 81.11% et une spécificité de 84.07%. Cependant, ce modèle dégrade la sensibilité de 81.14% à 78.94% par rapport à un HMM de 4 états. Généralement, un HMM/GMM de 5 états et de 16 Composantes Gaussiennes est nécessaire pour aboutir à un système de classification plus robuste.

4.3 Implémentation du SCAESP

Afin de simplifier la procédure de correction d'articulation, nous avons réalisé une application graphique SCAESP. Cette dernière offre plusieurs objets suivant le phonème à corriger. Ces objets sont montrés sous forme d'images et de fichiers audio. Le patient essaye d'écouter et de prononcer les différents mots avec plusieurs répétitions pour assurer le bon fonctionnement de son appareil phonatoire. L'enfant présentant ces erreurs de prononciation peut utiliser cette application à la maison en s'aidant d'un orthophoniste ou de ses parents.

4.3.1 Collection des données

L'application proposée est conçue pour répondre aux exigences spécifiques des ESP. Au cours d'une séance thérapeutique, un enfant serait capable de prononcer plusieurs mots montrés sous forme d'images (tableau 29).

Tableau 29 : Mots utilisés dans le processus thérapeutique
Ar : Arabe ; TOP : Transcription Orthographique-Phonétique

	Mots							
	Position 1		Position 2		Position 3		Position 4	
	Ar	TOP	Ar	TOP	Ar	TOP	Ar	TOP
[s]	سيارة	[sajja:ra]	سلم	[sullam]	سروال	[sirwa:l]	فانوس	[fa:nu:s]
[r]	راجل	[ra:dʒil]	رمان	[rumma:n]	رمال	[rima:l]	بدر	[badar]
[z]	زهرة	[zahra]	زحل	[zuħal]	زناد	[zina:d]	همزة	[hamza]
[dʒ]	جرة	[dʒarra]	جنود	[dʒunu:d]	جمال	[dʒima:l]	مهرج	[muharridʒ]
[k]	كعبة	[kaʕba]	كتاب	[kita:b]	كناش	[kunna:ʃ]	سمكة	[samaka]

4.3.2 Environnement du SCAESP

Le SCAESP est développé sous l'environnement C++ builder. Des objets créés peuvent être utilisés par l'enfant ; il peut entendre les fichiers audio enregistrés sous forme des signaux (.wav). Cette application fonctionne avec un ordinateur et un microphone. Au début, l'orthophoniste charge les informations nécessaires : le prénom, le nom et l'âge de l'enfant, en sélectionnant le phonème cible (figure 30).

Figure 30 : Fenêtre principale du SCAESP

Une fois que la cible est sélectionnée, une autre fenêtre apparait. Celle-ci montre quatre positions différentes pour la prononciation du phonème cible (figure 31).

L'enfant suit une procédure de prononciation avec plusieurs répétitions. Pendant chaque réalisation, le système utilise le signal de parole capté par le microphone pour calculer le taux de déviation et le stocke dans un fichier de sortie. A la fin de chaque séance thérapeutique, le système offre la possibilité d'imprimer le rapport final. Ceci garde la déviation totale et le niveau d'évaluation perceptuelle complétés par l'enfant. Si le taux de déviation s'améliore avec le temps, c'est un signe positif pour la correction d'articulation de l'enfant. Dans le cas contraire, le processus doit être répété en reprenant plusieurs séances. Le nombre des séances dépendra de l'intelligence et du degré de coopération de l'enfant.

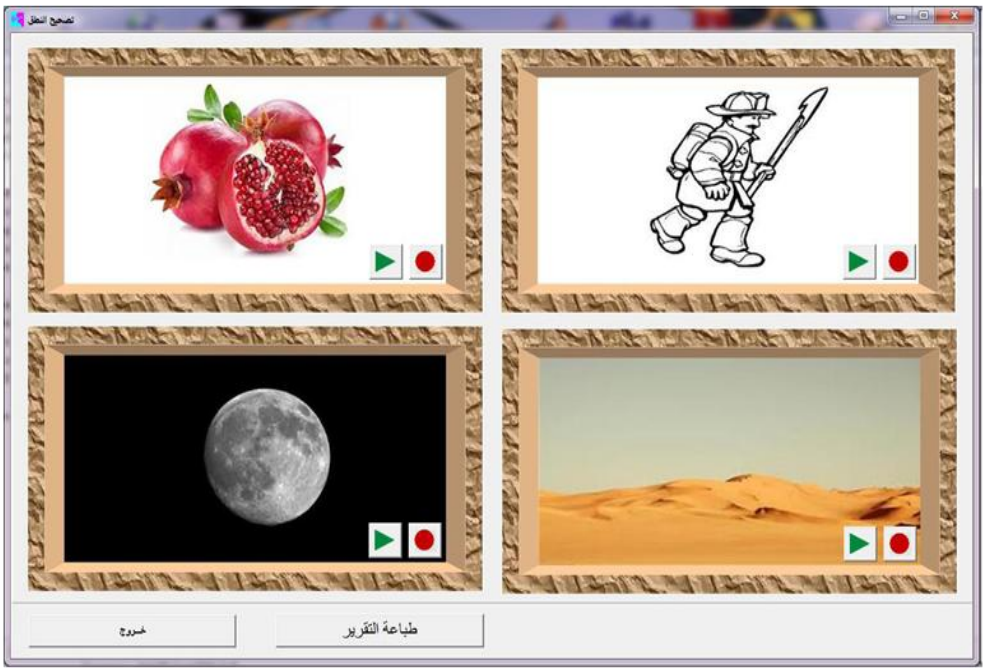


Figure 31 : Exemple de Séance de rééducation du son [r]

4.3.3 Evaluation du niveau de prononciation

Afin de montrer les performances du SCAESP, pour la rééducation des ESP dans l'arabe parlé, une étape de test est envisagée. Pour le test initial, nous l'avons essayé avec un enfant ayant une bonne prononciation, puis avec un groupe de cinq enfants (5 ans) ayant des ESP. Chacun d'eux montre une ESP différente de l'autre. Nous aidons les enfants dans le contrôle de ce système par des sessions de 30 minutes/semaine.

Le rapport final créé à la fin de chaque séance de rééducation montre le niveau de correction de prononciation de l'enfant. Ce rapport peut être stocké dans le dossier du patient pour une utilisation future. Il pourra aider l'orthophoniste lors des prochaines sessions, en sélectionnant les différentes activités suivant les informations stockées dans ce rapport.

La figure 32 donne le développement de la correction de prononciation de deux enfants. Le premier est un garçon qui substitue souvent le phonème [r] par [ʁ]. Le second est une fillette de 5 ans qui substitue souvent le phonème [s] par [θ]. Les résultats obtenus montrent l'efficacité du SCAESP.

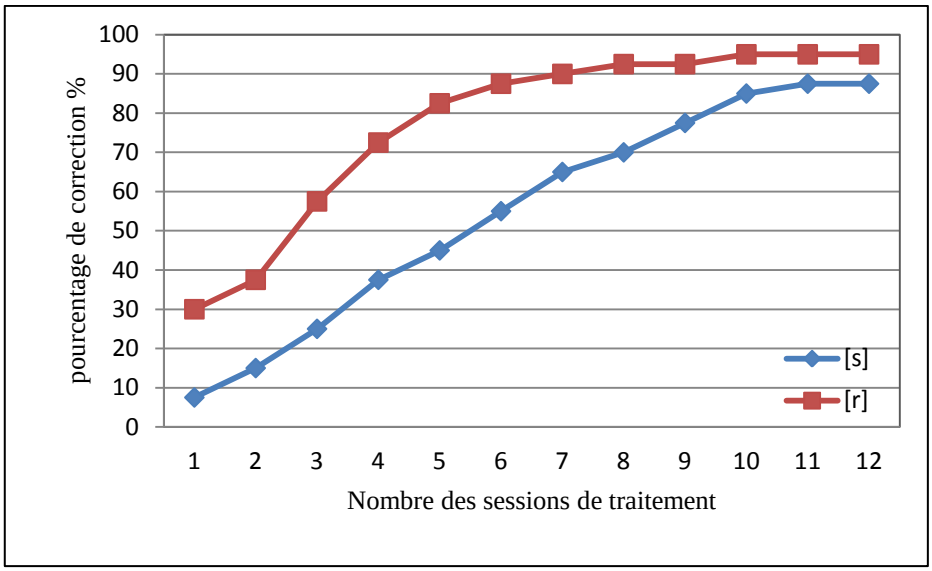


Figure 32 : Correction de la prononciation pour les deux enfants

4.4 Conclusion

Dans ce chapitre, nous avons étudié les performances du SCAESP en fonction des nombres d'états (HMM) et des Composantes Gaussiennes (GMM) en montrant l'amélioration apportée par les Delta-MFCC. Finalement, pour simplifier son utilisation et mettre en valeur son efficacité, nous l'avons implémenté sous forme d'une application graphique testée initialement avec un enfant ayant une bonne prononciation, puis sur cinq enfants ayant des ESP. Après trois mois de rééducation, avec une séance de 30 minutes/semaine, ces enfants ont montré des résultats satisfaisants dans leurs prononciations.

Conclusion Générale

Conclusion Générale

Ce travail s'intéresse essentiellement à la classification automatique des Erreurs de Substitution Phonémique (ESP) chez l'enfant arabophone. Celles-ci sont les plus fréquentes parmi les erreurs de prononciation durant les périodes d'apprentissage de la langue parlée. Ces erreurs sont les conséquences d'un changement dans une caractéristique distinctive, à savoir le lieu et le mode d'articulation.

A partir de l'âge de 5 jusqu'à 6 ans, l'enfant arabophone est capable de prononcer tous les sons de la langue arabe. La persistance de ces erreurs de prononciation au-delà de 5 - 6 ans peut influer directement sur la communication de l'enfant avec les autres et cela nécessite une intervention avant l'entrée en préscolaire.

Les enfants ayant ces erreurs de substitution phonémique rencontrent plusieurs problèmes durant les programmes de rééducation orthophonique. Ces problèmes peuvent être considérés les conséquences de plusieurs facteurs. A titre d'exemple nous citons : le nombre insuffisant de spécialistes dans le domaine, le délai nécessaire entre les différentes sessions de rééducation orthophonique et les moyens nécessaires pour cette opération (le temps, les efforts, le trajet, l'argent, etc.).

L'évolution remarquable des systèmes de Reconnaissance Automatique de Parole, nous a motivés pour les appliquer dans le traitement de ces erreurs. Ces systèmes reposent sur les Modèles de Markov Cachés et les Modèles de Mélange de Gaussiennes (HMM/GMM), dont l'objectif principal est la conception d'un Système de Classification Automatique des Erreurs de Substitution Phonémique (SCAESP).

Nous avons commencé par une étude concernant les notions de base sur l'Arabe standard. En outre, nous avons décrit les différentes erreurs de prononciation liées au retard simple de la parole qui regroupent à la fois l'addition, l'omission, la distorsion et la substitution phonémique. Cette étude a été effectuée en trois aspects différents :

- la phonétique de l'AS, en montrant la transcription, l'articulation et la prononciation des différents sons ;
- l'acquisition de l'AS, en analysant les périodes d'apprentissage (pré-linguistique et linguistique) ;

- le retard simple de parole, en présentant les différentes erreurs de prononciations (ESP, EOP, EAP et EDP).
- Dans une première étape, nous avons mené une analyse approfondie sur les HMM et les GMM et leurs applications dans les systèmes de la RAP. Nous avons décrit les techniques d'apprentissage et de reconnaissance, fondées respectivement sur l'algorithme de Baum-Welch et de Viterbi. La paramétrisation des signaux de parole est basée sur les MFCC et les Delta-MFCC. Ces paramètres spectraux sont obtenus habituellement par une transformation pour fournir aux trames une représentation faiblement corrélée et de dimensions réduites.

Ces techniques sont incorporées dans le système proposé pour la rééducation orthophonique des Erreurs de Substitution Phonémique. Il traite cinq oppositions, qui sont sélectionnées pendant quelques séances d'enregistrement au niveau de trois écoles primaires en Algérie. En outre, nous avons fait une étude comparative visuelle sur spectrogrammes de ces différentes oppositions.

Les oppositions sélectionnées montrent les cas de substitution les plus fréquents chez les enfants arabophones. Elles couvrent plusieurs cas en montrant le changement de lieux d'articulation et l'absence de vibrations des cordes vocales :

- Alvéolaire / Interdentale [s]/[θ] ;
- Alvéolaire / Interdentale [z]/[ð] ;
- Alvéolaire / Uvulaire [r]/[ʁ] ;
- Palatale / Alvéopalatale [ʤ]/[ʃ] ;
- Vélaire / Alvéolaire [k]/[t].

Chacun de ces phonèmes est différent de l'autre par son lieu ou son mode d'articulation. Nous avons illustré la configuration de l'appareil phonatoire pendant la prononciation de chacun de ces phonèmes. Nous avons estimé les paramètres pertinents : la durée, la fréquence fondamentale et l'énergie des phonèmes.

Pour réaliser notre objectif, nous avons élaboré un corpus de parole à l'aide d'un groupe de 50 enfants âgés entre 5 et 6 ans. Nous avons préalablement segmenté les signaux de parole

manuellement, en utilisant l'outil d'analyse PRAAT. Ce corpus est divisé en deux sous corpus l'un est utilisé pour l'apprentissage du système et l'autre pour la phase du test.

Les performances du SCAESP sont obtenues en estimant la sensibilité, la spécificité et le Taux de Classification Correcte TCC. Celles-ci montrent l'efficacité d'utilisation des HMM/GMM pour traiter ce type d'application. Le SCAESP montre des TCC satisfaisants concernant les oppositions consonantiques [s]/[θ] ; [z]/[ð] ; [r]/[ʁ] ; [ʤ]/[ʃ] et [k]/[t], qui sont respectivement de : 84.15% ; 80.27% ; 85.57% ; 90.10% et 81.46%.

En termes de perspectives, nous préconisons de modéliser d'autres oppositions pour couvrir tous les cas possibles des ESP. Nous proposons également de mettre le SCAESP à la disposition des orthophonistes pour l'intégrer dans les programmes de rééducation orthophonique. Enfin, la réalisation d'un corpus de parole orienté vers cet axe de recherche est impérative.

Bibliographie

[1] Deborah Lott, Super star speech : Speech Therapy Made Simple !, USA : Super Star DML Publishing - Huntsville, Alabama, 2007. 74 p. ISBN 13 978-0-9798041-0-6

[2] John Duncan Ferguson, Hidden Markov Analysis : An Introduction, Symposium on the Applications of Hidden Markov Models to Text and Speech, J. D. Ferguson, editor, October, 1980. IDA-CRD, Princton, NJ, pp. 143-179.

[3] Yu. Shun-Zheng, Hidden semi-Markov Models, Artificial Intelligence, Elsevier, 2010, Vol. 174, pp. 215-243.

[4] Djamel Bouchaffra, Embedding HMMs-based Models in a euclidean space: the topological hidden markov models, Pattern Recognition, Elsevier, 2010, Vol. 43, pp. 2590-2607.

[5] Karin C. Ryding, A reference grammar of modern standard Arabic, Cambridge university press, 2005. 731 p. ISBN 0 521 77151 X

[6] David Cohen, Langue arabe, Encyclopedia Universalis, Paris, 1990, pp. 707-732.

[7] Muhammad Alkhouli, Alaswaat Alaghawaiyah, in Proceedings of International Conference on Signal Processing, Jordan, 1990. pp. 646-651.

[8] Mustafa Elshafei, Toward an arabic text-to-speech system, The Arabian Journal for Science and Engineering, 1991, Vol. 16(4B), pp. 565-583.

[9] Yousef Ajami Alotaibi, Investigating the adaptation of arabic speech recognition systems to foreign accented speakers, Sid-Ahmed Selouani, 10th International Conference on Information Science, Signal Processing and their Applications, IEEE, 2010, pp. 646-649.

[10] Ahmed Zargua, Ousoul al-lugha al-arabia : Asrar al-huruf, Dar al-hasad, Syrie, 1993.

[11] Abdulrahman Ibrahim Alfozan, Assimilation in Classical Arabic ; A phonological study. 304 p. Thèse de doctorat, Faculty of Arts of the University of Glasgow, Scotland : 1989.

[12] Nahed Boukadida, Connaissances phonologiques et morphologiques dérivationnelles et apprentissage de la lecture en arabe (Etude longitudinale). 277 p. Education, Université Rennes 2, Université de Tunis : 2008.

[13] Khaoula Taleb Ibrahimi, Principes dans la linguistique, Algérie : Dar al-qasaba, 2006. 200 p. ISBN 4-207-64-9961

[14] Wai C. Chu, Speech Coding Algorithms Foundation and Evolution of Standardized Coders, USA : WILEY-INTERSCIENCE, a John Wiley & Sons, inc., publication, 2003. 578 p. ISBN 0-471-37312-5

[15] K. Bartkova, Production, description et perception du signal vocal, Rapport technique, France Telecom R & D Lannion, 2002.

[16] Jean Hennebert, Speaker Recognition, Overview. Encyclopedia of Biometrics, Springer reference, 2009, pp. 1262-1270. ISBN: 978-0-387-73002-8

[17] D. Schwarz, Spectral Envelopes in Sound Analysis and Synthesis, IRCAM Institut de la Recherche et Coordination Acoustique/Musique, 1998.

[18] J. P. Campbell, JR., Speaker Recognition: A Tutorial, Proceedings of the IEEE, Vol. 85, No. 9, pp. 1437-1462, 1997.

[19] Y. Mami, Reconnaissance de locuteurs par localisation dans un espace de locuteur de référence. Thèse de doctorat, Ecole Nationale Supérieure des Télécommunications, Paris, 2003.

[20] R. Boite, Traitement de la parole. Presses polytechniques et universitaires romandes, 2000.

[21] M. Delahaie, L'évolution du langage chez l'enfant : De la difficulté au trouble, Guide ressources pour les professionnels, INPES édition, France, 2004.

[22] M. Touzin, Les différents troubles d'apprentissage, ADSP No. 26, pp. 30-35, 1999.

[23] W. Lanier, Speech disorders (Diseases & disorders), Gale, Cengage Learning 2010. ISBN-13: 978-1-4205-0221-3

[24] J.S. Damico, N. Müller, and M.J. Ball , The Handbook of Language and Speech Disorders, Chapter 3, pp. 339-360, A John Wiley & Sons, Ltd., Publication, 2010. ISBN: 978-1-405-15862-6.

[25] A. Eshajhse, Articulation and speech disorders : types, treatment and diagnostic, Saudi Arabia, limited golden papers, 1997.

[26] M.E. Gordon-Brannan C.E. Weiss, Clinical Management of Articulatory and Phonologic Disorders, 3rd Edition, Lippincott Williams & wilkins, 2007.

[27] A. Abed, M. Guerti, Classification automatique des consonnes arrières arabe en vue de la correction de la substitution phonémique. National conference on speech processing NCSP'2014, pp. 127-134, Alger, 10-11 Dec, 2014.

[28] A. Abed, M. Guerti, Application des HMM à la substitution Phonémique dans l'Arabe Parlé, Journées d'Etudes Algéro-Françaises de Doctorants en Signal-Image & Applications, JEAFD'2012. pp. 18--23, Alger 5-6 Dec, 2012.

[29] M. Ghulam, T. Mesallam, K. Malki, M. Farahat, A. Mahmood, M. Alsulaiman, Multidirectional Regression (Mdr)-Based Features For Automatic Voice Disorder Detection, Journal of Voice, Vol. 26 , No. 6, pp. 817.19-817.e27, 2012.

[30] L. Salhi, M. Talbi, C. Adnene, Voice Disorders Identification Using Multilayer Neural Network. The International Arab Journal of Information Technology, Vol. 7, N° 2, pp. 177-185, 2010.

[31] G. Muhammad, M. AlSulaiman, A. Mahmood, Z. Ali, Automatic Voice Disorder Classification using Vowel Formants, International Conference on Multimedia and Expo (ICME), pp.1-6, 2011.

[32] M. Alsulaiman, Voice Pathology Assessment Systems for Dysphonic Patients : Detection, Classification, and Speech Recognition. IETE Journal of Research, Vol. 60, issue 2, pp. 156-167, 2014

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

[34] J. Makhoul, Linear prediction: A tutorial review, Proceedings of the IEEE, Vol.63, No.4, pp.561-580, April, 1975.

[35] Y. Kharin, Robustness in Statistical Pattern Recognition, Boston, Kluwer Academic, 1996.

[36] B.D. Ripley, Pattern Recognition and Neural Networks, Cambridge University Press, January, 1996.

[37] Calliope, La parole et son traitement automatique, Collection technique et scientifique des télécommunications, 1989.

[38] V. Barreaud, Reconnaissance Automatique de la Parole Continue : compensation des bruits par transformations de la parole, Phd Thesis, Nancy 1 (France), 2004.

[39] A. Hacine-Gharbi, Sélection de paramètres acoustiques pertinents pour la reconnaissance de la parole, thèse de doctorat de l'Université d'orléans, Dec. 2012.

[40] L. Rabiner, B. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1993.

[41] P. Lockwood, C. Baillargeat, J. Gillot, J. Boudy, G. Faucon, Noise reduction for speech enhancement in cars : non linear spectral subtraction, Eurospeech, Genova, Italia, 1991.

[42] A.V. Oppenheim, R.W. Schafer, Discrete-Time Signal Processing, Ehglewood Cliffs, NJ, USA, 1989.

[43] A. Hossan, Sh. Memon, M. Gregory, A novel approach for MFCC feature extraction, 4th International Conference on Signal Processing and Communication Systems (ICSPCS), pp. 1-5, IEEE, 2010.

[44] J.D. Ferguson, Hidden Markov Analysis: An Introduction, Proc. of the Symposium on the Applications of Hidden Markov Models to Text and Speech, IDA-CRD, pp. 8-15, Princeton, NJ 1980.

[45] Yu. Shun-Zheng, Hidden Semi-Markov Models, Artificial Intelligence, Elsevier, Vol. 174, pp. 215-243, 2010.

[46] B.H. Juang, S. Levinson, M. Sondhi, Maximum likelihood estimation for multivariate mixture observations of markov chains, IEEE Transactions in Information Theory, vol 32(02), pp. 307-309, 1986.

[47] X. Huang, A. Acero, H.W. Hon, Spoken language processing, A guide to theory, Algorithm, and System Development, Chapter 8, pp. 375-412, Printice-Hall, 2001.

[48] Viterbi, Error bounds for convolutional codes and an asymptotically optimal decoding algorithm, Transactions on Information Theory IEEE, Vol. 13, No. 2, pp. 260-269, 1967.

[49] J.A. Bilmes, 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.

[50] A. Abed, M. Guerti, Vectorial & Statistical Approaches for Automatic Speaker Identification Systems, 1st National Conference on Electronics and New Technologies NCENT'2015, M'Sila, Algeria, 19-20 May, 2015.

[51] Praat: doing Phonetics by Computer, http://www.fon.hum.uva.nl/praat, 2016.

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

[53] Kevin Murphy, 'Hidden Markov Model (HMM) Toolbox', 1998. Http://www.ai.mit.edu/~murphyk/Software/hmm.html

[54] A. Abed, M. Guerti, HMM/GMM Classification for Articulation Disorder Correction among Algerian Children, International Arab Journal of Information Technology, Vol. 13, N° 4, pp. 449-455, July 2016. IF: 0.582. 1683-3198 Print ISSN, 2309-4524 Online ISSN. Http://ccis2k.org//iajit/PDF/vol.13, no.4/9418.pdf