جامعة سيدي محمد بن عبد الله بفاس
UNIVERSITE SIDI MOHAMED BEN ABDALLAH DE FES
كلية العلوم ظهر المهراز فاس
FACULTE DES SCIENCES DHAR EL MEHRAZ FES

N° d'Ordre : **68 /2020**

Formation Doctorale : Sciences et Technologies de l'Information et de la Communication (STIC)

Discipline : Informatique

Spécialité : Informatique

Laboratoire : Informatique, Signaux, Automatique et Cognitivisme (LISAC)

THESE de DOCTORAT

Présentée par

ZEALOUK Ouissam

Système de reconnaissance automatique de l'Amazighe et analyse des formants pour le diagnostic vocal

Soutenue le 28 / 11 / 2020 devant le jury composé de :

Pr. ABARKAN Mustapha	***Faculté Polydisciplinaire - Taza***	***Président***
Pr. YOUSFI Abdellah	***FSJES - Rabat***	***Rapporteur***
Pr. OUANAN Mohamed	***Faculté des Sciences - Meknès***	***Rapporteur***
Pr. EL MOUTAOUAKIL Karim	***Faculté Polydisciplinaire - Taza***	***Rapporteur***
Pr. LOQMAN Chakir	***Faculté des Sciences Dhar El Mahraz - Fès***	***Examinateur***
Pr. EL ABDERRAHMANI Abdellatif	***Faculté des Sciences Dhar El Mahraz - Fès***	***Examinateur***
Pr. Satori Hassan	***Faculté des Sciences Dhar El Mahraz - Fès***	***Examinateur Directeur de thèse***
Pr. Satori Khalid	***Faculté des Sciences Dhar El Mahraz - Fès***	***Co-encadrant***

Année universitaire : 2019-2020

Résumé

La parole est la forme la plus naturelle de la communication humaine. Elle est délectable, déchiffrée et par conséquent reconnue. La reconnaissance automatique de la parole (RAP) est faisable grâce à des systèmes bien développés. Il s'agit d'une méthode de décodage du signal vocal capturé par le microphone pour le convertir ensuite en mots. Cette technologie est entrain de connaitre une grande évolution non seulement dans les domaines industriels et publics où ces techniques intègrent les appareils électroniques utilisés quotidiennement, mais aussi dans d'autres domaines tels que l'éducation, le médical, le militaire, etc. L'objectif de cette thèse est la réalisation des systèmes de reconnaissance et de diagnostic basant sur les formalismes statistiques Markovien pour modéliser les mots prononcés à base des unités phonétiques élémentaires. Notre première système de la reconnaissance était créé pour reconnaitre les locuteurs fumeurs via utilisation les techniques et les algorithmes de la reconnaissance automatique de la parole. Pour bien confirmer notre étude, nous avons examiné la voix humaine de 40 adultes (20 fumeurs et 20 non-fumeurs) afin de déterminer les effets du tabagisme sur les paramètres vocaux humain en s'appuyant sur 3 voyelles de la langue Amazighe (A, I, U). Après avoir acquis des résultats favorables concernant le diagnostic des fumeurs, nous avons commencé à construire un deuxième système de reconnaissance automatique de la parole permettant de diagnostiquer la parole des personnes atteintes de troubles de voix. Ce projet de recherche est réalisé en langue Amazighe. Son objectif est de distinguer entre les voix normales, celle des fumeurs et pathologiques. Afin de développer un système au milieu réel, nous avons étudié l'effet du bruit sur les dix premiers chiffres Amazighs dans des conditions bruyantes d'un point de vue RAP basé sur le rapport signal sur bruit en anglais signal-to-noise ratio (SNR). Nos expériences de tests ont été réalisées sous deux types de bruits et répétées avec un bruit environnemental supplémentaire avec différents rapports SNR pour chaque type allant de 5 dB à 45 dB. Les performances de nos systèmes ont été mesurées en utilisant des combinaisons des états MMC avec des distributions de mélange gaussiennes. Nos résultats obtenus sont très satisfaisants.

Mots clés :

Reconnaissance automatique de la parole, intelligence artificielle, modèle de Markov caché, algorithme de Viterbi, GMM, MFCC, langue amazighe, locuteur fumeurs, troubles de la voix, formants, bruit.

Abstract

Speech is the most natural form of human communication, of which Automatic Speech Recognition (ASR) is a method of decoding the voice signal captured by the microphone and converting it into words. This technology is undergoing a great evolution not only in industrial and public fields where these techniques integrate electronic devices used daily, but also in other fields such as education, medical, military, etc. The objective of this thesis is the realization of recognition and diagnostic systems based on statistical Markovian formalisms to model pronounced words based on elementary phonetic units. Our first recognition system was created to recognize smoking speakers using automatic speech recognition techniques and algorithms. To confirm our study, we examined the human voice of 40 adults (20 smokers and 20 non-smokers) to determine the effects of smoking on human vocal parameters based on 3 vowels of the Amazigh language (A, I, U). After acquiring one of the encouraging results concerning the diagnosis of smokers, we started to build a second automatic speech recognition system which allows to diagnose the speech of people who have voice disorders. This research project is carried out in Amazigh language in order to differentiate normal, smoking and pathological voices. In order to develop a system in the real environment, we studied the effect of noise on the first ten Amazigh digits under noisy conditions from a signal-to-noise ratio (SNR) -based ASR perspective. Our test experiments were performed under two types of noise and repeated with additional environmental noise with different SNR ratios for each type ranging from 5dB to 45dB. The performance of our systems was measured using combinations of HMM states with Gaussian mixing distributions. Our results obtained are very satisfactory.

Keywords:

Automatic Speech Recognition, Artificial intelligence, Hidden Markov model, Viterbi algorithm, GMM, MFCC, Amazigh language, Smokers' speakers, Voice disorders, Formants, Noise.

Remerciement

Je remercie ALLAH le tout puissant de m'avoir donné la santé et la volonté d'entamer et de terminer cette thèse.

Tout d'abord, ce travail ne serait pas aussi riche et n'aurait pas pu voir le jour sans l'aide et l'encadrement de mon directeur de recherche Prof. Hassan Satori. Je le remercie aussi pour la qualité de son encadrement exceptionnel, pour sa patience, sa rigueur et sa disponibilité durant ma préparation de la thèse.

Je voudrais remercier sincèrement mon Co-encadrant Prof. Khalid Satori, Professeur à l'Université Sidi Mohammed Ben Abdellah, Faculté des sciences Fès, pour les nombreux conseils judicieux qu'il m'a prodigués et pour sa disponibilité tout au long de cette recherche. Qu'il retrouve ici le témoignage de ma profonde gratitude.

Je voudrais aussi exprimer ma gratitude envers tous ceux qui m'ont accordé leur soutien, tant par leur gentillesse que par leur dévouement, en particulier Mohamed Hamidi qui au fil de nos discussions a fait progresser ma réflexion, Abdelkader Benzirar, Youssef Boutazart et Mimoun El Baghdadi.

J'adresse mes sincères remerciements aux rapporteurs de ma thèse, dont les remarques et suggestions permettront d'améliorer la qualité de ce travail. Je remercie également tous les membres de mon jury pour leur disponibilité afin d'évaluer ce travail.

J'ai une pensée toute particulière pour Monsieur le professeur Mustapha Zerfaoui, qu'il trouve ici l'expression de ma gratitude.

Je n'oublierai pas d'exprimer ma reconnaissance au groupe CMU Sphinx de l'université Carnegie Mellon qui ont développé l'open source CMU-Sphinx particulièrement M. Nickolay V. Shmyrev

Durant toutes ces années, j'ai eu l'occasion de rencontrer de nombreuses personnes, dans un cadre purement professionnel ou simplement amical. A leur façon, ils ont tous contribué à mon apprentissage. Je suis reconnaissant envers chacune de ces personnes.

Enfin, je voudrais exprimer mes plus profonds remerciements à mes parents, à mon frère, ma sœur, à toute ma grande famille et mes amis pour leurs sentiments, leurs soutiens et leurs encouragements dans pendant le temps où j'ai effectué cette thèse.

.

Table des matières

Liste des figures

Liste des tableaux

Liste des abréviations

ASR : Automatic Speech Recognition

CAN : Convertisseur Analogique- Numérique

CI : Context-Independant

DSVD Digital Simultaneous Voice and Data

DTW : Dynamic Time Warping

EM Expectation Maximization

FFT : Transformée de Fourier de Fourier

GMMs : Gaussian Mixture Models

HTK : Hidden Markov Model Toolkit

IRCAM : L'institut Royal de la Culture Amazighe

LPC : Linear Predictive Coding

MFCC : Mel Frequency Cepstrum Coefficients

MLE Vo: Maximum Likelihood Estimation

MMC : Modèles de Markov Cachés

NPC : Neural Predictive Coding

PLP : Perceptual Linear Prediction

RAP : Reconnaissance Automatique de la Parole

SNR : Signal-to-Noise Ratio

TDC : Transformation Discrète de Cosinus

TFD : Transformée de Fourier Discrète

VoIP : Voice over IP

WER : Word Error Rate

WRR : Word Recognition Rate

Introduction Générale

La parole est la forme de communication la plus naturelle que les humains utilisent pour échanger l'information. En effet, le signal vocal permet la transmission claire d'une grande quantité d'informations venant des locuteurs tels que le message linguistique, le sexe, l'âge, l'état psychologique, maladie, etc. Grace aux efforts des chercheures dans le domaine de la Reconnaissance Automatique de la Parole (RAP), il est actuellement possible qu'une machine puisse reconnaitre l'état émotionnel et réagir à un langage naturel humain où les énoncés sont prononcés par une personne.

La reconnaissance de la parole est l'un des sujets sensibles au centre de nombreuses études dans des domaines multidisciplinaires. Elle est de plus en plus répandue, alimentant les assistants virtuels populaires, facilitant le sous-titrage automatisé et offrant des plates-formes de dictée numérique. De nos jours, on a même observé l'intégration de reconnaissance vocale dans les domaines de la santé pour diagnostiquer des maladies. Au cours des dernières années, la technique de RAP a connu des progrès exceptionnels grâce au développement technologique dans les domaines de l'intelligence artificielle, méthodes de traitement de l'information, stockage de données et l'avènement des nouvelles moyennes de communication.

Notre premier objectif dans ce travail consiste à étudier et à préparer un système de reconnaissance de la parole de référence dédiée à reconnaitre les dix premiers chiffres de la langue Amazigh. Notre étude vise l'analyse de chaque modèle de RAP avec une adaptation à la langue Amazigh basée sur les méthodes statistiques et la modélisation Markovien. Notre deuxième objectif vise à modifier notre système de reconnaissance pour être capable de distinguer entre les chiffres prononcés par un locuteur fumeur ou non-fumeur. De plus, nous examinons la voix d'un utilisateur pour déterminer les effets du tabagisme sur les paramètres acoustiques tels que les formants, la hauteur, le Shimmer et la Jitter sur la base de 3 voyelles en langue amazighe (A, I, U). Notre troisième objectif est relatif à la construction d'un système de reconnaissance vocale automatique qui permettra de détecter les personnes atteintes des troubles de la voix. Nous adaptons et appliquons ces méthodes à la reconnaissance de phonèmes avec l'utilisation des échantillons vocaux des locuteurs malades. Notre quatrième objectif vise à prendre en considération l'effet des bruits sur la performance des systèmes de RAP précédents.

L'intérêt de présent travail c'est l'utilisation des formalismes statistiques Markovien pour modéliser les mots amazighs à base des unités phonétiques élémentaires. D'ailleurs, l'utilisation de cette approche a permis de produire un système reconnaissance puissant capable d'identifier la voix d'un fumeur et de diagnostiquer aussi des anomalies sur la base de l'approche d'apprentissage et des données vocales de la langue étudiée. De plus, Nous exploitons l'analyse des formants pour bien confirmer nos résultats de la reconnaissance vocale. Bien que l'approche statistique utilisée dans ce travail soit ancienne mais elle est largement appliquée dans le développement de systèmes RAP. L'application réussie des MMC à divers aspects de la modélisation de la parole a été soutenue par plusieurs études de recherche ces dernières années.

Cette thèse est organisée en sept chapitres. Dans le premier chapitre, nous allons mettre en évidence des généralités sur le signal vocal ainsi que la description des processus de production et de perception auditive de la parole.

Dans le deuxième chapitre, nous présenterons l'historique de la reconnaissance automatique de la parole et ses problèmes, ainsi que les différentes étapes constituant un tel système.

Le troisième chapitre, nous allons aborder les différentes bases des algorithmes constituant la théorie des modèles de Markov cachés (MMC), ainsi que les techniques utilisées dans la mise en œuvre des MMC dans la reconnaissance automatique de la parole.

Le quatrième chapitre sera consacré à l'étude des informations globales, des caractéristiques et des propriétés liées à la langue Amazighe au Maroc, dont la phonétique, le système d'écriture, la morphologie et l'utilisation informatique de Tifinagh.

Dans le cinquième chapitre, nous décrirons les différentes étapes que nous avons suivies pour la réalisation de notre système de la reconnaissance de la parole Amazigh, ainsi que la description de la base des données utilisée. D'autre part, nous présenterons nos expériences pour mieux adapter la langue Amazighe dans un système de reconnaissance vocale.

Dans le sixième chapitre, nous présenterons un système de reconnaissance de fumeurs et un système de diagnostic de la parole pathologique, ainsi que l'analyse spectrale d'onde vocale des fumeurs.

Le septième chapitre se concentrera sur l'analyse et l'évaluation des dix premiers chiffres amazighs dans des conditions bruyantes basées sur le rapport signal de bruit (SNR).

Chapitre 1: Généralités sur le signal vocal

1. Introduction

L'information transportée par le signal de parole peut être analysée sous différentes méthodes. A ce point, la distinction se fait pratiquement sous plusieurs niveaux de description non exclusifs à savoir le niveau phonétique, acoustique et bien d'autres.

Dans ce chapitre, nous allons décrire les processus consistants dans la production de parole et ses différentes étapes, les organes liés à cette production et aussi la perception auditive de la parole. Ensuite, nous allons discuter l'analyse qui s'axera sur le traitement de la conversion de la parole en signal électrique. Nous terminerons ce chapitre par l'aperçu de l'analyse du signal de parole et ses caractéristiques.

2. Production de la parole

Le constat s'apparente à bien des égards sur le plan de la parole qui, à son tour, peut être décrit comme le résultat efficient de l'action à la fois volontaire et coordonnée d'un certain nombre de muscles des appareils respiratoires et articulatoires (Boite et al. 1999). Ce va-et-vient concerne le contrôle du système nerveux central qui agrée, en permanence, des informations par rétroaction auditive et par les sensations kinesthésiques (Haton et al. 2006).

2.1. Le processus de production

D'abord, on peut affirmer que l'opération de la production de la parole est comme un système dans lequel une ou plusieurs sources excitent un certain nombre de cavités. La production de la source peut être faite soit au niveau des cordes vocales ou bien au niveau d'une constriction du canal vocal (Draper et al. 1959).

Premièrement, la source produit suivant une vibration quasi-périodique des cordes vocales ainsi elle produit une onde de débit quasi-périodique. Deuxièmement, la source sonore peut être un bruit de frottement ou bien un bruit d'explosion qui peut être vu dans le cas d'un fort abaissement dans le canal vocal ou si un brusque relâchement d'une occlusion du canal vocal a été produit. Les cavités situées après la glotte, nommées les cavités supraglottiques, vont ainsi être vivifiées par les sources et "filtrer" le son créé au niveau de ces sources. Pour cette raison, quand on change la forme de ces cavités, l'homme peut produire une différenciation au niveau des sons. Les producteurs de cette mobilité du conduit vocal sont souvent appelés les articulateurs.

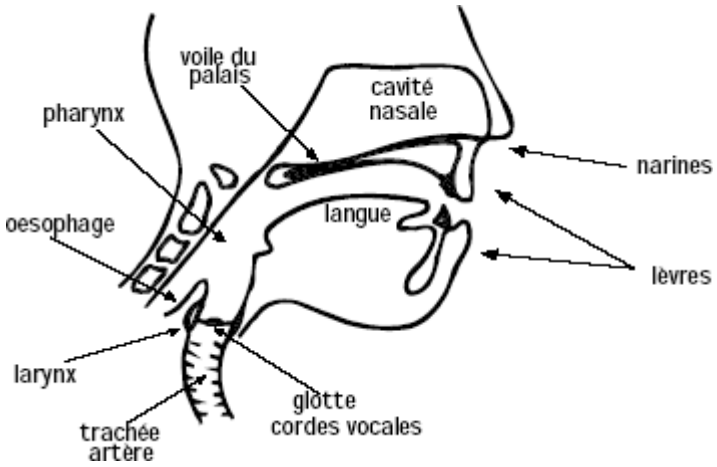


Figure.1. 1 : Coupe L'appareil phonatoire.

L'opération de production de la parole peut être résumée en trois étapes fondamentales:

- Naissance d'une source sonore via la génération d'un flux d'air.
- Production d'une source sonore sous la forme d'une onde quasi-périodique grâce à la vibration des cordes vocales ou bien sous la forme d'un bruit produit par une constriction, un relâchement ou bien par l'occlusion du canal vocal.
- La préparation des cavités supraglottiques afin d'avoir le son voulu.

2.2. Les différentes étapes de production de la parole

Le système respiratoire donne l'énergie essentielle à la production de sons, en envoyant de l'air via la trachée-artère. Au niveau de celle-ci, on trouve le larynx où la pression de l'air suit une modulation avant d'être accordée au conduit vocal. Le larynx est considéré comme un certain nombre de muscles et de cartilages mobiles qui encercle une cavité existant à la partie supérieure de la tranchée (figure.1.1) (Boite et al. 1999). Or, Les cordes vocales sont deux lèvres semblables situées au niveau du larynx. Pourtant, ces lèvres peuvent fermer définitivement le larynx et en s'écartant progressivement, trouver une ouverture triangulaire nommée glotte. Au moment de la respiration, l'air passe dans la glotte et la voix chuchotée, ainsi qu'au moment de la phonation des sons qui ne sont pas voisés. Enfin, les sons voisés se produisent grâce à la vibration périodique des cordes vocales.

Le larynx est en fait complètement fermé, ce qui permet d'accroître la pression des cordes vocales qui est en amont et les forces à s'ouvrir. Ce qui fait décroître la pression et permet aux cordes vocales de se refermer. Or, les impulsions périodiques de pression sont en fait accordées au conduit vocal contenant des cavités pharyngiennes et buccales pour la majorité des sons.

D'autre part, si la luette est placée en position inférieure, la cavité nasale vient de s'y insérer en dérivation. Dans ce qui suit, nous allons définir clairement les organes qui interviennent dans cette opération.

2.3. Les organes de production de la parole

D'une manière générale, la parole est produite par deux différents types de sources vocales. La première source, plus sonore, est celle qui naît au niveau du larynx grâce à la vibration des cordes vocales. La seconde source est moins sonore et prend naissance au niveau d'une constriction du canal vocal ou bien lors d'un relâchement ou d'une occlusion du canal vocal.

2.3.1. Le larynx

Le larynx est défini comme un organe situant dans le cou et joue un rôle important au niveau de la respiration et de la construction de la parole. Ensuite, le larynx (Voir figure.1.2) est plus précisément situé entre la trachée artère et le tube digestif sous la racine de la langue. La position de larynx change en fonction du sexe et de l'âge : Il s'abaisse de façon continue jusqu'à la phase de la puberté et il est aussi bien élevé chez la femme.

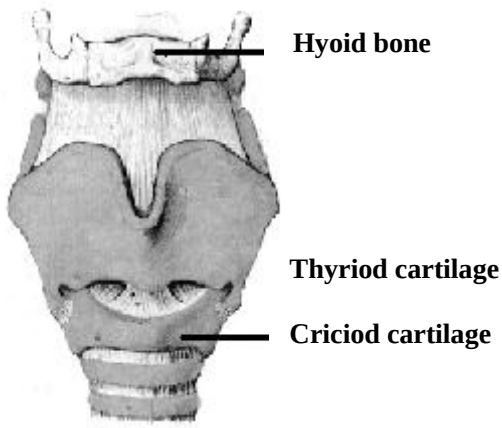


Figure.1. 2 : Schéma du larynx (Coleman J., 2001).

Il est caractérisé par un ensemble de cartilages entourés de tissus souples. Le côté le plus marquant du larynx est constitué de la thyroïde. Le côté antérieur du cartilage est souvent appelé la "pomme d'Adam". De plus, au-dessus du larynx, on trouve un os en forme de l'alphabet 'U' nommé l'os Hyoïde. Cet os se relie au larynx. Cet organe est en fait lié à la mâchoire par l'os Hyoïde grâce aux muscles qui jouent un rôle fondamental pour faire monter le larynx afin de réaliser la déglutition ou bien la construction de la parole.

Pourtant, le larynx est constitué au niveau de la partie inférieure par une collection de morceaux circulaires, le cricoïde, sous lequel on découvre les anneaux de la trachée artère.

Le larynx est ainsi caractérisé par trois fonctions principales. On peut les décrire comme suivant:

- Pendant la respiration, le larynx contrôle le flux d'air.
- Le larynx assure également la protection des voies respiratoires.
- La création d'une source sonore pour la parole.

2.3.1.1. Les muscles du larynx

Il existe deux ensembles de muscles permettant le contrôle des mouvements du larynx. Ainsi, on peut distinguer les différents muscles intrinsèques qui assurent le contrôle du mouvement des cordes vocales et celui des muscles dans le larynx, et également les muscles extrinsèques qui assurent de leur part le contrôle de la position du larynx au niveau du cou. La figure 1.3 représente les muscles intrinsèques.

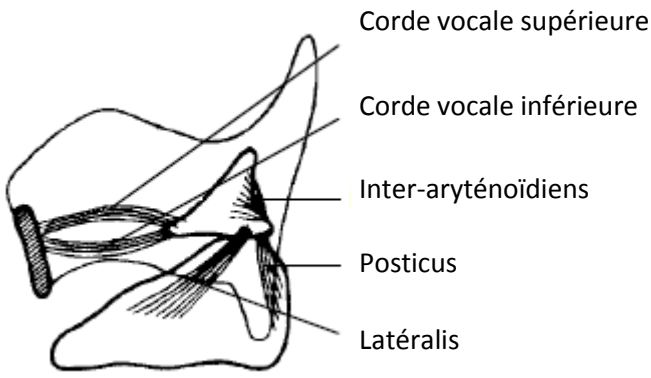


Figure.1. 3 : Schéma des muscles intrinsèques du larynx (Léothaud G., 2004).

2.3.1.2. Les cordes vocales

La production de la parole se réalise grâce aux cordes vocales situées au centre du larynx. De plus, les cordes vocales sont caractérisées par des muscles recouverts d'un tissu appelé la muqueuse. On trouve sur la partie arrière de chaque corde vocale une petite structure faite de cartilages « Les aryténoides », attachés à un ensemble de muscles permettant de les écarter pour garantir la respiration.

Au cours de la production de la parole, les aryténoides sont rapprochés. Les cordes vocales (voir figure 1.4) s'ouvrent et se ferment bien rapidement sous la pression de l'air provenant des poumons. Pour cela, si la pression soutenue de l'air d'expiration est maintenue, les cordes vocales vibrent et créent un son qui sera après modifié dans le canal vocal afin de donner lieu à un son voisé. Nous allons voir par la suite la description en détails de ce processus de vibration des cordes vocales.

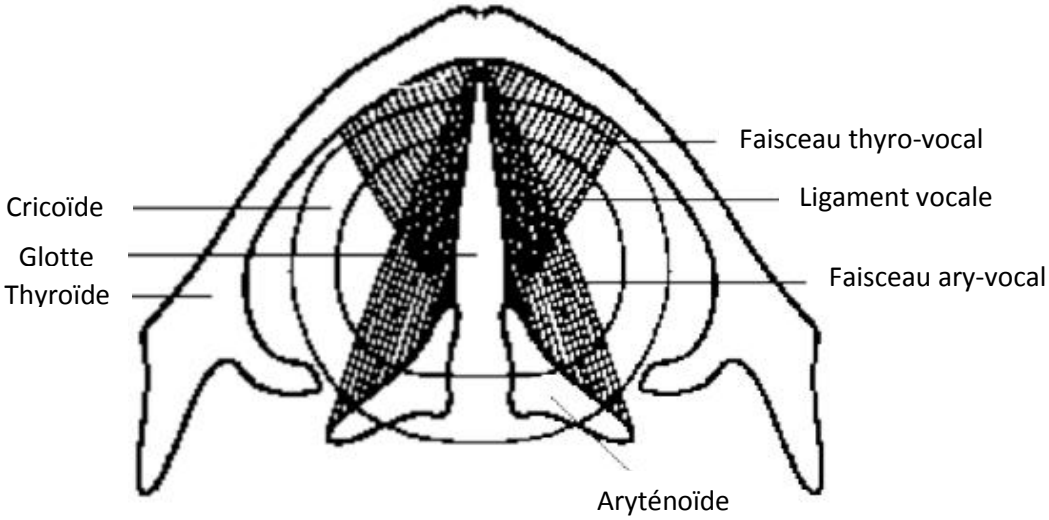


Figure.1. 4 : Structure de la corde vocale (Léothaud G., 2004).

Il y a tant de muscles qui permettent de fermer et de tendre les cordes vocales. Pourtant, les cordes vocales se constituent d'un muscle, le thyroaryténoide. De plus, le muscle nommé l'interaryténoide, aide à rapprocher ces deux cartilages. Enfin, le muscle appelé cricoaryténoide se situe entre l'aryténoide et le cartilage cricoïde permettant la fermeture du larynx. Le muscle cricothyroide se caractérise par le fait qu'il aille du cartilage cricoïde à cartilage thyroïde. On se contactant, le cartilage cricoïde bascule en avant et tend les cordes vocales, ce qui amène à un élevage de la voix.

Les muscles extrinsèques assurent l'augmentation ou l'abaissement de larynx dans sa globalité mais n'affectent pas le mouvement des cordes vocales.

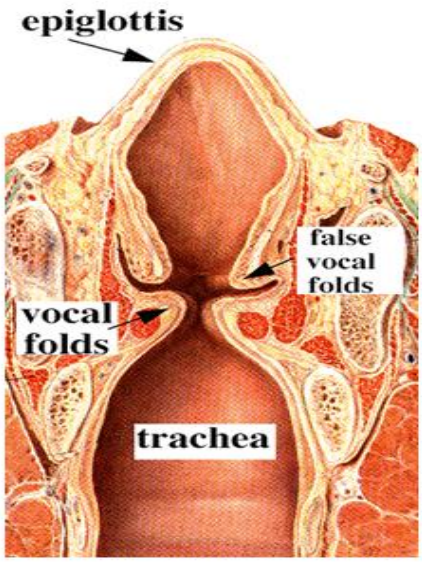


Figure.1. 5 : Vue longitudinale du larynx (Coleman J., 2001).

Dans la figure 1.5, se présente le schéma d'une coupe verticale du larynx. Dans ce schéma, les cordes vocales sont séparées, pendant la respiration. Or, on peut aussi constater au-dessus des cordes vocales, la présence des tissus qui ont comme principal rôle l'empêchement du passage de substances dans la trachée en cours de la déglutition : C'est ce qu'on appelle les fausses cordes vocales. D'ailleurs, on doit noter que ces cordes vocales ne jouent aucun rôle lors de la phonation. En effet, le cartilage appelé l'épiglotte qui se trouve au-dessus de la langue a également un rôle permettant la protection de l'accès de la trachée au moment de la déglutition.

2.3.2. Les cavités supraglottiques

La production du son nécessite l'intervention même à un degré moindre de certains organes situés au-dessus des glottes (organes supraglottiques) (Ghio et al. 2007). On peut donc distinguer:

- **Le conduit vocal**

Il est considéré comme un tube sonore de section variable qui va de la glotte jusqu'aux lèvres. En effet, le conduit vocal mesure pour un adulte environ 17 cm. De plus, la variation de sa forme dépend du mouvement des articulateurs qui consistent dans les lèvres, la langue, la mâchoire et le velum. Dans ce qui suit, nous allons montrer la description des articulateurs.

- **Le conduit nasal**

Ce conduit est considéré comme un passage auxiliaire pour la transmission du son. Il commence depuis le velum et se termine au niveau des fosses nasales. Cette cavité mesure environ 12 cm pour un homme adulte. La figure 1.1 montre le contrôle du couplage acoustique entre les deux cavités par l'ouverture au niveau du velum qui est largement ouvert. Dans ce cas, on aura la production d'un son nasal. Dans le cas inverse, quand le velum ferme le conduit nasal, le son produit est donc nommé non-nasal.

Enfin, les autres organes nommés articulateurs jouent un rôle fondamental, chacun en ce qui le concerne. Ces articulateurs sont alors:

- **La langue**

Cet articulateur est considéré comme une structure frontière qui appartient en même temps à la cavité buccale au niveau de la partie mobile et au glosso-pharynx au niveau de la partie fixe. Elle est appliquée contre les dents qui constituent de leur part un organe vibratoire

accessoire, intervenant dans la formation des consonnes. Elle joue donc un rôle important pour la phonation. Ainsi, il est bien clair que la langue est un articulateur principal grâce à sa position qui est déterminante dans le conduit vocal.

- **La mâchoire**

D'abord, la mâchoire ne possède pas un degré élevé de liberté comme la langue, car elle ne peut pas se déformer comme cette dernière. En outre, la mâchoire peut en plus de s'ouvrir et de se fermer, de s'avancer ou d'effectuer des mouvements rotationnels. Le rôle de la mâchoire dans la parole n'est pas principal dans la mesure où il est possible en bloquant cette dernière de parler de façon très claire.

- **Les lèvres**

Dans cette section, nous avançons que les lèvres, situées à la fin du conduit vocal, se caractérisent comme dans le cas de la langue par une mobilité considérable grâce aux différents muscles impliqués dans leur contrôle. Or, les lèvres, supérieure et inférieure, contiennent des points de jonction appelées les commissures. Elles jouent de leur part un rôle très important surtout dans le cas de sourire.

L'espace intérolabial est le plus important au niveau acoustique. Concernant la phonation, on peut observer des mouvements divers, dont:

- L'occlusion (lorsque les lèvres sont fermées)
- La protrusion (lorsque les lèvres sont avancées vers l'avant)
- La montée et la descente de la lèvre inférieure
- L'allongement, la descente ou la montée des commissures

2.4. Les sons de la parole par l'approche de la production des sons

Dans cette partie, nous allons parler des différentes classes de sons au niveau phonétique (Marchal et al. 1980) en expliquant comment ces sons se créent.

2.4.1. Notions de phonétique

D'une manière générale, la parole se compose par un nombre finis d'éléments sonores distinctifs. Or, ces différents éléments composent les unités linguistiques élémentaires et possèdent la propriété de changer le sens d'un mot. En effet, ces unités linguistiques élémentaires sont nommées phonèmes (Bekesy G., 1960).

Généralement, le phonème est défini comme la plus petite unité phonique fonctionnelle. De plus, le phonème n'est pas défini sur un plan articulatoire, acoustique, ou perceptuel, mais sur le plan fonctionnel. De plus, les phonèmes ne se caractérisent pas par une existence indépendante, c'est-à-dire qu'ils constituent un ensemble bien structuré dont chaque élément est différent des autres. Cependant, l'établissement de la liste des phonèmes pour la plupart des langues européennes a été faite dès la fin du 19ème siècle. D'ailleurs, les phonèmes sont vus en telle sorte comme des éléments pour le codage de l'information linguistique. En revanche, ces phonèmes peuvent être présentés en groupes de classes de telle façon qu'elles partagent des mêmes caractéristiques. C'est ce qu'on appelle "traits distinctifs".

- **Trait distinctif** : Il s'agit d'une expression de similarité au niveau articulatoire, perceptif ou acoustique des sons concernés.

  Par exemple, pour les voyelles, on distinguera 4 traits distinctifs :

  - La nasalité : Le conduit vocal et le conduit nasal prononcent la voyelle grâce à l'ouverture du velum.
  - Le degré d'ouverture du canal vocal.
  - La position de la constriction principale du canal vocal. Cette constriction se réalise entre le palais et la langue.
  - La protrusion des lèvres.

Pour leur part, les consonnes sont classées selon trois traits distinctifs :

- Le voisement : Grâce à une vibration des cordes vocales, la consonne se prononce.
- Le mode d'articulation dont les modes fricatif, occlusif, nasal, liquide ou glissant.
- La position de la constriction principale du conduit, généralement appelée lieu d'articulation qui n'est pas forcement atteinte avec le corps de la langue.

De surcroît, les phonèmes sont considérés comme des éléments abstraits liés à des sons élémentaires. De plus, les phonèmes ne sont pas identiques pour toutes les langues. Le /a/ du français n'est pas le même que le /a/ de l'anglais. Pour ce motif, est née l'idée de définir un alphabet phonétique international (alphabet IPA) (Durrand J., 2009) qui assure la description des sons et les prononciations de ces sons de manière universelle.

On trouve d'autres façons d'organiser les sons, par exemple on fait opposer, les consonnes nasales, les sons sonnants (voyelles), les glissantes ou liquides aux sons obstruant «

occlusives, fricatives ».

2.4.2. Les voyelles

Les voyelles se produisent à l'aide des vibrations des cordes vocales. Le son des voyelles est donc obtenu en changeant la forme du conduit vocal grâce à des différents articulateurs. D'ailleurs la forme du canal vocal en un mode d'articulation normal est maintenue relativement stable au cours de la durée de la voyelle.

2.4.3. Les consonnes

Les consonnes se regroupent en traits distinctifs de la même manière que les voyelles. Par contre, les consonnes ne sont pas exclusivement voisées comme les voyelles et ne sont pas nécessairement réalisées avec une configuration stable du canal vocal.

- **Les consonnes voisées**

On parlera de consonnes voisées, lesquelles sont produites avec une vibration des cordes vocales. En plus du voisement, lorsqu'une source de bruit est due à une constriction du canal vocal, on parlera ainsi des consonnes à excitation mixte.

- **Les fricatives**

Les fricatives sont produites par un flux d'air rude né au niveau d'une constriction du canal vocal. Par la suite, nous allons montrer des différentes fricatives suivant le lieu de cette constriction principale :

- Les labio-dentales, pour une constriction qui se fait entre les lèvres et les dents « comme avec l'alphabet /f/ français dans le mot "foyer" ».
- Les dentales, pour une constriction qui se réalise sur des dents « comme avec l'alphabet /t/ anglais dans le mot "think" ».
- Les alvéolaires, pour une constriction qui vient derrière les dents « comme avec l'alphabet /s/ dans le mot "son" ».

Pourtant, on trouve plusieurs langues dont quasiment tous les points d'articulations du canal vocal peuvent être exploités pour la réalisation des fricatives.

Cela présente une des difficultés de l'apprentissage des langues étrangères pour la raison qu'il n'est pas facile d'apprendre à construire des sons exigeant le positionnement de la langue à des endroits inhabituels.

- **Les plosives**

Les plosives connaissent un dynamisme considérable du canal vocal. Leur réalisation se fait à travers le canal vocal en un endroit. De plus, l'air qui vient des poumons crée une pression derrière cette occlusion qui est par la suite relâchée grâce au mouvement rapide des articulateurs ayant réalisé cette occlusion. Pour les fricatives, on trouve la même chose, telle que l'un des traits distinctifs entre le lieu d'articulation et les plosives. Concernant les plosives, on distingue:

- Les labiales, pour une occlusion réalisée sur les lèvres.
- Les dentales, pour une occlusion réalisée sur les dents.
- Les vélo-palatales, pour une occlusion réalisée au niveau du palais.

En plus du lieu d'articulation, les plosives peuvent également être voisées ou non voisées.

- **Les consonnes nasales**

Les consonnes nasales sont généralement voisées et produites par la réalisation d'une occlusion complète du canal vocal et également par l'ouverture du vélum qui assure au conduit nasal d'être un résonateur unique. De même que précédemment et suivant le lieu d'articulation, on aura les points suivants :

- Les labiales, pour une occlusion, étant réalisées sur les lèvres.
- Les dentales, pour une occlusion, étant réalisées sur les dents.
- Les vélo-palatales, pour une occlusion, étant réalisées au niveau du palais.

- **Les glissantes et les liquides**

Cette classe contient des sons qui ressemblent aux voyelles. De plus, les liquides sont, entre autres, nommés semi consonnes ou bien semi-voyelles. Or, les liquides et les glissantes, sont généralement, voisées et non nasales.

Pourtant, les glissantes sont caractérisées par leur mouvement et précédent toujours une voyelle ou bien un son vocalique. D'autre part, les liquides ou encore appelés semi-voyelles, sont très similaires aux voyelles mais avec une constriction conséquente et un apex de la langue aussi bien relevé.

3. Traitement de la parole

Le traitement du signal est en général défini comme un ensemble de techniques et de méthodes agissant sur un signal électrique dans le but d'en extraire l'information souhaitée (Calliope et al. 1989). En outre, ce signal doit en fait traduire le phénomène physique à étudier. Au niveau physique, la parole apparaît comme étant une variation de l'air produite et émise via le système articulatoire. Il s'agit d'un phénomène physique acoustique prenant une forme analogique. Ce signal est étudié par la phonétique acoustique en le transformant en signal électrique en se basant sur le transducteur approprié : Le microphone.

Actuellement, le signal électrique produit est le plus souvent numérisé. De plus, ce signal peut en fait être soumis à un certain nombre de traitements afin d'en extraire les informations nécessaires et les paramètres pertinents en liaison avec l'application. Ainsi, la conversion d'un tel phénomène de parole en un signal électrique exige les opérations suivantes (Boite et al. 1999).

3.1. Numérisation

D'abord, l'importance des systèmes numériques de traitement de l'information suit toujours le chemin de croissance (radio, télévision, téléphone…). Cela est justifié par les points positifs qu'ils connaissent, à savoir la stabilité considérable des paramètres, ainsi que l'excellence reproductibilité des fonctionnalités et des résultats ajoutés. D'ailleurs, la réalisation de la numérisation du signal de parole se fait par un convertisseur analogique-numérique (CAN). De plus, la conversion analogique numérique est le résultat de trois effets sur le signal analogique.

3.2. L'échantillonnage

Pourtant, l'échantillonnage se base sur le principe de transformation du signal à temps continu $x(t)$ en signal à temps discret $x(nT_e)$ défini aux moments d'échantillonnage, multiples entiers de la période d'échantillonnage T_e qui est de sa part l'inverse de la fréquence d'échantillonnage f_e.

Concernant le signal vocal, on choisit le T_e grâce au résultat d'un compromis. De plus, son spectre peut atteindre jusqu'à 12 kHz. Ainsi, et pour satisfaire au théorème de Shannon, il faut qu'on choisisse une fréquence f_e égale au moins à 24 kHz. Toutefois, le coût d'un traitement numérique, transmission, filtrage, ou bien l'enregistrement peut être réduit d'une façon considérable si on suit une limitation du spectre via un filtrage préalable. Il s'agit du rôle de

filtre de garde, telle que la fréquence de coupure f_c, prise en fonction de la fréquence d'échantillonnage retenue.

3.3. La Quantification

Dans cette étape, nous allons nous intéresser à l'approximation des valeurs réelles des échantillons via une échelle de *n* niveaux nommés échelle de quantification.

En outre, la quantification prend un nombre fini 2^n de valeurs espacées du pas de quantification δ, et le signal numérique résultant est noté *x(n)*. En effet, la quantification crée un bruit blanc qui est une erreur de quantification. Le pas de quantification est alors imposé par le rapport signal à bruit qu'on doit garantir. Ensuite, on adopte pour la transmission téléphonique une loi de quantification logarithmique dont chacun des échantillons est représenté sur 8 bits. D'autre part, la quantification du signal musical nécessite de sa part une quantification linéaire sur 16 bits.

3.4. Le Codage

Le codage de la parole est le processus d'obtention d'une représentation compacte des signaux vocaux pour une transmission efficace sur des canaux câblés et sans fil à bande limitée et / ou à un stockage. Aujourd'hui, les codeurs vocaux sont devenus des composants essentiels des télécommunications et de l'infrastructure multimédia. Les systèmes commerciaux qui reposent sur un codage vocal efficace comprennent la communication cellulaire, le protocole voix sur Internet (Voice over IP-VoIP), la vidéoconférence, les jouets électroniques, l'archivage, la voix et les données numériques simultanées (Digital Simultaneous Voice and Data-DSVD), ainsi que de nombreux jeux sur PC et d'applications multimédias.

Le codage de la parole est l'art de créer une représentation à redondance minimale du signal de parole qui peut être efficacement transmise ou stockée sur des supports numériques, et de décoder le signal avec la meilleure qualité de perception possible. Comme tout autre signal à temps continu, la parole peut être représentée numériquement par les processus d'échantillonnage et de quantification; la parole est généralement quantifiée en utilisant soit une quantification uniforme 16 bits, soit une quantification commandée 8 bits. Comme de nombreux autres signaux, cependant, un signal de parole échantillonnée contient une grande quantité d'informations qui sont soit redondantes (informations mutuelles non nulles entre les échantillons successifs du signal), soit non pertinentes sur le plan perceptuel (informations qui

ne sont pas perçues par les auditeurs humains). La plupart des codeurs de télécommunications sont avec perte, ce qui signifie que la parole synthétisée est perpétuellement similaire à l'original mais peut-être physiquement différente.

4. Analyse du signal de parole

Après avoir été numérisé, le signal de parole peut être traité de différentes façons en fonction des objectifs souhaités. Il existe de diverses techniques. Dans ce qui suit, nous allons présenter les outils liés au signal de parole.

4.1. Analyse temporelle

Le signal de parole est un signal quasi-stationnaire. De plus, sur un intervalle temporel supérieur, il est bien évident que les caractéristiques du signal se développent par les sons prononcés comme indiqué dans la figure ci-après. La figure 1.6 représente le signal de la parole des mots amazighs suivants « Sdes » et « Rrzam » avec son évolution temporelle. On y constate une alternance de zones assez périodiques et de zones bruitées, appelées zones voisées et non voisées.

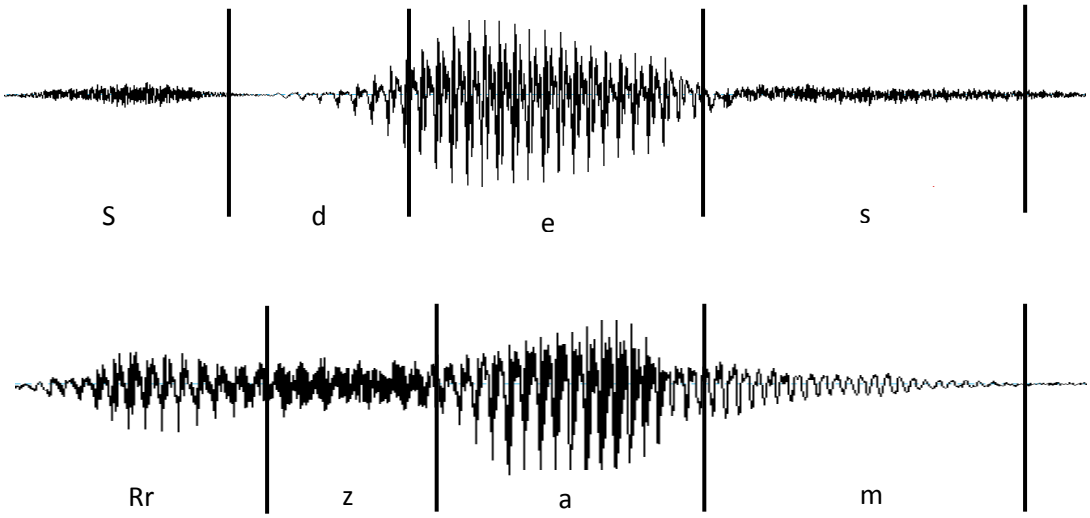


Figure.1. 6 : Représentation temporelle (Audiogramme) des Sdes et Rrzam prononcé par un locuteur marocain amazigh.

Pour étudier le signal de parole, nous commençons par une première approche qui vise à observer la forme temporelle du signal. On peut alors déduire un certain nombre de propriétés pouvant être utilisées pour le traitement de parole. Il est alors bien clair de distinguer les parties voisées, dans lesquelles on peut rencontrer une forme d'onde quasi-périodique, des parties non voisées dont un signal aléatoire de faible amplitude est observé. Conformément,

on peut observer que les petites amplitudes sont plus représentées que les grandes, ce qui démontre les choix faits en codage de la parole comme indiqué dans la figure 1.7:

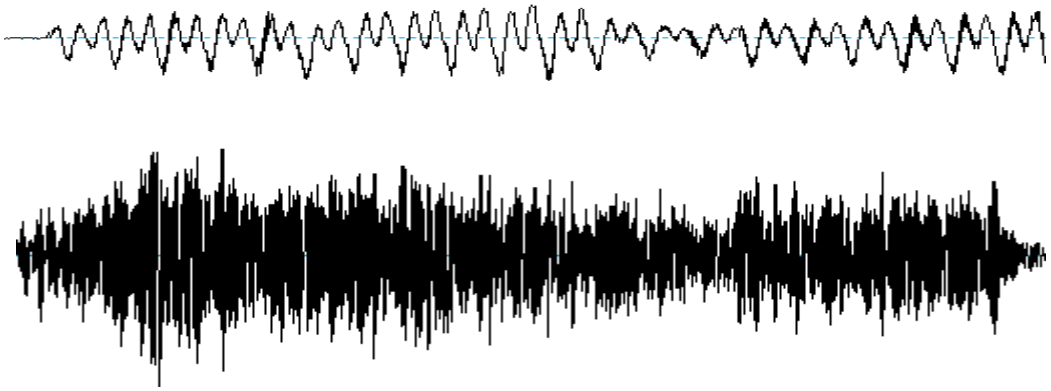


Figure.1. 7 : Exemple de son voisé (haut) et non voisé (bas).

4.2. Analyse fréquentielle

Pour caractériser et représenter le signal de parole, on utilise une seconde approche qui consiste dans la représentation spectrale.

Les méthodes du traitement du signal peuvent être classées en deux grandes catégories comme il est indiqué ci-dessous :

- **Les méthodes générales:** Ce sont des méthodes valables pour les signaux évolutifs dans le temps, particulièrement pour les analyses spectrales.

- **Les méthodes qui se réfèrent à un modèle:** Tel qu'un modèle de production du signal vocal ou celui d'audition.

- **Méthodes générales**

D'abord, les méthodes spectrales prennent une place majeure dans l'analyse de la parole. L'oreille fait une analyse fréquentielle pour le signal qu'elle reçoit, par la suite, les sons de la parole peuvent être bien décrits en fonction des fréquences. Or, la transformée de Fourier assure l'obtention du spectre d'un signal, surtout son spectre fréquentiel, c'est-à-dire sa représentation amplitude-fréquence.

La figure 1.8 suivante montre clairement la transformée de Fourier dans le cas d'une tranche non voisée et aussi d'une tranche voisée. Les parties voisées du signal se représentent en forme de successions des photos spectrales marquées, dont les fréquences centrales sont

multipliées de la fréquence fondamentale. Pourtant, le spectre d'un signal non voisé ne montre aucune structure particulière. De plus, la forme générale de ces spectres, nommée enveloppe spectrale, montre de sa part des pics et des creux qui correspondent aux résonances et aux antirésonances du conduit vocal et sont appelés formants et anti-formants.

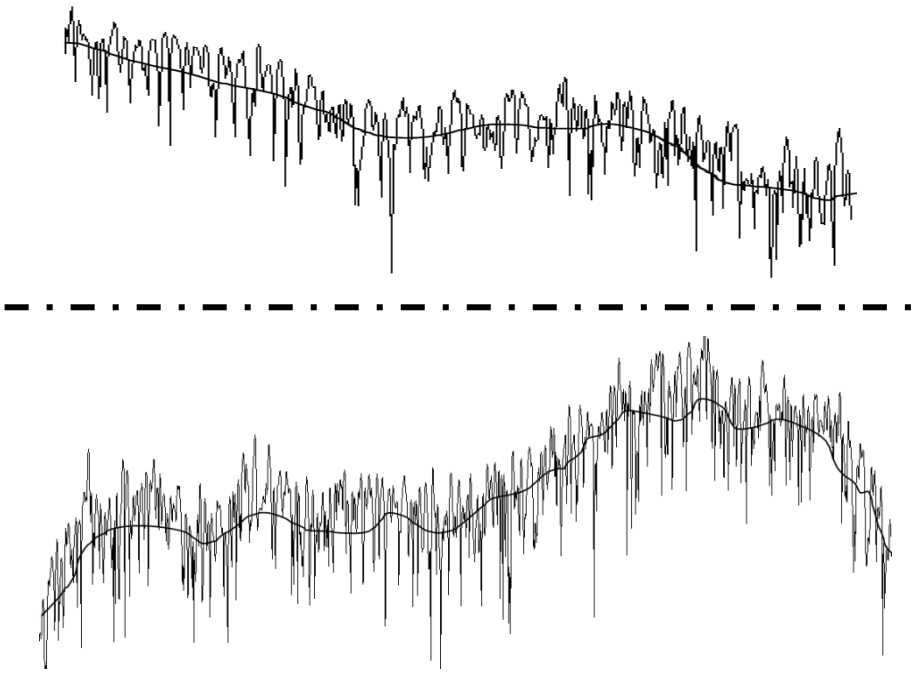


Figure.1. 8 : Evolution de la transformée de Fourier discrète du [a] et du [s] de chiffre amazigh « Sa ».

La parole est considérée comme un phénomène non stationnaire. Il importe de faire intervenir le temps comme étant la troisième variable dans la représentation. En outre, le spectrogramme est la représentation la plus répandue du tout.

- **Spectrogramme**

C'est une représentation tridimensionnelle d'un son dont l'énergie par bande de fréquences est représentée en fonction du temps (Ono et al. 2008), où le temps est représenté de sa part sur l'axe *X*, la fréquence figure sur l'axe *Y* et le niveau lié à chaque fréquence se représente sur l'axe *Z*.

D'autre part, le spectrogramme est déterminé par le module de la transformée de Fourier discrète qui est calculée sur une fenêtre temporelle plus ou moins longue. Ensuite, la transformée de Fourier discrète *TFD X(k)* de la première fenêtre du signal de parole *x(n)* est

représentée par l'équation suivant:

$$X_i\,(K)=\sum_{n=0}^{N-1} x(n)\, e^{-j2\pi nk/N} \qquad \text{(eq. 1.1)}$$

L'amplitude du spectre *y* est représentée sous la forme de niveaux de gris dans un diagramme bidimensionnel de la fréquence temporelle, comme on peut le voir dans les spectres de la figure 1.9. On parle d'un spectrogramme de large bande ou de bande étroite en fonction de la durée de la fenêtre de pondération. Les spectrogrammes à large bande sont trouvés avec des fenêtres de pondération de courte durée. Elles mettent en évidence l'enveloppe spectrale (les formants) du signal, les périodes voisées y apparaissent en forme de bandes verticales plus sombres. Les spectrogrammes à bande étroite sont obtenus à travers les fenêtres de l'ordre de 30 à 40 ms. De plus, ils offrent une bonne de résolution au niveau fréquentiel, les harmoniques du signal dans les zones voisées y apparaissent sous la forme de bandes horizontales.

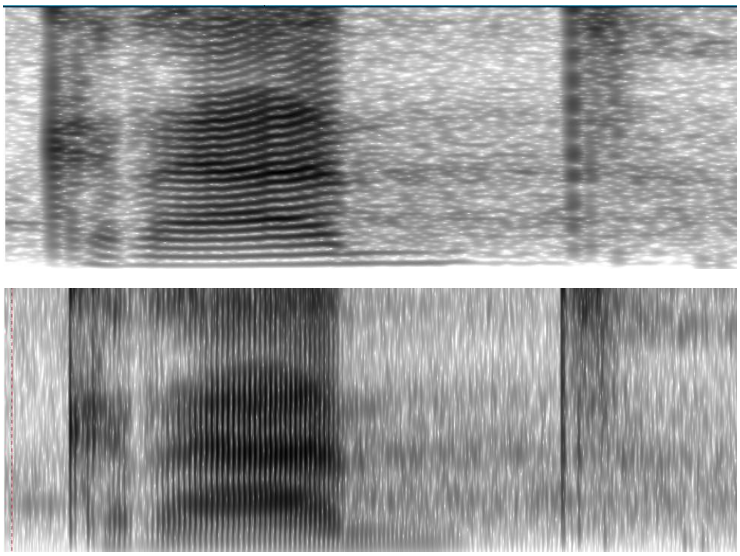


Figure.1. 9 : Spectrogramme à large bande (en bas), à bande étroite (en haut), et évolution temporelle du chiffre Amazigh « Krad », échantillonnée à 11.25 kHz (calcul avec fenêtre de Hamming de 10 et 30 ms respectivement).

5. Caractéristique du signal de parole

Le signal de parole est considéré comme un vecteur acoustique porteur d'informations

d’une grande complexité.

5.1. Traits acoustiques

Les traits acoustiques du signal de parole dépendent de sa production.

- ✓ **La fréquence fondamentale (Pitch)**

Il s'agit du premier trait acoustique et également de la fréquence de vibration des cordes vocales. Pour les sons voisés, la fréquence fondamentale correspond à la fréquence du cycle d’ouverture/fermeture des cordes vocales (Wertzner et al. 2005). Il est exprimé en hertz. Il correspond à la hauteur et produit des basses ou des aigues. Varie selon la longueur, l'épaisseur et la tension des cordes vocales. Il augmente lorsque la pression augmente sous le hautbois, et la gorge monte dans le cou en raccourcissant les dimensions du pharynx, entraînant une augmentation de la tension et de la longueur des cordes vocales. Il est mesuré à l'aide d'un logiciel informatique. Cité par (Nicholas et al. 2007) ont montré que la fréquence primaire diminue avec l'âge. Physiquement, la fréquence de base F0 correspond à la fréquence de vibration des cordes vocales sous l'influence de l'air traversant la glotte. Il existe une fonction spécifique à chaque locuteur qui varie en fonction de son âge et de son sexe (Wolfgang H., 1984) :

- De 50 – 500 Hz pour les hommes
- De 100 – 700 Hz pour les femmes
- De 200 – 600 Hz pour les enfants

Comme présentée dans (Roach P., 2010), "Aucune définition de la conscience n'est entièrement satisfaisante, mais toute tentative de définition doit avouer que la hauteur de la voix joue un rôle fondamental." Les variations de F0 sont fréquemment connectées aux phénomènes d'intonation, mais la compression et le rythme ainsi que divers facteurs, ont également un rôle dans la définition du graphique F0 du graphique. Ces modifications de F0 ou événements semblent se produire à différents niveaux de description. Au premier niveau, il est souvent appelé localement, dont certains semblent affecter les syllabes ou les groupes de syllabes. Cependant, d'autres événements F0 semblent affecter des unités plus grandes, telles que les phrases d'intonation ou même les phrases ou les paragraphes. Ces types d'événements sont souvent appelés globalement (Khalifa M., 2017).

✓ **Intensité**

L’intensité d’un son, appelée aussi volume (Wertzner et al. 2005), permet de distinguer un son fort d’un son faible. Elle correspond à l’amplitude de l’onde acoustique. Pour le son, onde de compression, cette grandeur est la pression. La source sonore crée une puissance acoustique dans tout l'espace environnant. Plus vous vous éloignez de la source sonore, plus la force acoustique est faible, c'est-à-dire l'intensité du son. Alexander Graham Bell a eu l'idée d'utiliser une échelle avec son nom : Bel. C'est le décibel (décibel) qui est généralement utilisé pour détailler la réalité de ce que l'oreille humaine réalise.

L'intensité est comme la hauteur et l'oreille humaine a une idée subjective et objective (dans la langue anglaise, il y a deux termes différents pour parler de ces deux types d’intensité : "loudness" pour la perception et "intensité" pour la puissance du son). En effet, l'oreille n'a pas la même sensibilité selon la fréquence entendue. Par exemple, un audio de 50 dB génère une sensation auditive plus forte lorsque sa fréquence est de 1000 Hz que lorsqu'elle est de 100 Hz.

✓ **La mesure de perturbation de la fréquence fondamental (Jitter)**

Jitter est une mesure des perturbations de la fréquence de base du signal audio. Nous nous plaçons ici au niveau de la période sonore, donc du cycle vibratoire, et notons les différences de durée entre une ou plusieurs périodes. Il décrit une perturbation acoustique, survenant selon les auteurs du fait d'un défaut neurologique, aérodynamique ou biomécanique (Farrús et al. 2007). Le jitter se calcule comme le rapport entre la moyenne de toutes les différences de durées entre deux cycles glottiques successifs et la durée moyenne d’un cycle.

✓ **La mesure de la perturbation de l’amplitude (Shimmer)**

Shimmer est une mesure des perturbations de l’amplitude qui signifie la réduction de la résistance glottale et des lésions de masse sur les cordes vocales. Elle est corrélée avec la présence d'émissions sonores et de respirations (Teixeira et al. 2013). La moyenne des différences entre l'amplitude maximale de deux cycles glottiques successifs est divisée par la moyenne des amplitudes maximales de chaque cycle.

✓ **Le spectre de fréquence**

C’est le deuxième trait acoustique qui dépend principalement du timbre de la voix. Il est obtenu par le filtrage dynamique du signal en provenance du larynx ou signal glottique par le conduit vocal.

- ✓ **Le timbre**

Le timbre est une caractéristique qui assure l'identification d'une personne à la simple écoute de sa voix. Il résulte particulièrement par la résonance dans la poitrine, la gorge, la cavité buccale et le nez. Le timbre est lié principalement à la corrélation entre la fréquence fondamentale et les harmoniques qui sont les multiples de cette fréquence.

- ✓ **L'énergie**

C'est le dernier trait acoustique lié à l'intensité sonore. L'énergie de la parole correspond à la pression de l'air en amant du larynx. Elle est par la suite plus forte pour les segments voisés de la parole que pour les segments non voisés.

- ✓ **Les formants**

Le spectre du signal vocal, résultant de l'action des sources de sons sur le conduit vocal, présente des maximums et des minimums qui correspondent aux résonances et aux antirésonances du conduit vocal, appelés formants (Coleman, 1971) et anti-formants. Du point de vue perceptif, seuls les trois premiers formants jouent un rôle essentiel pour caractériser le spectre vocal. On peut caractériser toute voyelle en n'utilisant que ses trois premiers formants. En général la fréquence du premier formant varie de 200 à 900 Hz, celle du second de 500 à 2500 Hz et le troisième se situe entre 1500 et 3500 Hz. Des formants d'ordre supérieur existent même si leur rôle sur le plan perceptif est limité, ils contribuent à caractériser la voix.

5.2. Méthodes avec modélisation

Dans cette catégorie, les méthodes dites de Codage Prédictif Linéaire LPC ont été largement utilisées pour l'analyse de la parole. Elles font référence à un modèle du système de phonation que l'on représente en général comme un tuyau sonore à section variable. L'analyse LPC est utilisée essentiellement en codage et en synthèse de la parole.

5.3. Méthodes cepstrales

L'analyse cepstrale (Haton J., 1992) est une méthode d'analyse du signal vocal fondée sur une modélisation. Elle est actuellement très répandue en reconnaissance automatique de la parole. La plupart des systèmes actuels de reconnaissance de parole utilisent un ensemble de paramètres appelés MFCC (Mel Frequency Cepstrum Coefficients) dont le principe d'obtention repose sur l'analyse cepstrale.

Cette méthode, appelée aussi analyse homomorphique, a pour but de séparer dans le signal vocal les contributions respectives de la source du signal à savoir la vibration des cordes vocales et du conduit vocal dont les fréquences de résonance conduisent notamment aux formants des voyelles, comme il est illustré dans la figure ci-dessous.

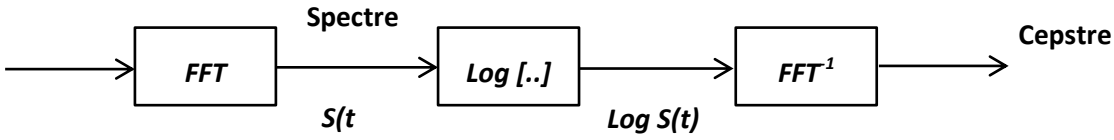


Figure.1. 10 : Principe de l'analyse homomorphique.

Les coefficients obtenus par MFCC sont robustes car, d'une part, ils assurent, comme il vient d'être dit, une séparation entre la fonction de transfert du conduit vocal et les caractéristiques du fondamental de la voix, et d'autre part, ils sont peu sensibles à la puissance acoustique du signal analysé.

5.4. Modèles d'oreille

Une famille de méthodes d'analyse de parole s'inspire des données de la psycho-acoustique et de la physiologie de l'audition humaine telles que courbes d'isotonie, bandes critiques de l'oreille, phénomènes non linéaires (saturation, masquage de sons, etc.), contrôle de gain, filtrage cochléaire, etc. Les modèles d'oreille (Caelen J., 1979), sont utilisés pour obtenir une représentation fréquentielle de la parole. On les trouve dans des systèmes de reconnaissance de parole, notamment en présence de bruits.

5.5. Analyse perceptive

En présence de bruit important, les méthodes d'analyse traditionnelles ont du mal à extraire les caractéristiques représentatives de la parole. De nombreuses méthodes ont été proposées pour améliorer cette situation. Elles se fondent sur différentes méthodes, notamment sur des propriétés de la perception auditive (Haton J., 1992). Un bon exemple est l'analyse RASTA-PLP, utilisée avec succès en reconnaissance de parole dans du bruit. Cette méthode intègre plusieurs opérations inspirées de données perceptives.

5.6. Analyse par ondelettes

Parmi les travaux menés pour améliorer les techniques d'analyse de signaux, l'analyse par ondelettes (Haton et al. 2006), présente un intérêt certain. Ce type d'analyse permet

d’obtenir une représentation temps-fréquence locale d’un signal comme alternative au spectre de Fourier. L’intérêt, pour des signaux non stationnaires comme la parole, est de pouvoir mener une analyse multi-résolution des phénomènes correspondant à des échelles de temps et de fréquence différentes. L’analyse par ondelettes a été appliquée à de nombreux types de signaux (biomédicaux, sismiques, etc.). Dans le cas de la parole, les applications actuelles concernent la synthèse, le codage, la suppression de bruit, etc. Peu de travaux ont trait à la reconnaissance.

6. Conclusion

Dans ce chapitre nous avons vu le mécanisme de production de la parole, le principe de son audition, le son de parole au niveau phonétique et les éléments qui permettent de produire la voix, ainsi que les caractéristiques générales du signal vocal. En effet, il existe deux types de son : Les sons voisés résultants de la vibration des cordes vocales et les sons non voisés qui ne nécessitent pas l'intervention du larynx.

Après avoir présenté les caractéristiques de base utilisées en traitement des signaux acoustiques. Nous allons décrire dans le chapitre suivant, la structure d’un système de reconnaissance de la parole, les méthodes d'extraction et l'évaluation de ces performances. Nous allons de plus, explorer les différentes approches liées au développement des systèmes de reconnaissance de la parole.

Chapitre 2: La Reconnaissance Automatique de la Parole

1. Introduction

La parole est l'un des moyens les plus directs d'échange de l'information utilisés par l'homme. Ceci a motivé plusieurs chercheurs à travers le monde à concevoir des systèmes capables de reconnaitre les mots parlés par un être humain ouvrant la voie de communication de l'homme avec la machine.

Un système de Reconnaissance Automatique de la Parole (RAP) est un système qui a la capacité de détecter et d'analyser la parole dans le but de générer une chaîne de mots ou de phonèmes représentant ce que la personne a prononcé. Cette analyse se fonde sur l'extraction des paramètres descriptifs de la parole. Cependant le signal de parole ne contient pas seulement des informations sur le texte parlé mais aussi des informations sur le locuteur, la langue, les émotions dont leur extraction n'est pas l'objectif de la RAP. Cette thèse s'intéresse à une étape primordiale de la RAP permettant l'analyse des différents types de la voix (fumeurs - pathologies - milieu bruités). Dans ce chapitre, nous allons présenter l'historique de la reconnaissance automatique de la parole. Ensuite, nous allons parler des problèmes, des principes généraux et des différentes étapes d'un système RAP. Enfin, nous aborderons la méthode d'évaluation due à ce système.

2. Bref historique de la Reconnaissance de la parole

Dans cette section, nous allons présenter l'historique de la reconnaissance de la parole et nous allons discuter les développements réalisés tout au long des dernières décennies dans ce domaine. Il présente brièvement la progression du système de reconnaissance de la parole. La reconnaissance de la parole est un côté actuel du domaine de l'informatique. En 1950 apparut le premier système de reconnaissance de chiffres qui est un appareil parfaitement câblé (Davis et al. 1952). De plus, il existe d'autres éléments électroniques de la reconnaissance de chiffres ou bien de voyelles qui ont été mis au point en France (Dreyfus-Graf J., 1950), aux Etats-Unis (Olson et al. 1956), en Grande-Bretagne (Dénes J. 1959), au Japon (Sakai et al. 1962), et également en Italie (Meo et al., 1965). En 1960, l'utilisation des ordinateurs et l'introduction des méthodes numériques guident à un changement au niveau de la dimension des recherches (Forgie et al. 1959).

A partir de la fin des années 1960, Martin et ses collègues ont lancé le premier projet de laboratoire RCA dans le but de trouver des vraies solutions pour les problèmes dépendant de la non-uniformité des échelles temporelles dans les discours. Dans ce contexte même, Martin a pu développer des méthodes de normalisation élémentaire de temps liées à la possibilité de

détecter de manière plus efficace les débuts et les fins d'une telle parole (Martin et al. 1964). Ensuite, Martin a amélioré cette méthode et créer une société de reconnaissance vocale nommée Threshold Technology, qui fonctionnait de façon à commercialiser et vendre des outils de reconnaissance vocale. Cependant, Vintsyuk a proposé l'emploi des méthodes de programmation dynamiques liée de sa part à l'alignement de temps d'une déformation temporelle dynamique de la parole (Dynamic Time Warping (DTW) (Vintsyuk, T., 1968), contenant de plus, des algorithmes pour la reconnaissance de mots connectés. Malgré les exploits de ce dernier, mettant en main l'essence des concepts de «warp» de temps dynamique, elle était en une importante partie inconnue en occident et n'a pas été montrée avant l'année 1980. D'autre part, Sakoe et Chiba ont déclaré le début de l'utilisation d'une technique de l'utilisation dynamique afin de trouver des solutions aux problèmes de non-uniformité (Sakoe et al. 1978).

Environ 1970, elle est apparue la nécessité d'utiliser des contraintes linguistiques dans le cadre du décodage automatique de phrases, pourtant la reconnaissance de la parole avait été prise comme étant une tâche d'ingénierie. Vers 1971, l'agence des Projets de Défense Avancée des États-Unis (DARPA) a entamé un projet de cinq ans sur la création de trois nouveaux systèmes (Skinner et al. 1976) afin de tester l'efficacité de la compréhension automatique d'une parole continue. En 1975, les modèles de Markov cachés ont été employés dans la reconnaissance automatique de la parole par les chercheurs de C.M.U et I.B.M. Ainsi, on a mis fin à la première génération des systèmes commercialisés de reconnaissance de mots isolés, et donner naissance aux systèmes de reconnaissance de phrases.

À partir des années 1980 les chercheurs travaillent sur la reconnaissance des mots connectés (Furui S., 2005). Le changement majeur de cette période est présenté par le passage des systèmes fondés sur les techniques liées aux systèmes des modèles statistiques (Arora et al. 2012). Enfin, des années quatre-vingts, les projets de recherche profondes de la défense (DARPA), qui inculquait un large programme de recherche dans le but d'avoir une précision fiable de mots dans le cas de mille mots, offrant une impulsion importante aux systèmes de reconnaissance vocale de la parole continue. Parmi les travaux de recherche qui sont mis au sein de la CMU, le système SPHINX a associé la méthode statistique. Pour cette raison, il a pu former et associer des modèles de phonème qui dépend du contexte dans un réseau de décodage lexical. Cependant, le programme DARPA s'est poursuivi dans les années 90. Au cours de cette période, l'accent a été mis sur le système de reconnaissance du langage naturel

et sur la tâche de conserver les informations sur les voyages aériens. De plus, la technologie de RAP est intégrée dans les systèmes téléphoniques.

Dans les années 2000, DARPA a lancé un programme focalisé sur la détection des frontières d'une phrase, d'une disfluence, d'un bruit, d'une obtention de résumés ou bien de traductions dans un cadre de parole spontanée et de différentes langues (Furui S., 2005). Ainsi, on a étudié des méthodes pour tester la confiance des hypothèses de reconnaissance (Nendaz et al. 2005).

Après avoir inventé les réseaux de neurones dans les années 1950, leur pratique n'a devenu forte que plus tard. Ainsi, on les utilisés dans la modélisation lexicale (Strik et al. 1999 ; Wester M., 2003; Bouallegue M., 2013). Les réseaux de neurones ont connu une amélioration considérable à partir des années 2010 lorsqu'ils ont commencé à utiliser de nombreuses couches cachées basant sur un algorithme de pré-entraînement non supervisé (Hinton et al. 2006.

Actuellement, les chercheurs au travers le monde s'intéressent bien fortement à rendre les systèmes prêts à résoudre divers types de problèmes, tel que l'apprentissage des différentes langues, l'assistant personnel, la traduction automatique, les pathologies et les personnes âgés, etc. On peut citer quelques recherches de l'actualité, qui consiste dans la détection de frontières de phrases (Dutrey et al. 2014), l'évaluation de la parole chez les patients dysphoniques pour la classification du type et de la gravité des pathologies vocales (Muhammad et al. 2011) et le test des applications informatiques en milieu bruité (Janicki A., 2013). Pourtant, une étude précédente a exploré dans quelle mesure la reconnaissance de la parole émotionnelle des émotions «de base» (colère, dégoût, peur, bonheur, agréable surprise, tristesse) diffère entre les différents groupes de sexe (masculin / féminin) et d'âge (jeune / d'âge moyen) dans une expérience comportementale (Paulmann et al. 2008).

3. Application de la Reconnaissance de la Parole

Dans la RAP, la transmission des informations est supérieure par rapport à l'usage du clavier. Les gens ne trouvent pas de difficulté en parlant, alors qu'à l'écriture le souci des fautes d'orthographe s'imposent énormément. L'oral est caractérisé par la fluidité, l'abondance et par son caractère éphémère. Dans ce sens, la reconnaissance de la parole à divers avantages. Elle s'attache à l'oral et nous épargne toutes les difficultés et les complications liées à l'écrit. En fait, elle libère l'utilisateur du clavier, économise ses gestes, ses mouvements tactiles. Seule la voix importe au détriment des autres sens en l'occurrence la

vue, l'odorat… Ces avantages ont abouti en quelque sorte à une variété d'applications telles que :

- ✓ L'aide des personnes handicapées.
- ✓ La composition de numéros de téléphone par la voix.
- ✓ L'avionique.
- ✓ La commande de robots.
- ✓ La dictée vocale.
- ✓ La saisie des données et le contrôle de qualité.
- ✓ L'accès à distance : Internet et téléphone.
- ✓ La validation des numéros de cartes ou de comptes bancaires.

Toutes les applications précédentes bénéficient de la progression technologique qui consiste dans l'invention de composants intégrés spécialisés du développement des techniques, ainsi que les méthodes algorithmiques fiables et performantes.

4. Caractéristique des systèmes de RAP

L'utilisation d'un système de reconnaissance vocale dans son domaine d'utilisation est attachée à son contexte d'application et aux conditions de déploiement qui rendent les problèmes aussi grands. Un système de reconnaissance automatique de la parole peut être présenté via cinq paramètres pour faire face à cette difficulté:

4.1. Mode de fonctionnement

Le système de reconnaissance automatique de la parole fonctionne d'une façon à indiquer la dépendance d'un tel système par rapport au locuteur. On en trouve deux types: Le premier système dépend du locuteur appelé système mono-locuteur caractérisé par sa capacité de savoir la voix d'un même locuteur. Le second, nommé système multi-locuteur qui est indépendant du locuteur, reconnaît la voix quel que soit le locuteur. Pourtant, le taux de reconnaissance des systèmes mono-locuteurs est plus fort que celui des systèmes multi-locuteurs, car la variabilité dans le signal vocal est plus compliquée au niveau de la gestion dans le deuxième cas. Néanmoins, les systèmes mono-locuteurs sont obligés à regrouper les principales caractéristiques transmises par les signaux de parole d'un nombre important de locuteurs pour atteindre les divers types de la voix, par suite, un mono-locuteur est entraîné à

la voix d'un locuteur particulier.

4.2. L'environnement

L'amélioration des performances d'un système consiste dans le taux de précision réel de la reconnaissance au niveau de l'enregistrement et dans le milieu d'utilisation. De nombreux aspects peuvent affecter les performances de la reconnaissance automatique de parole comme entre des salles calmes, chambres sourdes, les lieux bruyants, les transducteurs.

4.3. Mode d'élocution

Le mode d'élocution indique la manière avec laquelle le locuteur communique avec le système. Il s'est appelé mots isolés qui nécessite généralement que chaque énoncé soit calme des deux côtés des fenêtres d'échantillonnage et n'acceptant qu'un seul mot à la fois. D'autre part, les mots connectés ressemblent presque à des mots isolés, mais demande une pause minimale entre les énoncés pour former un flux clair de la parole. Par contre, le mode appelé mots continus apparait lorsque la communication est réalisée sans avoir besoin de faire des pauses au moment de l'enregistrement de la parole. D'ailleurs, la reconnaissance de la parole est plutôt facile avec les mots isolés mais difficile à atteindre quand on a des mots continus, car cela nécessite des méthodes spéciales et un son unique pour déterminer les limites des énoncés. Enfin, le discours spontané est un discours qui sonne naturellement et non répété. Un système de reconnaissance vocale qui permet de reconnaître la parole spontanée devrait être capable d'exploiter une diversité de caractéristiques de la parole naturelle telle que les mots exécutés ensemble.

4.4. Taille du vocabulaire

La forme du vocabulaire d'un système de RAP affecte la complexité, les besoins de traitement et la performance du système. Certains systèmes ne nécessitent que quelques mots (comme des nombres), tandis que d'autres nécessitent des dictionnaires très volumineux (tels que des machines à dicter). Dans les systèmes de la reconnaissance de la parole, les types de vocabulaire peuvent être classés comme suit :

- Petit vocabulaire : contient des dizaines de mots.
- Vocabulaire moyen : inclut des centaines de mots.
- Grand vocabulaire : comporte des milliers de mots.

- Vocabulaire volumineux : inclut des dizaines de milliers de mots.

4.5. Unités phonétiques

Les systèmes de la reconnaissance automatique de parole ne se basent pas sur le groupe des mots du vocabulaire lié à la quantité de mots qui se trouvent dans un ensemble de paroles, en plus de la nécessité d'avoir un nombre suffisant d'exemples prononcés pour chacun des mots à reconnaître. D'autre part, un système de la reconnaissance automatique de parole est limité à l'application de petites unités phonétiques qui sont formées par des spécialistes du domaine linguistique. On nomme ce corpus d'unités de sous-mot, Phonèmes. D'ailleurs, le phonème est défini comme étant la plus petite unité de l'oral contenant une valeur distinctive dans la langue.

Chaque système dépend d'un compromis entre ces différents axes. Il est choisi en fonction de l'objectif qu'on vise à atteindre. Les systèmes conçus de la reconnaissance automatique de la parole sont faits pour des applications précisées. Cela guide à réduire l'univers de la communication entre l'homme et la machine. De plus, la conception de systèmes qui comprennent la langue orale est jusqu'à présent très compliquée.

5. Difficultés de la Reconnaissance de la Parole

Le défi sous-jacent de la technique de reconnaissance vocale est la grande complexité existante dans le signal de la parole. Ce signal présente des caractéristiques qui rendent son interprétation difficile à faire. Le principal obstacle pour développer les performances d'un système de la reconnaissance automatique de parole vient de la complexité majeure du signal de la parole lié à la concaténation de plusieurs facteurs, en particulier la redondance, la continuité, les effets de coarticulation et l'ample variabilité intra et interlocuteurs, ainsi que les conditions d'enregistrement. Enfin, le système de la reconnaissance automatique de parole doit prendre en considération les données précédentes dans sa réalisation.

5.1. La Redondance

Le signal vocal présente un caractère redondant car il transporte une quantité importante d'informations (des informations liées au locuteur à savoir son état émotionnel, son timbre, etc. . .). Or, la présence de toutes ces informations n'est pas nécessaire pour réaliser la reconnaissance automatique de la parole. En effet, il est indispensable d'utiliser seulement les caractéristiques dépendant du message linguistique. Pour cela, on a pensé à faire la paramétrisation afin d'extraire les paramètres pertinents liés à la reconnaissance automatique

de la parole et également minimiser tant que possible la redondance du signal.

5.2. Variabilité

Le phonème de deux prononciations ayant une même information est varié même s'il a été produit par le même locuteur (variabilité intra-locuteur) ou pour des locuteurs différents (variabilité interlocuteur). La différence au niveau du signal vocal entre deux prononciations d'un même énoncé à contenu phonétique égal peut être causée par plusieurs facteurs, dont:

- L'état physique, comme le rhume ou la fatigue.
- L'état psychologique, par exemple la dépression et les addictions.
- Les émotions de la personne.
- La façon dont nous exprimons l'amplificateur et l'amplitude (normal, faible volume, volume élevé ...)

Pourtant, la variabilité la plus importante est celle d'interlocuteur. Elle s'explique par les points suivants:

- Les différences physiologiques entre les différents locuteurs.
- Le facteur social et géographique lié par exemple aux accents régionaux. Cette variabilité guide à compliquer la reconnaissance automatique de la parole.

5.3. Continuité et Coarticulation

Lorsque l'on entend parler une langue connue, on perçoit une continuité de mots, qui peuvent à leur tour être décrits comme une suite de sons élémentaires appelés phonèmes. Chaque phonème peut être représenté par un ensemble de caractéristiques articulatoires, comme l'harmonisation, le degré d'ouverture, le lieu d'articulation, etc. L'inertie du système de production de la parole humaine suggère que ces caractéristiques changent en douceur dans le temps, ce qui peut induire des variantes de prononciation. Ainsi que les règles de coarticulation intègrent ces contraintes articulatoires et prédisent les prononciations alternatives possibles de mots / phrases

5.4. Conditions d'enregistrement

Extraire les informations essentielles liées à la reconnaissance de mots prononcés est difficile lors de l'enregistrement dans de mauvaises conditions. En effet, les perturbations apportées par le type de microphone (à main, casque, microphone d'ordinateur et encore

d'autres) et l'environnement (bruit, réverbération) complique grandement le problème de la reconnaissance vocale. Pour clarifier toutes ces difficultés, le système RAP doit en conséquent être en mesure de décider "que [a] prononcé par un homme adulte est plus proche de [a] d'un enfant, dans un mot, environnement et un microphone différents, que le mot [i] prononcé par même adulte dans une même phrase" (Hamani et al. 2015).

6. Reconnaissance de la Parole

6.1. Principe de Reconnaissance de la Parole

Le principe de base de la reconnaissance de la parole est d'extraire des caractéristiques du signal d'entrée et de classer par des phonèmes utilisant des modèles statistiques. Une caractéristique spécifique de la reconnaissance vocale est qu'elle nécessite de fonctionner sur des caractéristiques séquentielles avec plusieurs niveaux de contraintes. En particulier, la distribution préalable du résultat de la reconnaissance, définie par la langue cible, est essentielle lors de la reconnaissance de la parole naturelle avec un vocabulaire. Un système de RAP contient deux modules primordiaux. Il s'agit d'un module d'analyse acoustique dont le but est de produire une représentation plus compacte et plus significative du signal vocal, et d'un autre module de décision qui sert à décoder les informations issues de l'analyse acoustique. La figure 2.1 illustre une architecture de base pour un système de RAP.

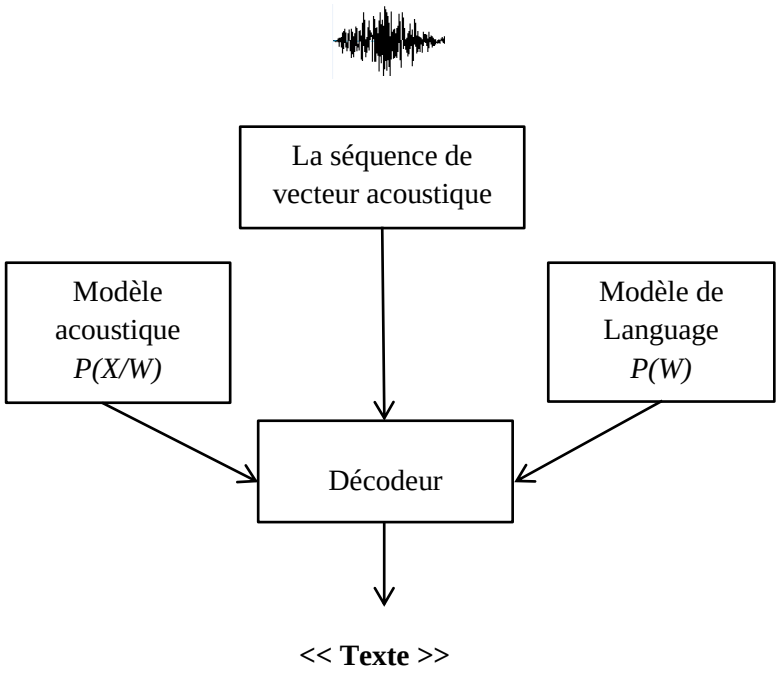


Figure.2. 1 : L'architecture générale de la reconnaissance de la parole.

La séquence de mots prononcés est convertie en un signal audio par un dispositif sonore humain. De plus, le signal audio est converti en un ensemble de vecteurs acoustiques ou bien

d'observations (chaque vecteur est un ensemble de paramètres acoustiques $X=X_1X_2X_3...X_n$). Enfin, le décodeur consiste à engendrer une série de mots reconnus à la séquence d'observations $\widehat{W} = w_1w_2w_3...w_k$. Le système vocal copie la séquence d'observation en une séquence de mots basée sur l'unité d'analyse vocale et l'unité de décodeur. Cette séquence doit maximiser l'équation suivante :

$$\widehat{W} = \underset{W}{argmax} \; P(W/X) \qquad \text{(eq. 2.1)}$$

Après, l'application de la formule de Bayes, on obtient :

$$\widehat{W} = \underset{W}{argmax} \; \frac{P(X/W).P(W)}{P(X)} \qquad \text{(eq. 2.2)}$$

La séquence d'observations *X* a été attaché et *P(X)* peut être utilisé comme une valeur fixe dont ne dépend pas l'équation 2.2 On obtient alors :

$$\widehat{W} = \underset{W}{argmax} \; P(X/W).P(W) \qquad \text{(eq. 2.3)}$$

La technique de RAP permet de maximiser le produit des probabilités :

- *P(X/W)*, probabilité d'une séquence d'observations vocales *X* sachant une séquence de mots *W*
- *P(W)*, probabilité a priori d'une série de mots.

Il est, en général, impossible de calculer directement la probabilité de la séquence de mots la plus probable. On peut cependant l'estimer à partir de corpus ou données d'apprentissage. Deux modèles probabilistes sont exploités pour cette phase : Un modèle acoustique qui nous permet d'obtenir la valeur de *P(X/W)*, et un modèle de langage qui permet d'avoir la valeur de *P(W)*. Pour produire un système vocal puissant, il est essentiel de générer les modèles les plus pertinents possibles pour l'estimation de *P(W)* et *P(X/W)*.

6.2. Les modules de RAP

6.2.1. Modèle acoustique

Le module acoustique est responsable du premier traitement du signal entre dans le système de reconnaissance. Son rôle principal est de produire les hypothèses liées à la probabilité de chaque unité du segment de parole, à partir des paramètres acoustiques. Aussi connu sous le nom de décodeur Acoustic-phonétique. Pour connaître les unités obtenues à l'étape de segmentation de la parole (souvent des phonèmes), l'unité doit avoir

appris à quoi ressemblent les réalisations phonétiques de ces unités en termes de vecteurs acoustiques. A partir de ces vecteurs, un formulaire statistique est créé pour chaque unité considérée en fonction de sa distribution. L'ensemble de modèles statistiques pour chaque unité vocale entrante est une forme de parole audio à stocker dans le système RAP. Le décodeur Acoustic-phonétique se compose généralement de deux sous-unités. La première consiste à extraire les paramètres distincts qui ont été choisis pour représenter le signal et la seconde consiste à connaître les modèles d'unités sonores à partir des groupes de paramètres (Satori, et al. 2014).

Le modèle acoustique permet de calculer le *P(X/W)*, c'est-à-dire la probabilité de générer une forme d'onde de parole pour le mode. Un modèle acoustique, en tant qu'élément important du système RAP, représente une grande partie de la surcharge de calcul et détermine également les performances du système. Les modèles acoustiques basés sur GMM-HMM sont largement utilisés dans les systèmes de reconnaissance vocale traditionnels. Dans ce modèle, GMM est utilisé pour modéliser la distribution des caractéristiques acoustiques de la parole et HMM est utilisé pour modéliser la séquence temporelle des signaux de parole.

6.2.2. Extraction des paramètres

Dans les systèmes de RAP, la première phase du traitement du signal vocale consiste à extraire les paramètres caractéristiques d'audio. Afin d'estimer un ensemble de vecteurs de caractéristiques qui donnent une représentation du signal de parole compressé. Il est généralement mis en œuvre en trois phases. La première étape, appelée analyse de la parole. Elle effectue l'analyse temporelle du spectre du signal de parole et adopte les premières caractéristiques qui donnent une description d'enveloppe du spectre de puissance du séparateur de parole court. La deuxième étape rassemble des vecteurs de caractéristiques étendus qui ont des caractéristiques dynamiques et statiques. Enfin, la dernière étape convertit ces vecteurs étendus en vecteurs plus puissants et plus puissants qui constituent des paramètres représentatifs du signal de parole. Dans la littérature, les principaux paramètres les plus utilisés sont :

6.2.2.1. Coefficients MFCC

Coefficients cepstraux Mel-Frequency (en anglais Mel-Frequency Cepstral Coefficients ou MFCCs) ce sont des coefficients cepstraux basés sur les variations connues des largeurs de bande critiques de l'oreille humaine avec des fréquences inférieures à 1000 Hz. C'est une des

méthodes qui permet d'extraire les caractéristiques. La technique la plus populairement utilisée d'extractions dans la reconnaissance de la parole est l'échelle de l'oreille humaine (échelle de Mel). En effet, le but principal du MFCC est de copier le comportement des oreilles humaines [(Satori H., 2009) (Shah et al. 2015)].

MFCC est une représentation du cepstre d'un signal avec une fenêtre de courte durée dérivée de la transformation rapide de Fourier (FFT) du signal. Ils sont robustes contre la variation de locuteurs et les différentes conditions liées à l'enregistrement comme déclaré précédemment. Le signal vocal est en effet segmenté en trames temporelles formées d'un nombre arbitraire de représentants. Chacune des trames est par la suite, fenêtrée avec la fenêtre de Hamming afin de se dégager des discontinuités au niveau des bords (Lahouti et al. 2006). L'usage du filtre de préaccentuation a pour but de relever les régions hautes fréquences qui sont moins énergétiques que celles de basses fréquences (voir figure 2.2). La préaccentuation du signal échantillonné à l'instant est calculée pour une valeur qui varie entre 0,9 et 1, par l'équation suivante:

$$P(n) = S(n) - a \times S(n-1) \qquad \text{(eq. 2.4)}$$

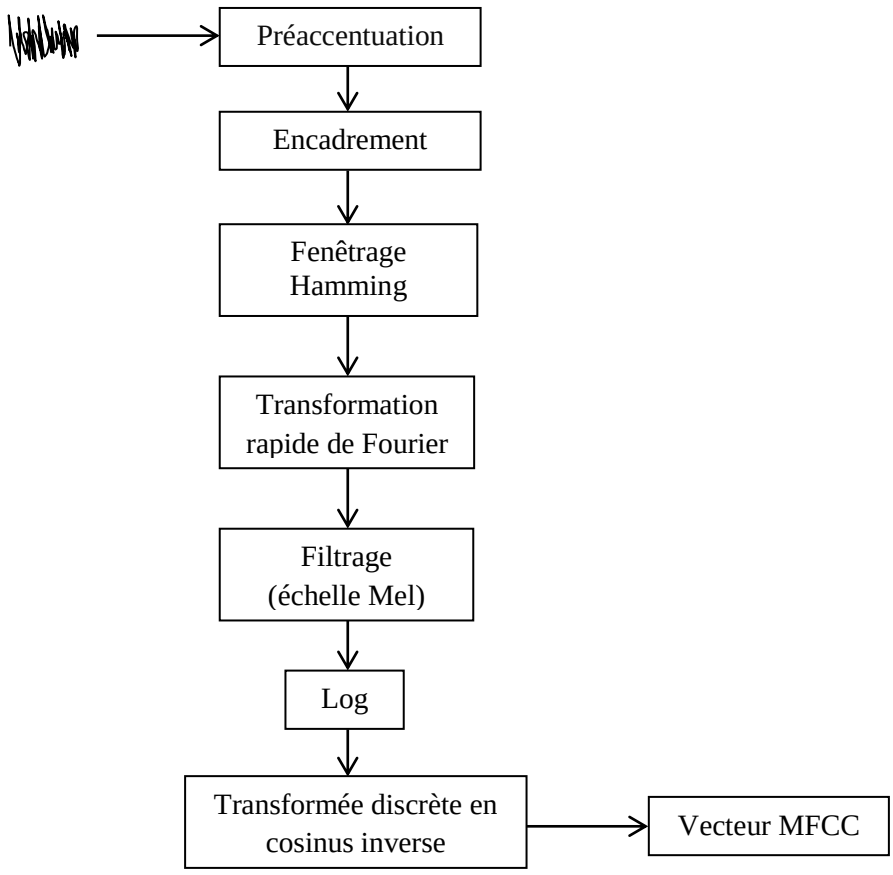


Figure.2. 2 : Schéma fonctionnel des techniques d'extraction MFCC.

Le calcul des coefficients d'une fenêtre de Hamming *W(n)* de longueur *n* se réalise selon la formule :

$$W(n) = 0.54 - 0.46 * cos\left(\frac{2\pi n}{N-1}\right) \qquad 0 \leq n \leq N-1 \qquad \text{(eq. 2.5)}$$

Où *n* est l'échantillon courant et *N* le nombre total d'échantillons. Après le fenêtrage, la transformation rapide de Fourier (FFT) est calculée sur chacune des trames afin d'extraire des éléments fréquentiels du signal vocal dans le domaine temporel. Cette transformée est appliquée pour avancer les traitements et acquérir du spectre. Un banc de filtres logarithmiques est adapté aux fenêtres transformées pour abreuver chacun une fréquence. Par l'usage des filtres, chacune des fenêtres générées est convertie à l'échelle de Mel. Cette échelle est logarithmique à de plus grandes fréquences. La relation entre la fréquence de la parole et l'échelle de Mel s'établit comme suit :

$$Mel(f) = x \times log\left(1 + \frac{f}{y}\right) \qquad \text{(eq. 2.6)}$$

Il existe tant des valeurs données aux variables *x* et *y*. Les plus couramment utilisées sont *x* = 2595 et *y* = 700. La dernière phase de calcul des coefficients MFCCs est le calcul de la transformation discrète de cosinus (TDC) sur les sorties du banc de filtre. Ce qui conduit à obtenir des coefficients MFCCs. Ainsi, pour chacune des trames de parole, un ensemble de coefficients MFCCs est calculé. Cet ensemble est appelé vecteur acoustique et représente les caractéristiques phonétiquement importantes de la parole. Il est très bénéfique pour une analyse plus approfondie et le traitement dans la reconnaissance de la parole.

6.2.2.2. Codage prédictif linéaire (LPC)

Le codage prédictif linéaire (Linear Predictive Coding -LPC) est une technique largement utilisée dans le traitement du signal audio, en particulier dans le traitement du signal vocal. Il a trouvé une utilisation particulière dans la compression du signal vocal, permettant des taux de compression très élevés.

L'objectif initial de LPC était de modéliser la production de voix humaines. LPC est un modèle de filtre de source en ce qu'il existe une source sonore qui passe à travers un filtre (voir figure 2.3). La source *e(n)* modélise les cordes vocales, tandis que le filtre résonnant *h(n)*, modélise le conduit vocal. Le signal résultant est :

$$x(n) = h(n) \times e(n) \qquad \text{(eq. 2.7)}$$

Il existe deux signaux possibles pour la source : un train d'impulsions ou un bruit blanc

aléatoire. Ces signaux modélisent respectivement les sons toniques et plosifs / fricatifs. La caractéristique commune du train d'impulsions et du bruit blanc est qu'ils sont spectralement plats. Toutes les informations spectrales sont modélisées dans le filtre. Le lecteur attentif remarquera que le signal source est étiqueté *e(n)*. Cela a été choisi pour des raisons qui seront révélées dans les sections suivantes.

LPC suppose que le filtre soit un filtre omnipolaire d'ordre *p*. Bien qu'il ne soit pas physiologiquement exact, il fournit une méthode extensible pour modéliser les résonances. Cela permet également une solution traitable lors de l'estimation de *h(n)* à partir de *x(n)*.

Bien qu'initialement développé pour les signaux vocaux, l'hypothèse d'un signal source spectralement plat et d'un filtre résonnant s'applique bien à la modélisation des signaux de la plupart des instruments tonaux ainsi que de nombreux sons naturels.

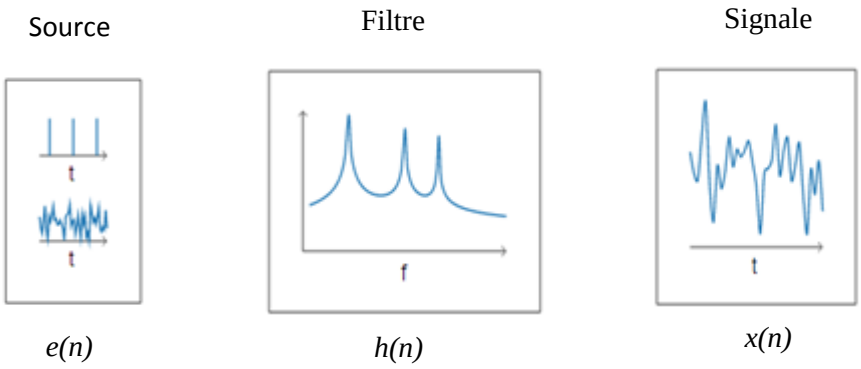


Figure.2. 3 : Le modèle LPC.

6.2.2.3. Prédiction linéaire perceptuelle (PLP)

Le modèle de Prédiction Linéaire Perceptuelle (Perceptual Linear Prediction-PLP) est développé par (Hermansky H., 1990). Le but de ce modèle est de décrire plus précisément la psychophysique de l'audition humaine dans le processus d'extraction des caractéristiques. Le PLP, similaire à l'analyse LPC, est basé sur le spectre à court terme de la parole. Contrairement à l'analyse prédictive linéaire pure de la parole, la prédiction linéaire perceptuelle modifie le spectre à court terme de la parole par plusieurs transformations basées sur la psychophysique.

6.2.2.4. Coefficient Cepstral de Fréquence Linéaire (LFCC)

Il s'agit d'une variante des MFCCs. La différence vient de l'utilisation d'un banc de filtres linéaire, contrairement à l'échelle de Mel des MFCCs.

Autres paramètres

Au-delà des principales approches de paramétrage évoquées ci-dessus, d'autres peuvent être trouvées dans la littérature comme NPC (Neural Predictive Coding - extension non linéaire du codage LPC), LSF (Line Spectral Frequencies - les fréquences des raies spectrales, du LPC).

6.2.3. Modèles de Language

Un modèle de langage statistique est une distribution de probabilité sur des séquences de mots. Etant donné une telle séquence, disons de longueur *m*, il attribue une probabilité display style $P\ (w_1, \ldots, w_m)$ aux séquences entières.

Le modèle de langage fournit un contexte pour distinguer les mots et les phrases qui semblent similaires. Par exemple, en anglais américain, les expressions « reconnaissent la parole » et « épave une belle plage » semblent similaires, mais signifient des choses différentes.

La rareté des données est un problème majeur dans la construction de modèles de langage. La plupart des séquences de mots possibles ne sont pas observées lors de la formation. Une solution est de faire l'hypothèse que la probabilité d'un mot ne dépend que des *n* mots précédents. Ceci est connu sous le nom de modèle n-gramme ou modèle uni-gramme lorsque $n = 1$. Le modèle uni-gramme est également connu sous le nom de modèle du sac de mots.

L'estimation de la probabilité relative de différentes phrases est utile dans de nombreuses applications de traitement du langage naturel, en particulier celles qui génèrent du texte en tant que sortie. La modélisation du langage est utilisée dans la reconnaissance vocale, la traduction automatique, l'étiquetage d'une partie du discours, l'analyse, la reconnaissance optique de caractères, la reconnaissance de l'écriture manuscrite, la recherche d'informations et d'autres applications.

Dans la reconnaissance vocale, les sons sont associés à des séquences de mots. Les ambiguïtés sont plus faciles à résoudre lorsque les preuves du modèle de langage sont intégrées à un modèle de prononciation et un modèle acoustique.

Les modèles de langage sont utilisés dans la recherche d'informations dans le modèle de probabilité de requête. Là, un modèle de langage distinct est associé à chaque document d'une collection. Les documents sont classés en fonction de la probabilité de la requête *Q* dans le

modèle de langage du document W_d : $P\ (Q \mid W_d)$.

7. Les approches utilisées en RAP

Le RAP présente un problème à deux dimensions : L'approche globale et l'approche analytique. La première approche prend la forme globale du mot ou de la phrase identifiable dans la comparaison avec des références enregistrées. La seconde, employée dans la parole continue, a pour but d'analyser une phrase en la segmentant par unités et en faisant un décodage acoustico-phonétique exploité par des éléments de niveau linguistique.

Par ailleurs, la majorité des systèmes RAP utilisent des techniques de statistiques basées sur des modèles de Markov. Ces méthodes sont mixtes (globale et analytique).

7.1. Approche globale

L'énoncé entier est considéré par l'approche globale comme une seule unité indépendamment de la langue. Ainsi, elle consiste à faire une abstraction parfaite des phénomènes linguistiques et retenir seulement l'aspect acoustique de la parole. En général, cette approche vise à réaliser la reconnaissance des mots isolés et séparés par au moins 200 ms (voir figure 2.4) ou enchaînés, appartenant à des vocabulaires réduits.

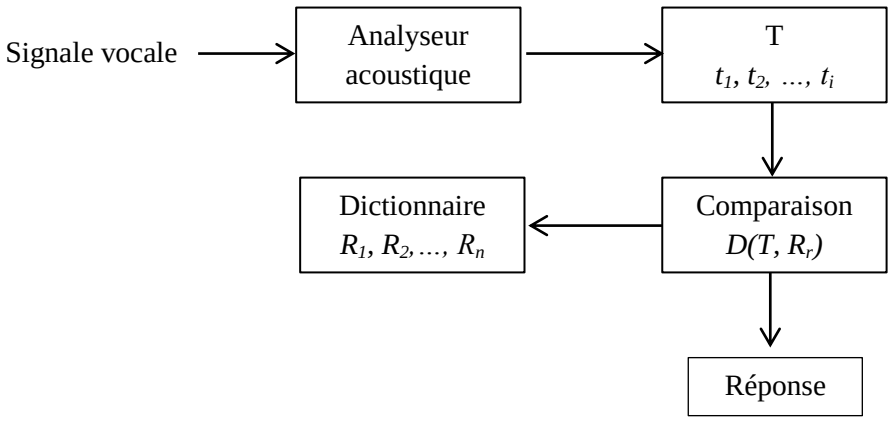


Figure.2. 4 : Reconnaissance de mots isolés.

Les systèmes de reconnaissance globale nécessitent une phase d'apprentissage. Lors de cette phase, l'utilisateur prononce les mots du vocabulaire de son application. À chacun des mots prononcés, on effectue une analyse acoustique qui permet d'extraire les données pertinentes sous la forme des vecteurs de paramètres acoustiques. Ensuite, le résultat se stocke en mémoire, ainsi les méthodes globales mettent en main une ou plusieurs images de

références acoustiques (R_1,.., R_n) a priori pour chaque mot.

Pendant la phase de reconnaissance, si l'utilisateur prononce un mot *T*, on constate qu'une même analyse s'effectue : L'image acoustique du mot concerné est alors comparée à toutes celles des mots de référence du vocabulaire via un indice de ressemblance D:

$$m = \underset{1 \leq r \leq N}{argmax} D[(T, R_r)] \qquad \text{(eq. 2.8)}$$

Le mot qui est plus ressemblant à celui prononcé est alors reconnu.

De manière générale, on peut trouver deux problèmes différents: Le premier est lié à la durée d'un mot qui varie d'une prononciation à l'autre, et le second dépend des déformations qui non linéaires en fonction du temps. Ces deux problèmes peuvent en telle ou telle sorte être résolus grâce à un algorithme classique de la programmation dynamique nommé alignement temporel dynamique (Dynamic Time Warping-DTW). Cet algorithme vise à obtenir une solution optimale à un problème de minimisation d'un critère d'erreur sans prendre en considération toutes les solutions possibles. Avec la reconnaissance automatique de la parole, cet algorithme consiste dans la recherche de l'alignement temporel le plus performant qui guide à minimiser la distance entre la représentation d'un mot de référence et celle d'un mot inconnu.

Cette approche devient insuffisante lorsque le vocabulaire est grand, ou bien de parole naturelle continue, ainsi il est indispensable de mettre en place une nouvelle approche (Hacine-Gharbi A., 2018).

7.2. Approche analytique

La résolution des problèmes de la reconnaissance de la parole continue et ceux du traitement de grands vocabulaires se base sur l'approche analytique. En effet, cette approche permet de segmenter le signal vocal en constituants élémentaires (phonème, mot, biphone, syllabe), puis de les identifier, et ensuite à reconstituer la phrase prononcée via des étapes successives en utilisant des modules d'ordre linguistique (niveaux lexical, sémantique ou syntaxique). Ces constituants élémentaires prennent des formes à savoir des phonèmes, des biphones, des syllabes ou bien des triphones. La reconnaissance de la parole s'effectue à travers un processus pouvant être décomposé en deux opérations: la segmentation qui consiste dans la représentation du message (signal vocal) en forme d'un ensemble de segments de parole, et l'identification qui consiste dans l'interprétation des segments trouvés en termes d'unités phonétiques.

7.3. Approche statistique

L'approche statistique se base sur une formalisation statistique simple extraite de la théorie de l'information qui permet la décomposition du problème de la reconnaissance de la parole continue.

Cette approche est fondée sur le principe de fonctionnement des méthodes globales (avec phase d'apprentissage et de reconnaissance) avec l'exploitation des niveaux linguistiques. Ainsi, une analyse acoustique est indispensable afin de convertir le signal vocal entièrement en une suite de vecteurs acoustiques. Dans la phase de reconnaissance et celle d'apprentissage des modèles statistiques ces vecteurs sont considérés comme des observations.

Soit X un ensemble d'observations acoustiques venant d'une analyse acoustique d'un signal de parole qui représente une suite de mots prononcés W. L'approche statistique vise en effet à chercher les mots $\widehat{W}$ ayant une probabilité plus puissante parmi toutes les séquences de mots EM sachant les observations X. Ainsi, la maximisation de la probabilité a posteriori $P(W/X)$ se réalise à travers la séquence de mots optimale.

8. Les systèmes de RAP

Au fil du temps, diverses boîtes à outils ont été construites pour contribuer à la recherche sur la technologie RAP. De nombreuses boîtes à outils en libre accès et commerciales sont disponibles pour créer un moteur RAP. Certains d'entre eux sont:

8.1. CMU Sphinx

Sphinx est une célèbre boîte à outils spécialement développée par Université de Carnegie Mellon (CMU) pour le moteur de reconnaissance vocale basé sur les modèles Markov cache à large vocabulaire, continu et indépendant du locuteur. Dans le cadre de ce travail, nous sommes intéressés au système de RAP développé au sein du CMU à Pittsburgh USA. Ce système (Lee K., 1988), diffusé sous licence libre depuis 2001. Il est actuellement l'un des systèmes de reconnaissance de parole les plus puissants. Le CMU Sphinx permet à des groupes de recherche avec des budgets modestes de développer et de conduire des applications de recherches dans la reconnaissance de la parole. Pour ces raisons, nous avons choisi ce système pour développer notre application pour la détection des fumeurs et pathologies dans des environnements différents. Le système CMU Sphinx est disponible sous deux versions :

– Sphinx 2 a été lancé et a étendu ses fonctionnalités pour inclure le traitement de la parole semi-continu basé sur HMM.

– Sphinx 3 : Il s'agit d'un décodeur qui a longtemps été le décodeur phare CMU
Sphinx. Cette une version qui a été développée en langage C. Il utilise des modèles de Markov continus.

– Sphinx 4 : la version 4 de Sphinx est une réécriture complète en langage de programmation Java. Une description détaillée se trouve dans (Walker et al. 2004).

– Sphinx 3 et Sphinx 4 utilisent les mêmes modèles acoustiques et modèles de langage.

8.2. HTK

La boîte à outils HTK est conçue par le département d'ingénierie de l'Université de Cambridge (CUED) qui permet de construire et de manipuler les MMCs. HTK est un ensemble de modules de bibliothèque et d'outils disponibles en langage C. Il est utilisé dans les études de reconnaissance vocale et de nombreuses autres applications comme synthèse vocale (Woodland et al. 1995). Ce cadre fournit des outils système pour la phase de formation et de test. Au départ, les outils d'apprentissage HTK sont utilisés pour traîner les MMCs à l'aide de la parole de formation à partir d'une base de données stockée. Les caractéristiques sont extraites de ces sons d'entraînement, puis ces caractéristiques sont utilisées pour modéliser le système. Enfin, des outils de reconnaissance HTK sont appliqués pour transcrire le son inconnu. Ils utilisent le modèle de système généré pendant la phase d'apprentissage pour la reconnaissance.

8.3. Kaldi

Kaldi est une boîte à outils open source conçue pour traiter les données vocales. il est utilisé dans des applications liées à la voix, principalement pour la reconnaissance vocale, mais aussi pour d'autres tâches, comme la reconnaissance du locuteur et la numérisation du locuteur (Povey et al. 2011). La boîte à outils est déjà assez ancienne (environ 9 ans) mais est toujours constamment mise à jour et développée par une communauté assez importante. Kaldi est largement adopté à la fois dans le milieu universitaire (plus de 400 citations en 2015) et dans l'industrie. Kaldi est écrit principalement en C / C ++, mais la boîte à outils est encapsulée avec des scripts Bash et Python. Pour une utilisation de base, cet emballage évite d'avoir à entrer trop profondément dans le code source. Au cours des 5 derniers mois, j'ai appris l'existence de la boîte à outils et son utilisation. Le but de cet article est de vous guider

tout au long de ce processus et de vous fournir les matériaux qui m'ont le plus aidé. Voyez-le comme un raccourci.

8.4. Dragon NaturallySpeaking

Dragon NaturallySpeaking (NatSpeak) est un système de reconnaissance vocale continue à grand vocabulaire à usage général. NatSpeak permet aux utilisateurs de dicter du texte en parlant dans le microphone d'un casque au lieu d'utiliser le clavier. NatSpeak a également la capacité de reconnaître les commandes vocales et de traduire ces commandes en séquences de touches ou en d'autres actions qui manipulent l'application actuelle ou Windows lui-même. Ce document a été (principalement) dicté par la voix en utilisant Dragon NaturallySpeaking. Ce dernier est vendu dans une variété d'éditions différentes, différenciées par l'ensemble de fonctionnalités et le prix.

9. Evaluation d'un système de reconnaissance automatique de la parole

Les performances du système de reconnaissance de la parole sont principalement déterminées en termes de précision et de vitesse. L'évaluation peut être calculée en termes de précision de performance qui est souvent évaluée avec le taux d'erreur de mot (*WER*) (Kalamani et al. 2014).

Le taux d'erreur de mots est une mesure dominante des performances de reconnaissance vocale ou d'un dispositif de traduction. La difficulté de performance déterminante réside dans le fait que la séquence de mots reconnue peut avoir plusieurs longueurs à partir de la séquence de mots de référence. Le *WER* est dérivé de la distance de Levenshtein, ce dernier est le nombre de suppressions, d'insertions ou de substitutions nécessaires pour transformer la chaîne source en chaîne cible. Le taux d'erreur de mot peut alors être calculé comme suit:

- de substitution (*S*) : mot reconnu à la place d'un mot de la transcription manuelle.
- d'insertion (*I*) : mot reconnu insérer par rapport à la transcription de référence.
- de suppression (*D*) : mot de la référence oublié dans l'hypothèse fournie par le système de RAP.
- *N* : est le nombre de mots dans la référence.

$$WER = \frac{S+I+D}{N} \times 100 \qquad \text{(eq. 2.9)}$$

Lorsque nous signons les performances d'un système de reconnaissance vocale, parfois on peut utiliser le taux de la reconnaissance de mots (*WRR*) comme suit:

$$WRR = 1\text{-}WER = \frac{N-S-D-I}{N} = \frac{H-I}{N} \qquad \text{(eq. 2.10)}$$

Où

H est *N- (S + D)*, le nombre de mots correctement reconnus.

10. Conclusion

Dans ce chapitre, on a introduit le cadre d'étude des travaux accumulés au cours de cette thèse et l'état de l'art sur la segmentation de la reconnaissance automatique de la parole. Nous avons commencé par un bref historique sur ce système. Afin de comprendre le fonctionnement de RAP, on a présenté les difficultés rencontrées lors de la réalisation d'un système de la parole, les modules utilisés et les approches de la reconnaissance, ainsi que les outils exploités. Le chapitre suivant décrit les méthodes permettant l'implémentation des modèles de Markov cachés qui sont utilisés comme supports de modélisations dans le cadre de cette thèse.

Chapitre 3: Les Modèles de Markov Cachées

1. Introduction

Les modèles de Markov cachés considérés comme des techniques statistiques très efficaces qui permettent de modéliser des phénomènes stochastiques. Ces modèles sont très riches en structure mathématique et peuvent donc constituer la base théorique pour une utilisation dans diverses applications (Huang et al. 2001) dont la reconnaissance de la parole, suivi des formants et de la hauteur de ton, amélioration de la parole, compréhension du langage parlé et traduction automatique, l'indexation de documents et le traitement d'images et prédiction de séries temporelles. Pourtant, il est nécessaire de connaître les principes pour pouvoir utiliser ces modèles d'une manière délicate.

Le but de ce chapitre est l'établissement des bases et les algorithmes constituant la théorie des modèles de Markov cachés (MMC), ainsi que les techniques utilisées dans la mise en œuvre des MMCs dans la RAP en se basant sur leurs algorithmes d'apprentissage et de décodage. Nous commençons en tout premier lieu par la présentation de l'histoire réelle de la théorie précédente connue par différentes phases. Après avoir fait une préface autour des chaines de Markov, afin de mieux traiter les phénomènes étudiés, il est primordial de prendre en compte un modèle puissant au niveau d'expression. Pour ce motif, nous avons choisi les modèles de Markov cachés (MMC) qui ont cette caractéristique. Par la suite, nous allons décrire les algorithmes classiques des MMC utilisés dans le décodage et la segmentation où la reconnaissance: Forward, Backward et de Viterbi. Dans la dernière section de ce chapitre nous allons parler des critères exploités dans l'apprentissage de MMC.

2. Historique du Modèle de Markov

Les modèles de Markov cachés ont une histoire ancienne qui a commencé au début des 19 siècles. En 1913 les premières études sur les chaines de Markov pour examiner le langage permettent à A.A. Markov d'établir la théorie des chaines de Markov (Markov A., 1913). De 1948 à 1951, on pratique les chaines de Markov (Shannon C., 1948), Shannon conçoit l'hypothèse de l'information.

Dès 1958, les modèles probabilistes d'urnes (Feller W., 1958), l'analyse de la continuité d'états dans une chaine de Markov la simplification et unification du maximum de vraisemblance, sont effectués (Hartley H., 1958). La mise en point des algorithmes fondamentaux pour l'estimation des états et des indices des modèles de Markov cachés il n'a pas eu lieu jusqu'en 1966 par L. E. Baum (Baum L., 1972). Depuis 1980, ces modèles sont

développés afin d'intégrer le concept de densité de probabilité continue multi-variable et de durée instable (Furguson J., 1980). Les recherches de Viterbi (Viterbi A., 1967) et G. D. Forney (Forney G., 1973) ont permis d'élaborer une technique performante et dont la complexité est linéaire, par subsistance à la longueur de la suite d'observations, pour déterminer la séquence d'états cachés. En 1970, les appellations « modèles de Markov cachés » ou « chaines de Markov cachées » (Hidden Markov Models) ont été Développés par L. P. Neuwirt afin de les renommer par « fonction probabiliste d'une chaine de Markov » utilisée jusqu'à une nouvelle étude (Slimane M., 2002).

L'utilisation des modèles de Markov cachés est apparue dans divers domaine à partir de l'année 1975, parmi ces domaines on peut citer la reconnaissance automatique de la parole (Rabinier L., 1989). Or, le groupe " International Business Machines" constitué de L. R. Bahl (Bahl L., 1975) et J. K Baker au CMU (Baker J., 1975) ont mis en main les travaux initiaux sur les modèles de Markov cachés afin d'être utilisés pour la reconnaissance automatique de la parole. Ainsi et grâce à ces travaux, on a pu reconnaitre les capacités de ces modèles pour la reconnaissance automatique de la parole. Les modèles de Markov cachés contenant des réseaux de neurones ont apparu en 1980 (Bourland et al. 1990). A partir de cette année ces modèles ont été employés dans le cadre de la reconnaissance de mots isolés ainsi que la parole continue (Rabiner et al. 1983; Bahl et al. 1983). En 1990 ont apparu les premières utilisations de la reconnaissance de l'écriture et de l'image (Siamaria et al. 1994). Or, les modèles de Markov cachés ont été employés dans les dernières années pour l'ordonnancement de tâches (Amini M., 2001).

De nos jours, les MMCs, sont un outil largement appliqué dans plusieurs domaines et sont considérés comme l'un des bases en termes d'efficacité et performances dans l'apprentissage automatique pour la modélisation de la reconnaissance vocale.

3. Les chaines de Markov discrètes

Une variable aléatoire « v. a. » réelle comme une fonction mesurable $X : \Omega \rightarrow \mathbb{R}$ a été définis pour le calcul des probabilités. Ω est appelé l'univers, celle-ci présente l'ensemble des réels $\mathbb{R}$ l'ensemble des entiers positifs $\mathbb{N}$ ou un de leurs sous-ensembles dans de nombreux cas de figures.

- **Processus stochastique** Un processus stochastique représente est une famille $\{X_t\}_{t \in T}$ de v. a. définies sur Ω.

$X_t : \Omega \rightarrow \mathbb{R}$

L'ensemble T désigne dans la plupart des cas la notion des temps mais il peut aussi être définit comme position spatiale en dimension 2 soit toute autre notion en autant de dimensions que nécessaires. On parle de processus stochastique en temps discret, Lorsque T correspond au temps et est discret, tandis que le processus est dit en temps continu, Quand T correspond au temps et est continu. Les états d'un processus stochastique défini par les v. a. $X_t : \Omega \rightarrow \mathbb{R}$ pour tout $t \in T$ sont les valeurs prises par ces v. a. lorsque t varie. On note S l'ensemble des « états » du processus.

A. A. Markov a été le premier à étudier et élaborer les principes mathématiques permettant l'étude des chaines qui portent son nom. Ces chaines ont été définies comme suit:

- **Condition d'une chaine de Markov** : Pour Qu'un processus $\{S_t\}_{t \in T}$ ($S_t : \Omega \rightarrow S$) soit une chaine de Markov il doit vérifier les trois conditions suivantes:

T est dénombrable ou fini. Dans ce cas et pour faciliter les notations postérieurs, il est toujours possible de prendre $T \subseteq N^* = \{1,2,\ldots\}$. Cette condition signifie que le processus ne change de valeur qu'à des instants déterminés a priori.

L'ensemble S des états du processus est dénombrable. Dans la suite, nous supposerons également que S est fini. Nous pouvons alors définir $S = \{s_1,\ldots, s_N\}$ cet ensemble.

Le processus est accordé à une fonction de probabilité P qui vérifier le modèle markovien: « la probabilité que l'état que ce processus à un instant t soit dépendu que de son état a l'instant $t - 1$ ».

Soit $Q = (q_t)_{t \in T}$ une suite d'états du processus $q_t \in S$. La propriété de Markov vérifie l'équation suivante, pour toute suite d'états Q et pour tout instant $t \in T$:

$$P(S_t = q_t / S_{t-1} = q_{t-1} \cdots S_1 = q_1) = P(S_t = q_t / S_{t-1} = q_{t-1}) \qquad \text{(eq. 3.1)}$$

La probabilité $P(S_t = q_t / S_{t-1} = q_{t-1})$ représente à la probabilité de transition de l'état q_{t-1} à l'instant t - 1 vers l'état q_t à l'instant t.

- **Homogénéité d'une chaine de Markov (dans le temps) :** Une chaine de Markov est dite homogène lorsque les probabilités de transition sont indépendantes du temps t « les probabilités de transition sont stationnaires »,

Ce qui signifie que pour tout $(t, t') \in T^2$, on a :

$$P(S_{t+1} = s_j / S_t = s_i) = P(S_{t'+1} = s_j / S_{t'} = s_i) \quad \text{(eq. 3.2)}$$

On note $a_{i,j}$ cette probabilité.

Une chaine de Markov est donc globalement caractériser par la donnée des états, des probabilités des états initiaux Π et des probabilités des transitions entre états A avec :

$$\Pi = \begin{pmatrix} \pi_1 \\ \vdots \\ \pi_N \end{pmatrix} = (\pi_1, \dots, \pi_N)' \quad \text{et} \quad \pi_1 = P(S_1 = s_j) \quad \text{(eq. 3.3)}$$

$$A = (a_{i,j})_{1 \leq i,j \leq N} \quad \text{et} \quad a_{i,j} = (S_{t+1} = s_i / S_t = s_j) \quad \text{(eq. 3.4)}$$

Ce sont des caractéristiques d'une chaine de Markov en général si de plus il est homogène c à d $a_{i,j}$ est indépendante du temps.

- **Vecteurs et matrices stochastiques:** un vecteur $V = (v_1, \dots, v_N)$ de dimension N est des stochastique lorsque:

  Pour tout $i = 1, \dots, N$ $0 \leq v_i \leq 1$, $\sum_{i=1}^{N} v_i = 1$

  Une matrice $M = (m_{i,j})_{1 \leq i,j \leq N}$ de dimension N*N est dite stochastique si et seulement :

  Si pour tout $i = 1, \dots, N$ et $j = 1, \dots, N$. $0 \leq m_{i,j} \leq 1$, $\sum_{j=1}^{N} m_{i,j} = 1$

  Pour tout $i = 1, \dots, N$

Particularité d'une chaine de Markov :

Une matrice est définie comme étant stochastique quand chaque ligne de cette matrice représente un vecteur de probabilité.

Pour que Π soit un vecteur stochastique, le système doit être obligatoirement dans un unique cas particulier au début. Ensuite, A est une matrice stochastique car le processus passe obligatoirement vers l'un des N états du système au temps $t + 1$ après avoir été dans un état S_i à l'instant t.

On peut associer une chaîne de Markov caractérisée par le couple (V, M) à tout couple constitué d'un vecteur stochastique V ayant une dimension N et d'une matrice stochastique M de dimensions $N \times N$.

- **Schéma représentant une chaîne de Markov** : On peut représenter une chaîne de

Markov graphiquement. Pour cette raison, on a la possibilité d'associer à la chaîne de Markov $\{S_t\}_{t \in T}$ une graphe *G* tel que l'ensemble des sommets et l'ensemble des états S sont en bijection entre eux et dont l'ensemble des arcs U est défini comme suit:

$$(s_i, s_j) \in U \leftrightarrow a_{i,j} > 0$$

On note S l'ensemble des sommets de graphe G (S_1= pluie, S_2= Neige, S_3= soleil). La figure 3.1 présente la structure graphique associée à la chaine de Markov (Π, *A*).

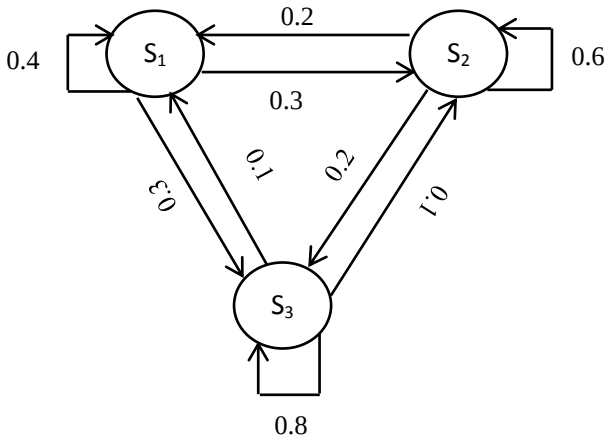


Figure.3. 1 : Représentation graphique de la chaine de Markov (*Π*, *A*).

$$A = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix} \quad \text{(eq.3.5)} \qquad \Pi = \begin{pmatrix} 0.02 \\ 0.4 \\ 0.58 \end{pmatrix} \quad \text{(eq. 3.6)}$$

Probabilité d'une séquence d'états:

$$P(S_n, S_{n-1}, \ldots, S_1) = P(S_n/S_{n-1}, \ldots, S_1) \times P(S_{n-1}, \ldots, S_1)$$

$$= P(S_n/S_{n-1}) \times P(S_{n-1}/S_{n-2}, \ldots, S_1) \times P(S_{n-2}, \ldots, S_1)$$

$$= P(S_n/S_{n-1}) \times P(S_{n-1}/S_{n-2}) \times \ldots \times P(S_2/S_1) \times P(S_1)$$

$$= P(S_1) \times \prod_{t-2}^{n} P(S_n/S_{n-1}) \quad \text{(eq. 3.7)}$$

Calculer la probabilité d'observer la séquence $S_3\ S_3\ S_3\ S_1\ S_1\ S_3\ S_2\ S_3$ sachant qu'aujourd'hui il fait à l'état S_3.

Le résultat de cette observation :

P (O/model) = P (S_3, S_3, S_3, S_1, S_1, S_3, S_2, S_3/model)

$$= P(S_3) * P(S_3/S_3) * P(S_3/S_3) * P(S_1/S_3) * P(S_1/S_1) *$$

$$P(S_3/S_1) * P(S_2/S_3) * P(S_3/S_2) \quad \text{(eq. 3.8)}$$

$$= 0.58 * a_{33} * a_{33} * a_{31} * a_{11} * a_{13} * a_{32} * a_{23}$$

$$= 0.58 * (0,8) * (0,8)(0,1)(0,4)(0,3)(0,1)(0,2)$$

$$= 8.9088 * 10^{-5}$$

4. Le Modèle de Markov Cachés

4.1. Les Modèles de Markov Cachés (MMC) (HMM)

Modèle Markov caché est une méthode efficace de caractérisation des échantillons de données observés d'une série temporelle discrète. Non seulement peut-il fournir un moyen efficace de construire des modèles paramétriques parcimonieux, mais peut également intégrer le principe de programmation dynamique dans son noyau pour une segmentation de modèle unifiée et une classification de modèle de séquences de données variant dans le temps. Les échantillons de données de la série temporelle peuvent être répartis de manière discrète ou continue; ils peuvent être des scalaires ou des vecteurs. L'hypothèse sous-jacente du MMC est que les échantillons de données peuvent être bien caractérisés en tant que processus aléatoire paramétrique et que les paramètres du processus stochastique peuvent être estimés dans un cadre précis et bien défini.

Un modèle de Markov caché discret est équivalent à la définition de deux processus stochastiques : un premier caché complétement modéliser par une chaine de Markov discrète et un seconde constaté dépendant des états du processus caché.

Soit $\mathbf{S} = \{s_1, \ldots, s_N\}$ l'ensemble des N états cachés du système. Soit $S = (S_1, \ldots, S_T)$ un T-uple de v.a définies sur $\mathbf{S}$. Soit $\mathbf{V} = \{v_1, \ldots, v_M\}$ l'ensemble M des symboles émissibles par le système. Soit $V = \{V_1, \ldots, V_M\}$ T-uple de v.a définies sur V.

Un modèle de Markov caché discret est déterminé par les probabilités :

- Les probabilités d'initialisation des états cachés : $P(S_1 = s_i)$
- Les probabilités de transition entre états cachés : $P(S_t = s_j / S_{t-1} = s_i)$
- Les probabilités d'émission des symboles dans chaque état caché :

$$P(V_t = v_j / S_t = s_j)$$

Les probabilités de transition entre états cachés et les probabilités d'émission des symboles dans chaque état caché sont indépendantes du temps $t > 1$, lorsque le modèle de Markov caché est invariable.

De ce fait en peut définir, pour tout $t > 1$ quelconque, $A = (a_{i,j})_{1 \leq i,j \leq N}$ avec

$a_{i,j} = P(S_t = s_j / S_{t-1} = s_i)$; $B = (b_i(j))_{1 \leq i \leq N, 1 \leq j \leq M}$ avec $b_i(j) = (V_t = v_j / S_t = s_j)$ et

$\Pi = (\pi_1, \ldots, \pi_N)'$ avec $\pi_i = P(S_1 = s_i)$. Le triplet (A, B, Π) détermine complétement Un modèle de markov caché invariable du premier ordre λ. Par la suite, nous emploierons la notation $\lambda = (A, B, \Pi)$ et le terme MMC pour des modèles de Markov cachés stationnaires du premier ordre. La figure schématisées les relations de dépendance entre les différentes variables aléatoires d'un MMC. Dans ce schéma, les flèches partent du v. a. qui détermine et se finalisent au niveau de la variable aléatoire conditionnée. Dans la figure 3.2, seules les transitions au temps t -1, t et t + 1 sont représentées.

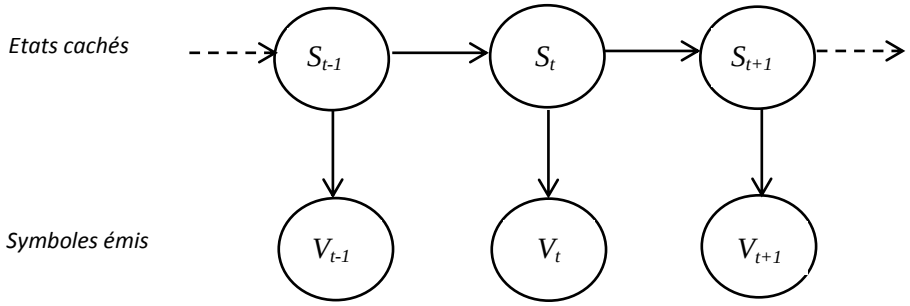


Figure.3. 2 : Les variables aléatoires d'un MMC et leur relation de dépendance.

Supposons que $Q = (q_1, \ldots, q_t) \in S^T$ est une séquence d'états cachés et $O = (o_1, \ldots, o_T) \in V^T$ est une séquence de symboles observés. La probabilité de réalisation de ces deux séquences Q et O par rapport au MMC **λ** est définie comme suit :

$$P(V = O, S = Q / A, B, \Pi) \qquad \text{(eq. 3.9)}$$

Or,

$$P(V = O, S = Q / \lambda) \qquad \text{(eq. 3.10)}$$

En employions les dépendances des probabilités conditionnelles, on constate que :

$$P(O, Q / \lambda) = P(O/Q, \lambda)\, P(S = Q / \lambda) \qquad \text{(eq. 3.11)}$$

Tel que,

$$P\,(V= O/ S = {}_Q\,, \lambda) = \prod_{t=1}^{T} P(\,V_t = o_t/S_t = q_t, \lambda)$$

$$= \prod_{t=1}^{T} b_{q(t)}\,(o_t)$$

$$= b_{q1}(o_1)\;\; b_{q2}(o_2)\ldots.\; b_{qt}(o_t) \qquad \text{(eq. 3.12)}$$

De même

$$P(S = {}_Q / \lambda) = P\,(S_1 = q_1/ \lambda) \prod_{t=1}^{T-1} P(S_t = q_t/S_{t-1} = q_{t-1}\,, \lambda)$$

$$= \pi_{q1}\; a_{q1q2}\; a_{q2q3}\; \ldots.\; a_{qt-1\,qt} \qquad \text{(eq. 3.13)}$$

Pour calculer l'adéquation entre le modèle $\boldsymbol{\lambda}$ et les deux séquences ${}_Q$ et *O*, nous dépendons sur les MMC $\boldsymbol{\lambda}$.

Pour cela, il suffit de calculer la probabilité $P\,(V = O,\; S = Q/\,\boldsymbol{\lambda})$ qui correspond à la probabilité déterminée par le modèle $\boldsymbol{\lambda}$.

Quand la séquence d'états cachés indéfinie, il peut examiner la vraisemblance d'une séquence d'observation *O* par rapport à un modèle λ. La vraisemblance correspond à la probabilité $P\,(V = O/\lambda)$ que la séquence d'observations ait été produite par le modèle pour l'ensemble des séquences d'états cachés possibles détermine la vraisemblance et on conclue que :

$$P\,(V = O/\,\lambda) = \sum_Q P\,(V = O,\; S = Q/\boldsymbol{\lambda}) \qquad \text{(eq. 3.14)}$$

Substituons les équations (3.12) et (3.13) dans (3.14) nous obtenons:

$$P\,(V= O/\,\lambda) = \sum_Q a_{q0q1}\;\; b_{q1}(o_1)\; a_{q1q2}\; b_{q2}(o_2)\; \ldots\; a_{qt-1\,qt}\; b_{qt}(o_t) \qquad \text{(eq. 3.15)}$$

4.2. Intérêt de MMC pour la reconnaissance automatique de la parole

La reconnaissance vocale est une chose très compliquée. Cependant, il en a plusieurs sources de variabilité (locuteur, environnement, accent, contexte, etc.), pouvant être modélisées suivant une quantité importante des données d'apprentissage. De plus, au niveau d'une base de données volumineuse, la manipulation manuelle reste impossible, il est par suite souhaitable que le système de reconnaissance vocale puisse généraliser automatiquement à partir d'une quantité de données importante. Or, les modèles de Markov cachés (MMCs) sont caractérisés par une telle capacité en faisant des hypothèses structurelles, et par utilisation d'une telle estimation paramétrique dans le but d'améliorer la probabilité générée

par les modèles grâce aux données d'apprentissage. Cette opération est alors caractérisée par plusieurs propriétés souhaitables, dont:

- Il demande une surveillance minimale.
- Seule une transcription orthographique du discours est nécessaire.
- Il possède une base mathématique permettant de garantir la convergence vers un point critique.
- Il est adapté à un entraînement accru, ne nécessitant que de calculs linéaires.

La nature probabiliste des MMCs en font la représentation idéale pour la reconnaissance de la parole. La reconnaissance vocale demande en effet une recherche dans un espace d'état dont la solution est optimale ou bien presque optimale. Or, les recherches basées sur les MMCs diffèrent des approches ascendantes propageant les erreurs et ne peuvent pas intégrer des connaissances de nature descendante, ainsi que de haut en bas. D'autre part, il en a les approches qui sont souvent intraitables. De plus, il est tout à fait possible de représenter des phonèmes, sons, mots, syllabes, et aussi des états de grammaire en termes de MMC. Par intégration de nombreux niveaux de connaissance, la recherche en MMC est une stratégie globale qui se focalise sur les objectifs, où toutes les sources de connaissances participent à chaque décision.

Enfin, par utilisation d'un cadre probabiliste, nous avons bien un mécanisme de notation cohérent. En résumé, les modèles de Markov cachés ont des propriétés très puissantes. La capacité des MMCs pour optimiser automatiquement les paramètres des données est extrêmement forte, la recherche intégrée de MMC liée aux connaissances sources à chaque étape est très efficace, et l'absorption des hypothèses structurelles défectueuses est le plus indulgent. Le MMC constitue l'un des paradigmes d'apprentissage les plus puissants de nos jours.

4.3. Topologie des Modèles de Markov Cachés

Beaucoup de topologies de modèles MMC ont été testées et sont actuellement utilisées en pratique, parmi ces topologies il existe le modèle de Bakis (Gauche-Droite) et le modèle ergodique, ces topologies permettons de limiter l'erreur au sens de la pertinence de symboles émis et en termes d'incertitudes dans sa conception.

- **Le Modèle ergodique** (figure 3.1). C'est un modèle sans contrainte sur les connexions entre les états. Chaque état peut être atteint à partir de n'importe quel autre état c'est-à dire

toutes les transitions d'un état vers un autre état sont possibles.

- **Le Modèle de Bakis** (Figure 3.3) C'est un modèle avec contrainte sur les connections entre les états (Rabinier L., 1989). La transition d'un état ayant un indice bas vers un état ayant un indice haut, c'est-à-dire n'autorise aucune transition d'un état d'indice supérieur vers un autre état d'indice inférieur (pas de retour arrière). Dans le cas de modèles de mots, ceux-ci sont souvent représentés par des MMCs à plusieurs états. Dont le nombre est parfois proportionnel au nombre de phonèmes dans le mot ou à la longueur de celui-ci. La topologie générale est toujours de type strictement gauche-droite avec les transitions possibles limitées à la boucle, au passage à l'état suivant et éventuellement au saut d'un état. Dans le cas de modèles de phonèmes, les MMCs sont souvent à 3 ou 5 états, le but initial étant d'avoir l'état central modélisant la partie stable du phonème alors que les deux états extrêmes modélisent la partie transitoire. Il a alors été observé qu'il est souvent préférable de ne pas permettre le saut d'état, de façon à également introduire une certaine contrainte sur la durée minimale de chaque unité phonétique, Des modèles de phonèmes plus complexes ont également été proposés.

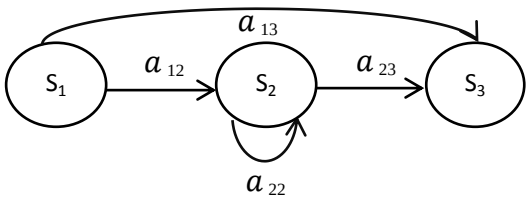


Figure.3. 3 : Modèles de Markov Cachés de type Bakis.

5. Mise en œuvre des Modèles de Markov Cachés

Pour utiliser les modèles Markov Cachés en reconnaissance de la parole, un certain nombre de problèmes pratiques doivent être résolus. Historiquement, les difficultés rencontrées avec ces problèmes ont retardé l'emploi des MMCs dans RAP, bien que la théorie des processus markoviens existe depuis longtemps. Les trois problèmes sont les suivants:

Le problème de l'évaluation de la vraisemblance : prenant un modèle et une séquence d'observations, comment calculer la probabilité que ce modèle ait produit cette séquence ?

Le problème du décodage : prenant un modèle et une séquence d'observations, quelle est la plus probable suite d'états ayant engendré les observations ?

Le problème de l'apprentissage : prenant un modèle et un ensemble de séquences d'observations, quels sont les paramètres du modèle qui conduise à maximiser la probabilité que l'ensemble des séquences ait été produit par le modèle ?

5.1. Evaluation de la vraisemblance

D'après l'équation précédemment (eq.3.15), calculer la vraisemblance d'une séquence de observations de duré T par apport à un λ consiste à évaluer la probabilité *P (V= O/ λ).* La complexité de cet algorithme étant de l'ordre $2NT^N$ opérations[1]. Ce calcul peut s'effectuer en utilisant différentes méthodes et cela en faisant introduire deux algorithmes supplémentaires, à savoir l'algorithme Forward qui procède à une estimation directe et l'algorithme Backward qui procède à une estimation rétrograde (Devijver P., 1985).

5.1.1. Estimation directe via l'algorithme Forward :

Cet algorithme de calcul fait appel à une variable intermédiaire $\boldsymbol{\alpha_t}$(i) qui résulte d'un produit d'un grand nombre de valeurs comprises entre 0 et 1. Conséquemment, après quelques observations, une dizaine ou vingtaine par exemple, on remarque que la valeur de cette variable tend exponentiellement vers 0. La variable Forward α est définie comme la probabilité que les observations jusqu'à l'instant *t* ont été émises par le modèle λ et que l'état atteint en *t* soit s_i :

$$\alpha_t(\text{i}) = \text{P}\ (V_1 = o_1,\ldots., V_t = o_t\ , S_t = \text{s}_\text{i}\ /\ \lambda) \qquad \text{(eq. 3.15)}$$

Le calcul de la probabilité $P\ (V = O/\ \lambda)$ se fait en utilisant une récurrence sur les variables α_t(i) pour tous les états *i*. La récurrence employée est définie par:

- **Initialisation**

$$\alpha_1(i) = \pi_i\ b_i(o_1) \qquad \text{Avec } 1 \le \text{i} \le N \qquad \text{(eq. 3.16)}$$

- **Induction**

---

1 Par exemple pour 5 états et une séquence de 100 observations, cela représenterait $2\times\ 100\times 5^{100} = 10^{72}$ séquences (opérations).

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i)\, a_{ij}\right] b_j(O_{t+1})$$

Avec $1 \leq t \leq T$ *et* $1 \leq j \leq N$ (eq. 3.17)

- **Finalisation**

$$P\,(V = O/\lambda) = \sum_{i=1}^{N} \alpha_T(i) \qquad \text{(eq. 3.18)}$$

Dans ce cas, le calcul de la probabilité $P\,(V = O/\lambda)$ ne nécessite que N^2T opérations, ce qui est optimal par rapport au calcul direct.

5.1.2. Estimation directe via l'algorithme Backward :

Le calcul de la probabilité $P\,(V = O/\lambda)$ peut effectuer à base d'une estimation rétrograde en faisant intervenir une variable $\beta_i(t)$ dite variable Backward. La fonctionnalité de cette variable est similaire à celui de Forward, à l'exception que celle de Backward utilise une estimation rétrograde dans le temps qui démarre à l'instant t = *T*, sa définition est donnée par :

$$\beta_i(t) = P\,(V_{t+1} = o_{t+1},\ldots,V_T = o_T\,, S_t = s_i/\lambda) \qquad \text{(eq. 3.19)}$$

Le calcul de β se fait par une récurrence sur le temps en partant de l'état final au temps *T* :

- **Initialisation**

$$\beta_T(i) = 1 \qquad \text{Avec } 1 \leq i \leq N$$

- **Induction**

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij}\beta_{t+1}(j)b_j(O_{t+1}) \qquad \text{(eq. 3.20)}$$

Avec $t = T-1, T-2, \ldots, 1$ *et* $1 \leq j \leq N$

- **Finalisation**

$$P\,(V = O/\lambda) = \sum_{i=1}^{N} \beta_t(i) \qquad \text{(eq. 3.21)}$$

L'algorithme Backward est modelé similairement à l'algorithme Forward. Cependant, il ne permet pas de calculer directement la probabilité $P\,(V = O/\lambda)$ La différence s'existe au niveau des observations dès l'instant initial. Autrement dit, dans l'algorithme Forward la première

observation de la séquence correspond à o_1 la deuxième à o_2, ainsi de suite jusqu'à la dernière correspondant à o_T, tandis que l'algorithme Backward la première observation correspond à o_T, la deuxième à o_{T-1}, ainsi de suite jusqu'à la dernière qui correspond à o_1. L'application commune des deux algorithmes (Forward et Backward) nous permet de calculer la probabilité $P\ (V = O/\lambda)$ d'observer la séquence O à chaque instant t. et aussi déterminer plus facilement la probabilité $\gamma_t(i) = P\ (S_t = s_i/V = O, \lambda)$ d'être dans l'état s_i à un certain temps t étant donné la séquence d'observations O.

$$P\ (S_t = s_i/V = O, \lambda) = \alpha_t(i)\ \beta_t(i) \qquad \text{(eq. 3.22)}$$

$$\gamma_t(i) = \frac{P\ (S_t = s_i/V = O, \lambda)}{P(V = O/\lambda)} = \frac{\alpha_t(i)\ \beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\ \beta_t(i)} \qquad \text{(eq. 3.23)}$$

5.2. Décodage

Le décodage ou la segmentation de séquences d'observations consiste à trouver la séquence d'états cachés qui a engendré une séquence d'observations (Lefevre F., 2000). Deux approches sont possibles. La première consiste à rechercher, à chaque instant, l'état qui a le plus probablement engendré le symbole observé. La deuxième approche (Viterbi) consiste à trouver la séquence complète d'états cachés qui a le plus probablement engendré la séquence d'observations.

5.2.1. Etats cachés les plus probables à chaque instant

Dans cette approche, on cherche la séquence $Q^* = (q_1^*, \ldots, q_T^*) \in S^T$ vérifiant, pour tout $t=1, \ldots, T$. l'équation :

$$q_t^* = argmax_{i=1\ldots N}\ P\ (V = O, S_t = s_i / \lambda) \qquad \text{(eq.3.24)}$$

Tout d'abord, d'après l'équation (3.15) il est nécessaire d'appliquer les variables Forward et Backward, Malgré sa formulation simple, pour trouver l'état caché le plus probable à chaque instant qui a une complexité en $O(N^2T)$. En outre, la séquence Q^*calculer peut être incompatible, dans le sens où $P\ (V=O, S=Q^*/\lambda) = 0$. Vraiment, il peut être une transition entre deux états s_i et s_j existe dans la séquence Q^*, alors que la probabilité $a_{i,j}$ est nulle.

5.2.2. Algorithme de Viterbi

Le décodage de Viterbi (Viterbi A., 1967) permet de déterminer l'alignement optimal d'une forme à connaître sur un modèle de Markov, c'est-à-dire le chemin qui conduit à la plus forte probabilité d'émission de la forme considérée**.** La recherche de la séquence d'états cachés Q^* qui le plus probablement engendré une séquence d'observations O consiste à résoudre :

$$Q^* = argmax_{Q \in S^T} P(V = O, S = Q / \lambda) \quad \text{(eq. 3.25)}$$

Il permet de trouver la meilleure séquence d'états, pour une séquence d'observation donnée o_1, . . ., o_T. Soit la quantité $\delta_t(i)$ avec i = 1, 2, ..., N, définie par :

$$\delta_t(i) = \max_{q_1 \ldots q_{t-1}} p\left(S_1 = q_1, \ldots, S_{t-1} = q_{t-1}, S_t = s_i, V_1 = o_1, .., V_t = o_t \middle| \lambda\right) \quad \text{(eq. 3.26)}$$

On définit $\delta_t(j)$ la probabilité maximale sur tous les chemins aboutissant à l'état j au temps t, l'algorithme de Viterbi permet de calculer $\{\delta_t(j), \forall j\}$ à partir de $\{\delta_{t-1}(j), \forall j\}$. A tous les instants, et pour tous les états du modèle, on va démontrer la formule suivante, explicitant le fait qu'un chemin de longueur t résulte du prolongement d'un chemin de longueur t-1 par une transition entre deux états et l'émission d'une trame :

$$\delta_t(j) = \max_{q_1 \ldots q_{1-1}} P\left(S_t = s_j \middle| S_{t-1} = s_i, \lambda\right) P\left(V_t = O_t | S_t = s_j, \lambda\right)$$

$$P\left(S_1 \ldots S_{t-2}, S_{t-1} = s_i, V_{t-1} = O_{t-1} | \lambda\right)$$

$$= \max_{S_i}\left[a_{ij} b_{S_j}(O_t) \delta_{t-1}(i)\right]$$

$$= \max_{S_i}\left[a_{ij} \delta_{t-1}(i)\right] b_{S_j}(O_t) \quad \text{(eq. 3.27)}$$

Nous Considérons $\psi_t(i)$ est le meilleur chemin de transition à l'état s_i au temps $\boldsymbol{t}$ à partir du temps t — *1*. L'algorithme de Viterbi est présenté par la procédure suivant:

- **Initialisation**

$$\delta_1(i) = \mu_i b_i(O_1) \; et \; \psi_1(i) = 0 \qquad \text{Avec } 1 \leq i \leq N$$

- **Induction**

$$\psi_t(j) = \arg\max_{\mathbf{1} \leq \mathbf{i} \leq \boldsymbol{N}} [a_{ij} \delta_{t-1}(i)] \quad \text{(eq. 3.28)}$$

Avec $2 \leq t \leq T$ *et* Avec $1 \leq j \leq N$

$$\delta_t(j) = \max_{\mathbf{1} \leq \mathbf{i} \leq \boldsymbol{N}} \left[\delta_{t-1}\left(\psi_t(j)\right) a_{\psi_t(j)j}\right] b_j(O_t) \qquad \text{(eq. 3.29)}$$

Avec $2 \leq t \leq T$ et Avec $1 \leq j \leq N$

- **Finalisation**

$$q_T^* = \operatorname{argmax}_{1 \leq i \leq N} \delta_T(i) \qquad \text{(eq. 3.30)}$$

$$\delta_t\left(q_T^*\right) = \max_{\mathbf{1} \leq \mathbf{i} \leq \boldsymbol{N}} \delta_T(i) \qquad \text{(eq. 3.31)}$$

- **Construction (séquence d 'états)**

$$q_t^* = \psi_{t+1}(q_{t+1}^*)$$

Avec $t=T-1$

5.3. Apprentissage

L'algorithme d'Expectation Maximization (EM) est une méthode itérative de calcul d'estimations du coefficient de distribution du maximum de vraisemblance à partir de données incomplètes (éléments manquants dans les vecteurs de caractéristiques). Les équations de mise à jour EM qui donnent une action sont utilisées pour maximiser la probabilité que les données d'entraînement soient enregistrées à plusieurs reprises selon le modèle.

La procédure d'apprentissage à base du critère du maximum de vraisemblance (en anglais Maximum Likelihood Estimation) (Celeux G., 1995), qui sera noté MLE, a pour objectif de chercher le meilleur ensemble de paramètres λ' qui maximise la probabilité d'émission de la séquence d'observations *O*.

$$\lambda' = \ argmax_{\lambda} P\ (V = O/\ \lambda) \qquad \text{(eq. 3.32)}$$

L'équation précédente est résolue d'une manière itérative en pratiquent les formules de Baum-Welch qui est dérivée de l'algorithme EM (Expectation Maximization). Les équations de Baum-Welch donnent des formules consistent à augmenter la valeur des paramètres très probables et à diminuer celle de ceux peu probables (Aupetit S., 2005).

L'idée principale de l'algorithme de Baum-Welch basée sur l'estimation des nouveaux paramètres du modèle λ', à partir du modèle initial λ afin que la probabilité soit maximale où. $P\ (V = O/\ \lambda') > P\ (V = O/\ \lambda)$. La méthode de calcul est réalisée d'une manière itérative telle que paramètres du nouveau modèle vont utiliser comme une initiation pour l'itération suivante et ainsi de suite, jusqu'à atteindre un seuil de convergence prédéfini.
La démonstration de La convergence vers un optimum local a été réalisée, Cependant, les valeurs initiales des paramètres à estimer sont crucial pour assurer une convergence exact et rapide la plus tôt possible du maximum global.

- ***Ré-estimation à base de l'algorithme de Baum-Welch***

Pour la ré-estimation de paramètres à base de l'algorithme de Baum-Welch (Baggenstoss P., 2001), deux variables doivent être introduites, à savoir:

$x_t(i,j)$ représente la probabilité d'être dans l'état s_i au temps *t* et dans l'état s_j au temps *t+1* sachant le modèle λ' et la séquence d'observation *O*.

$$x_t(i,j) = P\ (S_t = s_i, S_{t+1} = s_j / V = O, \lambda') \qquad \text{(eq. 3.33)}$$

$y_t(i,j)$ représente la probabilité d'être dans l'état s_i au temps t sachant le modèle λ' et la séquence d'observation O.

$$y_t(i,j) = P\ (S_t = s_i / V = O, \lambda') \qquad \text{(eq. 3.34)}$$

D'après l'équation (3.33) on obtient :

$$x_t(i,j) = \frac{P\ (S_t = s_i, S_{t+1} = s_j, V = O/\lambda')}{P(V = O/\ \lambda')} \qquad \text{(eq. 3.35)}$$

D'ailleurs à partir de l'équation (3.34) on obtient :

$$y_t(i,j) = \frac{P\ (S_t = s_i / V = O, \lambda')}{P(V = O/\ \lambda')} \qquad \text{(eq. 3.36)}$$

Basant sur les variables des algorithmes Forward et Backward décrits précédemment :

$$x_t(i,j) = \frac{\alpha_t(i)\ a_{ij}\ \beta_{t+1}(j)\ b_{S_j}(O_{t+1})}{P(V = O/\ \lambda')} \qquad \text{(eq. 3.37)}$$

$$y_t(i) = \frac{\alpha_t(i)\,\beta_t(i)}{\mathrm{P}(V=O/\,\lambda\prime)}$$

$$= \frac{\sum_{j=1}^{N} \alpha_t(i)\; a_{\mathrm{ij}}\; \beta_{t+1}(j)\; b_{S_j}(O_{t+1})}{\mathrm{P}(V=O/\,\lambda\prime)}$$

$$= \sum_{j=1}^{N} \frac{\alpha_t(i)\; a_{ij}\; \beta_{t+1}(j)\; b_{S_j}(O_{t+1})}{P(V=O/\,\lambda\prime)}$$

$$= \sum_{j=1}^{N} x_t(i,j) \qquad \text{(eq. 3.38)}$$

La somme de deux variables au cours de temps peut être interprétée comme :

$\sum_{\mathrm{t=1}}^{\mathrm{T-1}} y_t(i) = \sum_{\mathrm{t=1}}^{\mathrm{T-1}} P\,(S_t = s_i/V = O, \lambda')$= nombre de transitions depuis l'état s_i

$\sum_{\mathrm{t=1}}^{\mathrm{T-1}} x_t(i,j) = \sum_{\mathrm{t=1}}^{\mathrm{T-1}} P\,(S_t = s_i/V = O, \lambda')$= nombre de transitions depuis l'état s_i vers s_j

En utilisant ces deux sommes, les nouveaux paramètres ré-estimés $\lambda^* = (A^*, B^*, \Pi^*)$ pour un model $\lambda = (A, B, \Pi)$ sont :

$$\Pi^* = y_1(i) = \text{nombre de passage par l'état } s_i \text{ au temps } t=1$$

$$a_{ij}* = \frac{\sum_{\mathrm{t=1}}^{\mathrm{T-1}} x_t(i,j)}{\sum_{\mathrm{t=1}}^{\mathrm{T-1}} y_t(i)} = \frac{\text{nombre de transitions de l}'\text{état } s_i \text{ a l}'\text{instant t=1}}{\text{nombre de fois où l}'\text{on quitte l}'\text{état } s_i}$$

$$b_j(k) = \frac{\text{nombre d}'\text{apparition simultanées de l}'\text{état } s_j \text{ et du symbole } v_k}{\text{nombre de fois où l}'\text{on quitte l}'\text{état } si}$$

On peut observer que l'algorithme d'apprentissage avec les approches "d'étiquetage" (voir annexe B) est similaire à celui "Baum-Welch", mais il existe une petite différence au niveau de la priorité de calcul de la probabilité, tel que, l'apprentissage avec "l'étiquetage" s'effectue avant la probabilité de ré-estimation.

La complexité de l'algorithme de Baum-Welch est $O\,(N^2T + NMT)$.

Les probabilités exploitées pour ré-estimer les matrices qui peuvent être obtenues par les méthodes Forward et Backward. Cependant, toujours pour des problèmes de la mise en œuvre numérique, on utilise plutôt leurs versions qui ont basé sur les algorithmes avec ré-échelonnement.

L'algorithme de Baum-Welch est présenté par la procédure suivant:

Choisir un modèle initial λ_0 $t = 0$

Répéter $t = t + 1$

Calculer les variables Forward et Backward pour le modèle λ_{t+1}

Calculer Π de λ_t

Calculer A de λ_t

Calculer B de λ_t

Tant que ($P(V = O/\lambda_t) > P(V = O/\lambda_{t-1})$) et $t < t_{max}$

6. Conclusion

Dans ce chapitre nous avons présenté brièvement les principes théoriques des modèles MMCs et leur intérêt d'utilisation dans la reconnaissance automatique de la parole, ainsi que l'élaboration des trois problèmes fondamentaux adressés lors du développement d'un système de RAP à base de MMC et les algorithmes de l'évaluation de la vraisemblance, de décodage et d'apprentissage utilisés dans leur résolution.

Dans le chapitre suivant, nous allons présenter la langue sur laquelle nous nous sommes basés dans cette thèse. Nous dériverons ses propriétés et caractéristiques spécifiques.

Chapitre 4: Propriétés et caractéristiques de la langue Amazighe

1. Introduction

L'objectif de ce chapitre est de présenter les informations globales, les caractéristiques et les propriétés liées à la langue Amazighe au Maroc, dont la phonétique, le système d'écriture, la morphologie, l'utilisation informatique de Tifinagh et le processus de la mise en œuvre de ce dernier, lesquels nous ont permis de mieux comprendre la langue et également la construction d'un système aussi puissant et capable de fournir des données plus efficaces.

2. La distribution géographique de la langue Amazighe

La langue Amazighe est considérée comme une langue polyvalente vu son aspect pluriel qui regroupe une pléthore de dialectes parlés dans différents pays d'Afrique du Nord et du Sahel y compris le Maroc, l'Algérie, le Niger, le Mali, la Libye, l'Egypte et d'autres. Des régions plus ou moins petites témoignent de la présence berbère en Tunisie et en Libye [(Galand L., 1988), (Chaker S., 1992)]. Ainsi, la langue amazighe est peu utilisée dans l'oasis de Siwa en Egypte car elle est parlée par un nombre restreint de population. Un autre groupe berbérophone s'ajoute à ces communautés: Les Touaregs résidant au Mali, au Sud de l'Algérie, au Niger et aussi Burkina Faso. Le berbère est principalement parlé en Algérie et au Maroc. Or, il est difficile de mentionner des chiffres bien définis en ce qui concerne les populations berbérophones. Nous signons encore qu'au Maroc le berbère est parlé par 35% jusqu'à 40% de la population. Pourtant, ce pourcentage se rétrécit en 25% en Algérie [(Galand L., 1988), (Chaker S., 1995)].

La plupart des linguistes Amazighes sont unanimes sur le fait qu'il existe généralement trois régions linguistiques berbères au Maroc (Ridouane R., 2003) :

- Première zone: Nord et Nord-Est où l'on parle le tarifit, du nom de la chaîne montagneuse 'le Rif'.
- Deuxième zone: Sud-Est où l'on parle le tamazight, appelé aussi le berbère.
- Troisième zone du chleuh ou le chleuh est parlé dans le Sud et le Sud-Ouest du Maroc.

Cette répartition montre une difficulté à tel point qu'il ne s'agisse pas d'un système des isoglosses entre ces trois dialectes. Ceci est dû plutôt aux clivages géographiques, mais non linguistiques qui définissent ces dialectes. L'intercompréhension entre ces trois dialectes est contrainte à des limites. Les divergences concernent plus particulièrement le système phonologique et lexical. En dépit de ces divergences entre ces dialectes, les données structurales fondamentales sont les mêmes.

3. Propriétés de la langue Amazighe

3.1. Inventaire des phonèmes de l'amazigh standard

3.1.1. Système d'écriture

Tableau.4. 1 : Tableau officiel des alphabets de la langue Amazighe avec leurs syllabes et leur transcription en anglais et en arabe (Ameur el al. 2004).

Tifinagh	English transcription	Arabic transcription	Syllabes	No. Of syllabe	Tifinagh	English transcription	Arabic transcription	Syllabes	No. Of syllabe
ⵄ	YA	يا	CV	1	ⵍ	YAL	يال	CVC	1
ⵅ	YAB	ياب	CVC	1	ⵎ	YAM	يام	CVC	1
ⵇ	YAG	ياڤ	CVC	1	ⵏ	YAN	يان	CVC	1
ⵇⵯ	YAGG	ياڤ	CVC	1	ⵓ	YU	يو	CV	1
ⵃ	YAD	ياد	CVC	1	ⵔ	YAR	يار	CVC	1
ⵄ	YADD	ياض	CVCC	1	ⵕ	YARR	يارً	CVC	1
ⵉ	YEY	ياي	CVC	1	ⵖ	YAGH	ياغ	CVC	1
ⴼ	YAF	يافـ	CVC	1	ⵙ	YAS	ياس	CVC	1
ⴽ	YAK	ياك	CVC	1	ⵚ	YASS	ياص	CVC	1
ⴽⵯ	YAKK	ياكك	CVCC	1	ⵛ	YAC	ياش	CVC	1
ⵀ	YAH	ياه	CVC	1	ⵜ	YAT	يات	CVC	1
ⵃ	YAHH	ياح	CVC	1	ⵟ	YATT	ياط	CVC	1
ⵄ	YAAA	ياع	CVC	1	ⵡ	YAW	ياو	CVC	1
ⵅ	YAX	ياخ	CVC	1	ⵢ	YAY	ياي	CVC	1
ⵇ	YAQ	ياق	CVC	1	ⵣ	YAZ	ياز	CVC	1
ⵉ	YI	يي	CV	1	ⵣ	YAZZ	ياز	CVCC	1
ⵊ	YAJ	ياج	CVC	1					

La langue Amazighe Standard possède 33 phonèmes. Ces phonèmes sont classés sous trois catégories : 27 consonnes, 2 semi-consonnes et 3 voyelles pleines et une voyelle neutre.

En prenant l'exemple du système de l'IRCAM[2] qui a réalisé un système alphabétique sous le nom Tifinaghe-IRCAM s'écrivant de gauche à droite. Ce standard alphabet se base sur un système graphique à orientation phonologique. Pourtant, il n'a pas la capacité de mettre en garde toutes les réalisations phonétiques produites, mais seulement celles qui sont fonctionnelles. Le tableau précédent (Tableau 4.1) présentés les phonèmes Amazighe et leur transcription arabe et anglaise.

3.1.2. Les voyelles et les consonnes de la langue Amazighe

- **Les Voyelles**

Le système de la langue Amazighe est composé de quatre voyelles : Trois voyelles pleines □, □ et □ (a, i et u) et une voyelle neutre (ou schwa) □(e) qui est caractérisée par un statut très particulier au niveau phonétique amazighe. Parmi ces quatre voyelles, si le u à une prononciation invariable, ce n'est pas le cas pour le a, le e et le i. Chacune de ces trois voyelles se distinguent en effet par deux prononciations distinctes. Cette différentiation d'ordre phonétique n'affectera pas l'écriture de ces trois voyelles, l'usage seul décidant de la prononciation de chacune d'elles, ainsi qu'il apparaîtra à travers les exemples ci-après. [(Arbouz C., 2016), (Satori H., 2015)]

La voyelle «a» se prononcera d'une manière claire selon l'éloignement des mâchoires ou leur rapprochement. Pour cela, dans le mot tavla (table), le premier « a » se prononcera comme dans le mot français « tamis » et le second le sera comme dans le mot anglais « latter » (dernier). La voyelle «e» peut être courte ou longue selon la situation. Pour cette raison, le « e » du mot ilef (sanglier) sera bref, et celui de *sers* (pose) sera long et appuyé. La voyelle «i» du mot thira (l'écriture) se prononcera de la même façon que le «i» du mot français livre. Cependant, le « i» se prononcera dans certains cas presque comme le «ei» du mot français peine. Exemple : thajehnit (une queue). La voyelle «u» se prononcera toujours comme le «ou» de la langue française. Exemple : afus (la main) (Arbouz C., 2016).

L'amazigh est une langue qui est basée dans certains cas sur la consonne seulement, alors que les langues latines ou anglo-saxonnes sont syllabiques. Ainsi on ne peut pas trouver dans Ces langues de consonnes indépendantes, ce qui veut dire que toute consonne dépend en amont ou en aval d'une voyelle. D'autre part, on trouve dans les langues autres que l'amazigh,

[2] L'Institut royal de la culture amazighe ou IRCAM est un institut académique de l'État marocain chargé de la promotion de la culture amazighe et du développement de la langue berbère. L'institut a son siège à Rabat, il a été fondé en octobre 2001

des voyelles isolées. Par exemples, en français, nous avons « a », « eu », « ou », « au » et « ai », en espagnol, nous avons « y » et « a », en anglais « a » et la diphtongue « I ». En amazigh, on trouve tellement des consonnes isolées (n, d, s), des doubles et triples consonnes isolées (rs, ls, ns, ml, zd, ... et : frn, brn, frs, krs, srs etc.). On a même des mots dont quatre ou cinq consonnes se suivent (frfr, sfrfd, msbrid, msflid, etc). Pourtant, les voyelles isolées sont presque inexistantes. Le système vocalique de la langue Amazigh est présenté par le tableau ci-dessous (Tableau 4.2).

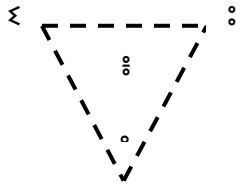
Lieu d'articulation / Degré d'aperture	Antérieures	Postérieures
Aperture minimale		
Aperture maximale		

Tableau.4. 2 : Le système vocalique de l'Amazigh standard (Dameur el al. 2004).

Remarque 2 : un schwa prononcé ne sera noté que dans deux cas :

- dans des suites de plus de deux consonnes identiques (□□□□□) "elle a demandé").
- dans les radicaux verbaux se terminant par deux consonnes identiques (□□□□ "être blanc").

➢ **Les Consonnes**

Les consonnes d'abord peuvent être prononcées avec une fermeture locale ou une étroitement du conduit vocal. Ensuite deux types différents de consonne qui existent, des consonnes sonores et d'autres sourdes, selon que l'air provenant des poumons est modulé par les cordes vocales ou non. Les autres facteurs de classement des consonnes sont le mode d'articulation et le lieu d'articulation.

Les consonnes de la langue amazighe sont distinguées selon leur mode d'articulation (occlusif, nasales, etc..), leur lieu d'articulation (labials, dentals, etc..) et leur Voisement (sonores ou sourds) (voir tableau 4.3). Or, les consonnes sont dites sonores ou voisées quand les cordes vocales participent à l'émission du son et vibrent. Des consonnes comme [□, ⵖ, □] sont sonores, et les consonnes comme [□, ⵃ, ⵛ] sont sourdes.

Tableau.4. 3 : La phonologique des consonnes de l'amazigh standard (Boukous A., 1995).

Mode d'articulation \ Lieu d'articulation			Labiales	Dentales	Alvéolaires	Palatales	Vélaires	Labiovélaires	Uvulaires	Pharyngales	Laryngale
Occlusives	Non emphatiques	Sourdes		ⵜ			ⴽ	ⴽᵘ	ⵇ		
		Sonores	ⴱ	ⴷ			ⴳ	ⴳᵘ			
	Emphatiques	Sourdes		ⵟ							
		Sonores		ⴹ							
Constrictives	Non emphatiques	Sourdes	ⴼ		ⵙ	�right			ⵅ	ⵃ	ⵄ
		Sonores			ⵣ	ⵊ			ⵖ	ⵃ	
	Emphatiques	Sourdes			ⵚ						
		Sonores			ⵣ						
Nasales			ⵎ	ⵏ							
Vibrantes	Non emphatiques			ⵔ							
	Emphatiques			ⵕ							
Latérale				ⵍ							
Semi-consonnes			ⵡ			ⵢ					

En effet, la production des consonnes comporte l'émission d'un bruit de constriction ou d'explosion. De ce point de vue, les consonnes sont des bruits, qui évoquent des explosions ou des frottements, produits par le souffle heurtant divers organes (points phonatoires d'articulation).

De manière générale, les consonnes sont considérées par une énergie plus faible que les voyelles. Le premier formant monte généralement pendant la transition d'une consonne à une voyelle (et inversement décroît pendant la transition d'une voyelle à une consonne) particulièrement si la consonne est sonore.

Remarque 1 : la gémination (ou tension) concerne toutes les consonnes; elle est rendue, au niveau de l'écrit, par le dédoublement du graphème. Pour les labiovélaires

géminés, seul le deuxième graphème porte l'indice de la labiovélarisation (ⴽⴽᵘ et ⵅⵅᵘ) (Arbouz C., 2016).

3.1.3. Critères retenus dans l'élaboration de l'alphabet

Les phonèmes qui constituent l'alphabet de l'amazighe sont choisis à partir d'une analyse phonologique et sont basés sur les points suivants :

- L'univocité du signe: un graphème pour un son et un son pour un graphème.
- L'extension géographique : une particularité phonétique très localisée ne peut pas être retenue dans le système graphique.
- Le rendement fonctionnel: si elle est isolée et peu productive, une opposition de deux phonèmes ne peut prétendre à un statut phonologique. Elle relèvera de la variation régionale.
- La neutralisation de la variation linguistique de surface toutes les différences phonétiques superficielles (et n'ayant donc pas d'incidence sur l'intercompréhension entre les usagers de la langue) ne seront pas prises en compte par le système graphique. Par contre, différentes latitudes de réalisation restent possibles au niveau du code oral.

La syllabe est une unité fondamentale de la structure suprasegmentale. Elle joue un rôle très important en linguistique ou elle est concédée comme une base de regroupement de phonèmes dans la chaine prononcée (Boukous A., 1995). C'est une chaine d'une ou plusieurs consonnes tournant autour d'une voyelle qui peut à elle seule d'attacher au lieu de syllabe, et la syllabe peut être ouverte. La voyelle peut être survenue d'une ou plusieurs consonnes alors la syllabe est fermée. La structure de la syllabe est définie par un ensemble de règles qui changeant avec le changement de la langue. La langue amazighe support 7 types de syllabes classées selon les traits ouvert et fermé. Une syllabe est considérée fermée si elle contient une coda. Sinon elle est ouverte. Exemple le mot « amya: zero » inclut deux syllabes : [am] est une syllabe fermée et [ya] est une syllabe ouverte. Le système syllabique de l'amazigh a les caractéristiques suivantes (Ridouane R., 2003).

– Une syllabe en Amazigh contient une voyelle ou bien ou une consonne comme noyau.

– Une syllabe doit avoir une attaque (sauf après pause où elle peut ne pas contenir d'attaque). Exemple : la première syllabe de [ifta] ne contient pas d'attaque : elle commence par un noyau /i/.

– Les attaques composées ne sont pas autorisées, pourtant une attaque contient une

seule consonne pas deux ou trois, etc.

– La sonorité joue un rôle principal dans la structuration du noyau (les éléments les plus sonores ont priorité pour occuper la position de noyau.

– L'Amazigh contient des syllabes de type : CV, CVCC, CCVCC, CCVC, V, VC, CVC.

3.2. Morphologie de langue Amazighe

La langue Amazighe considérée par une morphologie riche. Elle est également présentée comme étant une langue complexe dont les mots sont classés selon trois formes morphosyntaxiques différentes: Nom, Verbe et Particules [(Boukhris et al. 2008), (Nejme et al. 2013)].

3.2.1. *Nom*

En Amazighe, le nom est défini comme étant une unité lexicale qui se compose par une un schéma et également une racine. Il regroupe deux caractéristiques, la première prend des formes à savoir : Une forme simple (argaz [argaz] "homme"), forme composée (buhyyuf [buhyyuf] "la famine") ou bien forme dérivée (amsawa [amsawaḍ] "la communication"). La deuxième caractéristique liée à la variation : Elle varie en genre, en nombre et en état (libre, annexion).

- Le genre

Le nom Amazighe tient en compte deux genres, le féminin et le masculin.

- Le nom féminin: Qui est de la forme [t....t], sauf quelques noms qui ne portent que le [t] au début du nom ou bien à la fin du morphème du féminin: [tadla] "gerbe", [ṛmuyt] "fatigue".

- Le nom masculin : En générale il commence par une des voyelles : [a], [i] ou bien [u], par exemple : [argaz] "homme", [ixf] "tête" et [udm] "visage".

En général, le féminin est composé à partir du radical d'un nom masculin par l'ajout du morphème discontinue [t....t], par exemple : [isli] "marié" -> [tislit] "mariée".

- Le nombre

Le nom Amazighe est généralement composé par un singulier et un pluriel. Ce dernier est obtenu de sa part des types suivants: Le pluriel externe, interne, mixte et le pluriel en id [id].

- Le pluriel externe : Il est acquis par une alternance de voyelles ayant un suffixe de [n] ou l'une des variantes ([in], [an], [ayn], [awn], [wan], [win], [yin]), par exemple : [tarbat] -> [tirbatin] “filles”.

- Le pluriel interne: Il est obtenu de sa part par un changement de voyelle internes et une alternance vocalique ([adrar] -> [idurar] “montagnes”).

- Le pluriel mixte : Il s’agit d’un pluriel composé par une alternance d'une voyelle interne et / ou d'une consonne plus un suffixe par [n] ([izri] - <[izran] “couplets”), ou par une succession voyelle initiale en fonction d'un changement final de voyelle [a] plus une alternance interne ([amggaru] - <[imggura] “derniers”).

- Le pluriel en [id]: Ce type du pluriel est acquis par une préfixation de [id] du nom au singulier. Il est également pratiqué à une collection de cas de noms tels que: Un nom ayant une consonne initiale, nom propre, nom de parenté et également un nom composé et des numéral...

- L’état

Nous classifions deux états différents pour les noms Amazighs, tels que l’état libre et l’état d’annexion.

– L’état libre: Dans ce cas, la voyelle initiale du nom ne contient pas de modification. Le nom est considéré en état libre s’il s’agit des cas suivants: Mot isolé de tout contexte syntaxique, complément d’objet direct, ou bien d’un complément de la particule prédictive.

- L’état d’annexion: Il est fondé sur un changement de l’initiale du nom dans des contextes syntaxiques définis. De plus, l'état d'annexion prend l'un des types suivantes: Alternance vocalique [a] et [u] ou maintien de la voyelle initiale et ajout d’un [w] dans le cas des noms masculins. Dans le cas des noms féminin, cet état est défini par la chute ou la conservation de la voyelle initiale.

3.2.2. *Verbe*

En langue Amazighe, le verbe peut être présenté sous les formes suivantes: Simple ou dérivée. Le premier est composé de la racine et du radical. Pourtant, le verbe dérivé est constitué des verbes simples et une préfixation d’un morphème: [s]/ [ss], [kk] et [n]/ [nn].

La première forme correspond à la forme factitive, la deuxième liée à la forme passive et la troisième montre la forme réciproque.

3.2.3. *Particules*

Les particules sont présentées comme étant une collection de mots Amazighs qui ne correspondent ni à des noms ni à des verbes, et qui ont un rôle dans la signification d'une phrase. De plus, cette collection est constituée de plusieurs parties telles que: Les particules d'orientation, de la négation, possessifs et interrogatifs, les pronoms personnels autonomes, les pronoms démonstratifs, les pronoms indéfinis, les affixes du sujet, les affixes d'objet direct et indirect, les compléments prépositionnels, les compléments du nom de parenté et ordinaire, les adverbes du temps, du lieu, de quantité et de manière, les subordonnants, les prépositions et les conjonctions (Boukhris et al., 2008).

En effet, les particules sont invariables. Cependant, dans le cas de l'Amazighe, également dans le cas de la langue française, on trouve des particules flexionnelles à savoir les pronoms possessifs.

4. Utilisation informatique de Tifinagh

A partir du moment où on a adopté le Tifinagh comme graphie officiellement attestée au Maroc pour la langue Amazighe, l'encodage Tifinagh est d'une nécessité principale (Ataa Allah et al. 2012). Ainsi, le centre des études et systèmes d'information et de communication de graphie officielle au Maroc IRCAM a fourni des efforts importants pour la langue Amazighe. Ces efforts déployés ont abouti à un codage uniforme et standardisé constitué de quatre sous-ensembles de caractères Tifinaghe à savoir le groupe de base et l'ensemble étendu de l'IRCAM, ainsi que d'autres lettres néo-Tifinaghe et aussi des lettres Touareg modernes. Les deux premiers sous-types composent les différents caractères choisis par l'IRCAM.

La langue Amazighe et grâce à la confirmation de Tifinagh a pu entrer en telle ou telle sorte dans l'ère de mondialisation liée au traitement et à l'utilisation de l'information comme les autres systèmes d'orthographe à travers le monde, d'où est l'obligation d'établir des normes de codage. D'autre part, et pour garantir l'unicité de l'identifiant numérique, le Consortium Unicode a développé un standard informatique, ainsi on aura vraiment l'opportunité d'avoir un caractère de propriété tout à fait différente des autres, sans prendre en considération la plate-forme ou le logiciel utilisé. Cependant, les spécialistes [(Andries P., 2008), (Silberztein M., 2007)] déclarent que la bonne lecture des codes des lettres dépend de

l’association avec des polices qui fournissent des images visuelles, ou des glyphes, liés aux codes. Ainsi, la lettre est considérée comme une unité abstraite d'informations textuelles. Le glyphe présente la forme graphique des lettres. Pourtant, le codage a généralement connu dès l'avènement d'Internet, un problème au niveau de l’utilisation de l’information pour l’affichage textuel. Pour cette raison, on trouve divers codes qui ont été développés dont l'utilisation est limitée au niveau international, comme le cas du code ASCII ou UTF-8. Unicode est devenu dans les dernières années la norme du codage du texte dans la plupart des langues (Zenkouar L., 2004). Or, Unicode a pour but de développer un code universel afin de coder les lettres de manière plus efficace.

Grâce au codage de Tifinagh, la langue Amazighe peut donc mettre en place une standardisation des produits électroniques et de donner un code informatique pour les différentes lettres dans le but de pouvoir assurer l'échange de documents électroniques de manière efficace (El Yachi el al. 2010). De plus, l'écriture Tifinagh peut être, dans ce cadre, intégrée dans des logiciels informatiques. Enfin, cette évolution liée à Tifinagh permettra de la rendre connue au niveau international.

5. La mise en oeuvre de Tifinagh

Dans l'enseignement, l'apprentissage du Tifinagh offre les résultats suivants:

Motivation de la communauté Amazighophone pour apprendre Tifinagh en leur indiquant que ce standard est une partie du patrimoine culturel du Maroc, ce qui lui permet la légitimité historique.

Monstration aux étudiants que la langue Amazighe, basée sur le standard Tifinagh, ressemble en quelques part aux langues latines, car elle s'écrit de gauche à droite et ne nécessite pas l'ajout des diacritiques à certains graphèmes comme le cas de la langue arabe. De plus, il existe des caractères en Tifinagh qui n'ont pas des lettres correspondantes en langue arabe.

En somme, malgré tous les efforts fournis pour développer le Tifinagh et le rendre plus connue et employé. La langue amazighe rencontre toujours des problèmes dans ce cadre. Ce qui ne lui permet pas d'atteindre le niveau des autres langues fameuses au niveau international comme le Français, l'Arabe, etc. Ainsi, recommandation est impérative. De plus, dans la planification linguistique, la phase d'implémentation d'un produit standard nécessite une vision évolutive caractérisée par la flexibilité imposée par les exigences de suivi, de test et de régulation, notamment dans le domaine de l'enseignement de l'apprentissage des langues.

6. Conclusion

Dans ce chapitre, nous avons indiqué les différents détails liés à la langue Amazighe qui consistent dans l'attestation de rendre le Tifinagh comme graphie officiellement attestée au Maroc, également l'intégration du standard Tifinagh dans le domaine de l'informatique, et discussion de la structure de la syllabe qui va nous servir dans la réalisation de notre système de reconnaissance automatique.

Dans le chapitre suivant, on va présenter notre contribution pour la création d'un système de reconnaissance vocale de la langue Amazigh, ainsi que la présentation de notre extension de configuration pour implémenter le modèle explicite proposé.

Chapitre 5 : Implémentation de Système de Reconnaissance de la Langue Amazighe

1. Introduction

Dans ce chapitre, nous allons décrire les différentes étapes qu'on a suivies pour la réalisation d'un système de la reconnaissance de la parole Amazigh, ainsi que la description de la base de données qui contient les dix premiers chiffres Amazigh (Amdigits). D'autre part, nous présentons nos expériences pour mieux adapter la langue Amazigh en utilisant la base de données amdigits, une procédure détaillée est donnée pour installer, modifier et faire l'apprentissage avec le système SphinxTrain. A cet égard, plusieurs paramètres seront ajustés pendant l'apprentissage comme : nombre d'états par MMC, nombre de distributions de probabilités Gaussiennes et les coefficients MFCC (Mel-Frequency Cepstral Coefficients).

2. Base de données Audio

2.1. Préparation du corpus :

Les corpus de parole (ou corpora) jouent un rôle fondamental dans la création des systèmes de RAP du fait que ces systèmes utilisent la parole comme entrée et / ou sortie. La plupart de ces systèmes sont actuellement basés sur des modèles statistiques qui exigent, pour leur apprentissage et leur test, une quantité énorme de données audio enregistrées et préparées d'une manière qualitative. On a collecté une base de données vocale nommée Amdigits qui nous permettra l'évaluation d'algorithmes pour une application de la reconnaissance des chiffres de 0 à 9 connectés à locuteurs indépendants. Une cinquantaine de locuteurs marocains natifs de la région du Rif, âgés entre 25 et 50 ans, sont invités à prononcer les chiffres dix fois. Le corpus comprend dix répétitions du même chiffre par chaque locuteur. Ainsi, le corpus est constitué de 6200 fichiers audio (10 chiffres x 10 répétitions x 62 locuteurs). Notre corpus inclus la voix des locuteurs normaux, les fumeurs et pathologique.
Afin de faciliter et d'alléger la tâche d'enregistrement sur les locuteurs, chacun a été invité à répéter le chiffre Amazigh dix fois consécutivement dans un même enregistrement vocal. Par suite, nous avons séparé chaque chiffre prononcé en coupant les extrémités du signal.

Les bases d'apprentissages et de tests ont été enregistrées en utilisant un microphone simple mono, placé entre 4 et 10 centimes de la bouche des locuteurs. La voix a été enregistrée à un taux d'échantillonnage de 16 KHz sur le canal mono à l'aide de l'outil Wavesurfer pour conserver la quasi-totalité de l'information (théorème d'échantillonnage de Nyquist/Shannon). L'amplitude est alors quantifiée sur 16 bits afin d'obtenir une bonne qualité et pour diviser la position de l'élément d'un échantillon en 65 536 (2^{16}) valeurs

possibles. Dans la table 5.1 il y a plus de détails sur les caractéristiques techniques du corpus. Table 5.2 présente les dix premiers chiffres avec transcription anglaises, arabes et amazighes.

Paramètres	**Valeur**
Taux d'échantillonnage	16 kHz
Nombre de bits	16 bits
Wave format	Mono, wav
Corpus	10 Amazigh-digits
Accent	Tarifit marocain berbère

Tableau.5. 1 : Paramètres d'enregistrement utilisés pour la préparation de la base de données amdigits.

Chiffres	**English transcription**	**Arabic transcription**	**Tifinaghe transcription**	**Syllables**
0	AMYA	أميا	ⴰⵎⵢⴰ	VC-CV
1	YEN	يان	ⵢⴰⵏ	CVC
2	SIN	سين	ⵙⵉⵏ	CVC
3	KRAD	كراض	ⴽⵔⴰⴹ	VC-CVC
4	KOZ	كوز	ⴽⵓⵣ	CVC
5	SMMUS	سموس	ⵙⵎⵎⵓⵙ	CCV-VC
6	SDES	سضيس	ⵙⴷⵉⵙ	CCVC
7	SA	سا	ⵙⴰ	CV
8	TAM	تام	ⵜⴰⵎ	CVC
9	TZA	تزا	ⵜⵣⴰ	CC-CV

Tableau.5. 2 : Les chiffres Amazigh avec leurs transcriptions.

2.2. Organisation de la base de données- Structure et fichiers :

La base de données est stockée dans le répertoire racine « Amdigits » qui contient les sous répertoires et les fichiers audio relatifs aux 10 chiffres amazighs. Les sous-répertoires de la base de données sont pour indiquer le type, le sexe et l'âge des locuteurs. Le dernier sous répertoire dans l'architecture de la base de données contient des 10 fichiers .wav, et pour des raisons de confidentialité, chaque fichier a un nom comme XYN.wav, où : X présent les deux premières lettres du prénom du locuteur, Y indique les deux premières lettres du nom de famille. N est l'Identifiant de fichier wav dans la base de données (de 00 à 99). La figure 5.1

donne un aperçu de l'organisation de la base de données. Nous avons renommé les fichiers wav de telle sorte qu'il n'y a pas de confusion en ajoutant deux chiffres A et B tels que : A représente le chiffre à prononcer et B représente les répétitions.

Exemple : « mora17.wav » : c'est le 17ème enregistrement du chiffre « YEN » de la Mr Mohamed RAHOUTI.

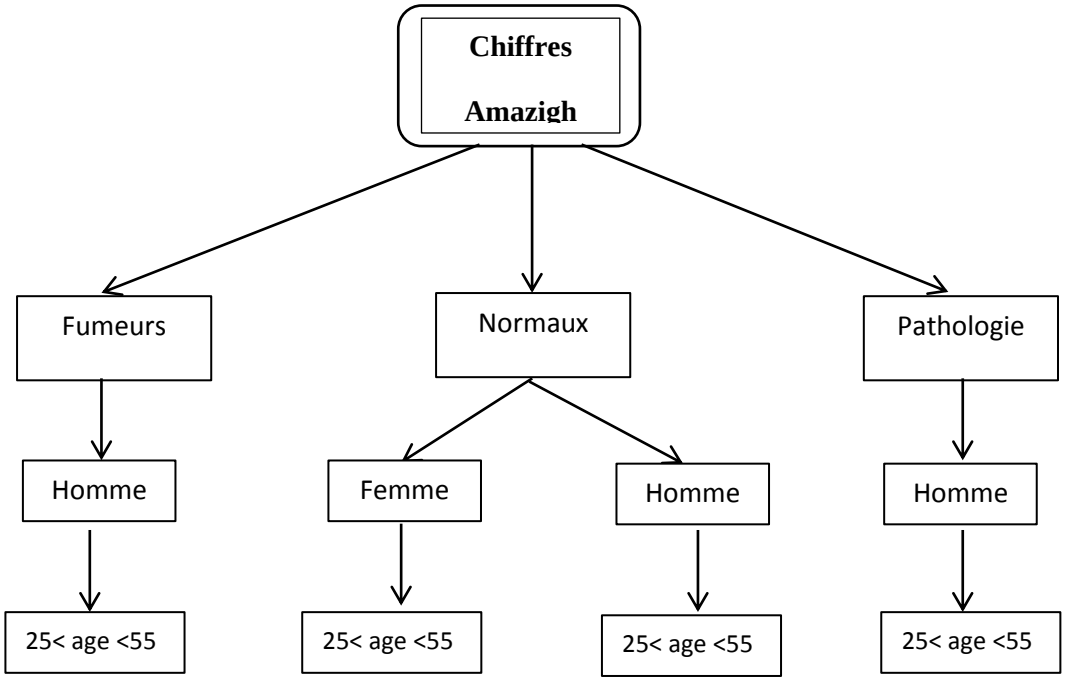


Figure.5. 1 : Description de la base des données.

3. Compilation des packages necessaries

Les packages suivants sont requis pour l'apprentissage :

- Sphinxbase
- SphinxTrain

Les packages suivants sont externes et sont également requis:

- Perl
- Python

Fondamentalement, il faut tout mettre dans le dossier racine unique. Après l'extraction des packages, on exécute et configure « make » et « make install » dans chaque dossier du package. On met le dossier de base de données dans ce dossier root. On aura donc un répertoire nommée chiffres dans notre cas avec le contenu suivant :

Chiffres\

- SphinxTrain
- SphinxTrain.tar.gz
- Sphinxbase
- Sphinxbase.tar.gz

4. Préparation de la Configuration de SphinxTrain

SphinxTrain est un formateur de modèle acoustique open source de l'Université Carnegie Mellon. Ce répertoire contient les scripts et les instructions nécessaires à la création de modèles pour le CMU Sphinx Recognizer. L'installation de Sphinxtrain nécessite en plus des codes sources Sphinxtrain, des logiciels supplémentaires comme un compilateur C++ et aussi active perl22 pour manipuler les scriptes perl qui sont fournis avec Sphinxtrain. Les différentes librairies qui composent SphinxTrain :

- ActivePerl: L'outil pour éditer des scripts pour SphinxTrain et permet de travailler dans un Unix-like environnement pour Windows plateforme.
- Microsoft Visual Studio : Pour compiler les sources en C afin de produire les exécutables.

Après la création du répertoire Amdigits, le script setup_sphinxtrain.pl va créer dans ce même répertoire un ensemble de sous répertoires:

- Bin : Il copie dans ce répertoire tous les exécutables nécessaires pour son fonctionnement.
- Bwaccumdir : Répertoire temporel utilisé lors de l'exécution de l'algorithme de Baum Welch pour accumuler les résultats.
- Etc : Répertoire utilisé pour la configuration du modèle acoustique.
- Feat : Répertoire où SphinxTrain va mettre les fichiers contenant les coefficients MFCC correspondant à la base de données d'apprentissage.
- Model_Architecture : Les définitions du modèle acoustique seront dans ce répertoire.
- Model_Parameters : Les paramètres du modèle (Matrices de transitions, poids de mixage entre les distributions gaussiennes pour chaque état, moyennes et variances des distributions gaussiennes) seront dans ce répertoire.
- Scripts_pl : Contient les scripts Perl à utiliser durant le processus d'apprentissage.

- Trees : Répertoire utilisé pour la classification des phonèmes lors de la transformation du modèle de CI à CD.

L'exécution du script make_feats.pl permet de convertir les fichiers wav en coefficients cepstraux MFCC.

5. Apprentissage du modèle acoustique

Une fois que l'on a préparé un corpus, on peut passer à l'étape de la création du modèle acoustique. Pour cela, il faut savoir quelle base sonore à utiliser (phonèmes, syllabes, mots). Par exemple, l'utilisation des phonèmes permet par la suite de pouvoir rajouter de nouveaux mots dans le dictionnaire (en spécifiant quels phonèmes interviennent pour les mots donnés) sans même avoir à recréer un nouveau modèle.

Il est à noter que même si Sphinx est pourvu d'une communauté relativement importante, les étapes de la création d'un modèle acoustique sont difficilement trouvables (les informations sont assez dispersés sur l'Internet et on trouve plusieurs bribes dans les différents forums du CMU). Par ailleurs, une certaine expertise est nécessaire avant toute création puisqu'il faut non seulement définir plusieurs paramètres qui influenceront la qualité de la reconnaissance, mais aussi apporter certaines modifications dans le code des scripts de SphinxTrain corrigeant quelques bugs.

L'apprentissage des modèles acoustiques passe par plusieurs étapes préalables. La durée du traitement étant assez variable. Pour avoir une idée, elle varie de quelques minutes, dans le cas d'un petit corpus, et à plusieurs semaines de traitements si l'on souhaite avoir un modèle de grand vocabulaire. L'opération s'exécute indépendamment du locuteur et mais quant à la base de phonèmes, ça dépend du contexte.

5.1. La préparation des fichiers d'entre :

La base de données contient des informations nécessaires pour extraire les statistiques de la parole sous la forme du modèle acoustique. SphinxTrain nécessite de lui fournir les unités sonores dont vous voulez lui apprendre leurs paramètres, et au moins l'ordre dans lequel ils apparaissent dans chaque signal de parole dans notre base de données d'apprentissage. Cette information est fournie à SphinxTrain dans un fichier appelé fichier de transcription, dans lequel la séquence de mots et des sons non-vocaux sont écrits exactement

comme ils se sont produits dans un signal de parole suivi d'une étiquette qui peut être utilisée pour associer cette séquence avec le signal de parole correspondant.

SphinxTrain cherche alors dans un dictionnaire qui fait correspondre chaque mot à une séquence d'unités de sons pour obtenir la séquence d'unités sonores associés à chaque signal. Ainsi, en plus des signaux de parole, il faut que vous donneriez également un ensemble de transcriptions pour la base de données (dans un seul fichier) et deux dictionnaires, l'un dans lequel les mots dans la langue légitimes sont cartographiés par des séquences d'unités sonores, et un autre dans lequel des sons non-vocaux sont cartographiés pour correspondre les non-paroles comme unités sonores. Nous ferons référence au premier comme le dictionnaire de la langue et le second comme le dictionnaire de remplissage « filler dictionary ». La structure du fichier de la base de données est la suivante :

- Etc

Amdigits.dic - Phonetic dictionary

Amdigits.phone - Phoneset file

Amdigits.lm.DMP - Language model

Amdigits.filler - List of fillers

Amdigits _train.fileids - List of files for training

Amdigits _train.transcription - Transcription for training

Amdigits _test.fileids - List of files for testing

Amdigits _test.transcription - Transcription for testing

- Wav

speaker_1

file_1.wav - Recording of speech utterance

speaker_2

file_2.wav

5.2. Modèle de Language :

Comme pour la majorité des systèmes de reconnaissance automatique de la parole, les modèles du langage de notre système sont des modèles n-grammes. Nous avons créé un fichier texte qui contient une liste des 10 premiers chiffres Amazigh que nous voulons utiliser pour former le modèle du langage. Les 1-grammes, 2-grammes, et 3-grammes ont une pluralité de probabilités existantes. Cependant, pour avoir une bonne estimation des

probabilités de la modélisation de langage n-grammes, on nécessite de disposer de corpus textuels de qualité et de taille suffisante. La façon la plus simple de construire un modèle du langage est d'utiliser l'outil Web en ligne LMTOOL où il suffit de cliquer sur le bouton "Choose File...", puis sélectionner le fichier qu'on a créé Amdigits.txt, ensuite cliquer sur <COMPILE KNOWLEDGE BASE>. On doit normalement voir une page avec quelques messages d'état, suivie d'une page intitulée «base d connaissances Sphinx ». Cette page va contenir des liens intitulés "Dictionnaire" et "modèle de langage". On télécharge ces fichiers et on fait une note de leurs noms (ils devraient être composés d'un numéro à 4 chiffres suivi par les extensions. DIC et. Lm). Le figure 5.2 présente un extrait des modèles[3] 1-grammes, 2 grammes et 3- grammes utilisés dans ce travail.

Pour charger rapidement des grands modèles, nous souhaitons probablement les convertir au format binaire, ce qui permettra d'économiser le temps d'initialisation de notre décodeur. Ce n'est pas nécessaire avec les petits modèles. Sphinx4 exige que vous soumettiez le modèle DMP au composant TrigramModel. Le format DMP peut être converti mutuellement. On a produit le fichier amdigits.lm.DMP avec la commande: sphinx_lm_convert -i amdigits.lm -o amdigits.dmp.

```
\data\
ngram 1=14
ngram 2=24
ngram 3=12

\1-grams:
-0.7782 </s> -0.3010
-0.7782 <s> -0.2218
-1.8573 AMYA -0.2218
-1.8573 KRAD -0.2218
-1.8573 KUZ -0.2218
-1.8573 SA -0.2218
-1.8573 SEDISS -0.2218
-1.8573 SEMUS -0.2218
-1.8573 SIN -0.2218
-1.8573 SMUS(1) -0.2218
-1.8573 TAM -0.2218
-1.8573 TZA -0.2218
-1.8573 YEN -0.2218
-1.8573 YEN(1) -0.2218

\2-grams:
-1.3802 <s> AMYA 0.0000
-1.3802 <s> KRAD 0.0000
-1.3802 <s> KUZ 0.0000
-1.3802 <s> SA 0.0000
-1.3802 <s> SEDISS 0.0000
-1.3802 <s> SEMUS 0.0000
-1.3802 <s> SIN 0.0000
-1.3802 <s> SMUS(1) 0.0000
-1.3802 <s> TAM 0.0000
-1.3802 <s> TZA 0.0000
-1.3802 <s> YEN 0.0000
-1.3802 <s> YEN(1) 0.0000
-0.3010 AMYA </s> -0.3010
-0.3010 KRAD </s> -0.3010
-0.3010 KUZ </s> -0.3010
-0.3010 SA </s> -0.3010
-0.3010 SEDISS </s> -0.3010
-0.3010 SEMUS </s> -0.3010
-0.3010 SIN </s> -0.3010
-0.3010 SMUS(1) </s> -0.3010
-0.3010 TAM </s> -0.3010
-0.3010 TZA </s> -0.3010
-0.3010 YEN </s> -0.3010
-0.3010 YEN(1) </s> -0.3010

\3-grams:
-0.3010 <s> AMYA </s>
-0.3010 <s> KRAD </s>
-0.3010 <s> KUZ </s>
-0.3010 <s> SA </s>
-0.3010 <s> SEDISS </s>
-0.3010 <s> SEMUS </s>
-0.3010 <s> SIN </s>
-0.3010 <s> SMUS(1) </s>
-0.3010 <s> TAM </s>
-0.3010 <s> TZA </s>
-0.3010 <s> YEN </s>
-0.3010 <s> YEN(1) </s>

\end\
```

Figure.5. 2 : les modèles de langage 1-gramme, 2-gramme et 3-gramme utilisé dans notre système.

[3] http://www.speech.cs.cmu.edu/tools/lmtool-new.html

5.3. Dictionnaire de Prononciation :

Le dictionnaire fournit des prononciations pour chaque mot défini dans le modèle de langue et inclut les mots que nous voulons former suivis de leur transcription. Dans le cadre de ce travail, l'ensemble des mots à reconnaitre sont les dix premiers chiffres de 0 à 9 de l'amazigh. Dans la figure 5.3 nous représentons le dictionnaire de prononciation utilisé dans notre système. Le mot TAM figure avec deux variantes TAM et TAM(2) dont les descriptions phonétiques sont respectivement T A M et T T A M.

AMYA	A M Y A
YEN	Y E N
YEN (1)	Y A N
SIN	S I N
KRAD	K R A D
KOZ	K O Z
SMMUS	S M M U S
SDES	S D E S
SDES (1)	S D E SS
SA	S A
TAM	T A M
TAM(1)	T T A M
TZA	T Z A

Figure.5. 3 : Dictionnaire de prononciation.

Le système consulte le dictionnaire pour en extraire la prononciation de chaque entrée observée. Il force le décodage acoustique-phonétique à ne reconnaître que des mots qui sont présents dans le dictionnaire de prononciation.

5.4. Dictionnaire de Phonétisation

Après la sélection du vocabulaire, la deuxième étape consiste à définir le rôle de phonèmes. Dans notre système, amdigits.phone va quant à lui contenir la liste de tous les phonèmes contenus dans le fichier amdigits.dic, en prenant soin de ne mettre qu'un phonème par ligne. Nous utilisons un jeu de 17 phonèmes définis dans le fichier de phonétisation « .phone » (voir la figure 5.4). Le jeu de phonèmes choisi est utilisé pour phonétiser manuellement toutes les entrées du vocabulaire.

I L E M K R	D O Z S A SS	U N T Y SIL

Figure.5. 4 : Les phonèmes utilisés dans notre système de reconnaissance.

5.5. Les fichiers d'entrée (filler, transcription et fileids) :

Les fichiers de types de filler contient tous les sons contenus dans les fichiers audio mais qui ne représentent pas de la parole à proprement dit (silence, bruit, inspiration, etc.). La figure 5.5 présente notre fichier amdigits.filler.

<s>	SIL
</s>	SIL
<sil>	SIL

Figure.5. 5 : Fichier amdigits.filler.

<s> silence qui marque le début d'un mot ou d'une phrase.

<sil> silence au milieu d'un mot ou d'une phrase.

<s/> silence qui marque la fin d'un mot ou d'une phrase.

N'oublions pas d'inclure à la fin du fichier un retour à la ligne. En effet, il est impératif et fait partie des erreurs presque invisibles.

Tous les fichiers audio étant dans la base de données sont accompagnés de leurs correspondantes transcriptions. L'ensemble de toutes les transcriptions doit être dans un fichier d'extension transcription. Chaque ligne de ce fichier représente un enregistrement de l'ensemble d'apprentissage. Les transcriptions doivent être classées dans le même ordre d'apparition dans le fichier de contrôle d'extension fileids et chaque ligne doit être terminée par le nom du fichier écrit entre deux parenthèses.

Dans notre travail, on va créer les fichiers amdigits.transcription (Voir figure 5.6). Amdigits.fileids (Voir figure 5.7) sont vraiment très durs à écrire manuellement. Prenons notre exemple pour bien comprendre le problème :

La base d'apprentissage et tests utilisée dans notre système est constituée de 10 chiffres prononcés par 62 personnes dont chacune est invitée à prononcer 10 fois le même chiffre: 10 (chiffres) * 62 (locuteurs) * 10 (répétitions) = 6200 fichiers audio. Chacun de ses deux

fichiers doit être composé de 6200 lignes : 6200 (.fileids) + 6200 (.transcription) = 12400 lignes, sans oublier qu'il n'est pas facile de corriger les erreurs sur les deux fichiers, et qu'une erreur peut causer l'échec de l'entrainement. Pour éviter tous ces problèmes et gagner du temps, on a pensé à créer un script qui demande à l'utilisateur d'entrer le nom de la base de données (amdigits) puis le nombre et les noms des locuteurs, après il génère automatiquement les deux fichiers amdigits.fileids et amdigits.transcription.

```
<s> YEN </s> (amha10)
<s> YEN </s> (amha11)
<s> YEN </s> (amha12)
<s> YEN </s> (amha13)
<s> YEN </s> (amha14)
<s> YEN </s> (amha15)
<s> YEN </s> (amha16)
<s> YEN </s> (amha17)
<s> YEN </s> (amha18)
<s> YEN </s> (amha19)
<s> YEN </s> (elsi10)
<s> YEN </s> (elsi11)
<s> YEN </s> (elsi12)
<s> YEN </s> (elsi13)
<s> YEN </s> (elsi14)
<s> YEN </s> (elsi15)
.
.
.
```

Figure.5. 6 : Extrait du fichier amdigits.transcription.

```
amdigits/amha/amha10
amdigits/amha/amha11
amdigits/amha/amha12
amdigits/amha/amha13
amdigits/amha/amha14
amdigits/amha/amha15
amdigits/amha/amha16
amdigits/amha/amha17
amdigits/amha/amha18
amdigits/amha/amha19
amdigits/elsi/elsi10
amdigits/elsi/elsi11
amdigits/elsi/elsi12
amdigits/elsi/elsi13
amdigits/elsi/elsi14
amdigits/elsi/elsi15
.
.
```

Figure.5. 7 : Extrait du fichier amdigits.fileids.

5.6. Configuration du format audio de la base de données

Dans le fichier de configuration, on cherche les lignes suivantes :

```
$CFG_WAVFILES_DIR = "$CFG_BASE_DIR/wav";
$CFG_WAVFILE_EXTENSION = 'sph';
$CFG_WAVFILE_TYPE = 'nist'; # one of nist, mswav, raw
```

Puis on remplace « sph » et « nist » par l'extention et le type des fichiers audio enregistrés, dans notre cas, on change « sph » par « wav » et « nist » par mswav.

5.7. Configuration du chemin vers les fichiers

Pour cela, on vérifie les lignes suivantes dans le fichier etc /sphinx_train.cfg :

```
# Variables used in main training of models
$CFG_DICTIONARY = "$CFG_LIST_DIR/$CFG_DB_NAME.dic";
$CFG_RAWPHONEFILE = "$CFG_LIST_DIR/$CFG_DB_NAME.phone";
$CFG_FILLERDICT = "$CFG_LIST_DIR/$CFG_DB_NAME.filler";
$CFG_LISTOFFILES = "$CFG_LIST_DIR/${CFG_DB_NAME}_train.fileids";
$CFG_TRANSCRIPTFILE =
"$CFG_LIST_DIR/${CFG_DB_NAME}_train.transcription"
```

Ces valeurs seraient déjà comme si nous avons configuré la structure de fichiers identiquement à la description fournie précédemment. Mais, il faut assurer que les fichiers sont vraiment appelés de cette façon. La variable $ CFG_LIST_DIR est le répertoire / etc de notre projet, et la variable $ CFG_DB_NAME est le nom de notre projet lui-même (chiffres).

5.8. Configuration des paramètres caractéristiques du son

La valeur par défaut pour les fichiers audio utilisés dans Sphinx est un taux de 16000 échantillons par seconde (16 KHz). Si c'est le cas, le fichier etc/feat.params sera généré automatiquement avec les valeurs recommandées.

5.9. L'apprentissage de système de la parole Amazigh

Les modèles acoustiques employés par le système de RAP de l'Amazigh, à base de modèles de Markov cachés, utilisent un jeu de 10 chiffres de l'Amazighe.

Après la préparation des fichiers nécessaires à la phase d'apprentissage (training), le script runall.pl automatise l'exécution d'une série de scripts, chacune représente une phase

importante de l'opération d'apprentissage ou chaque unité acoustique ou phonème est représenté par un modèle statistique décrivant la distribution des données. Chaque chiffre prononcé est transformé en une série de vecteurs de caractéristiques (feature vectors) comprenant les coefficients MFCC (Mel-Frequency Cepstral Coefficients).

Dans notre réalisation, chaque mot a été utilisé pour l'apprentissage des états HMM correspondant au modèle acoustique de la démonstration Amdigits. Le système doit savoir à quel HMM correspond chaque variable (phonème). Ces informations sont stockées dans le fichier Amdigits.dic qui permet de faire une représentation symbolique pour chaque mot. Au cours de la formation, l'algorithme Baum-Welch a été utilisé pour estimer les probabilités de transition. Le modèle acoustique est entraîné en utilisant une densité de probabilités d'état continue allant de 4 à 32 distributions de mélange gaussien avec 3 et 5 HMMs. La figure 5.8 présente le déroulement de la phase du training

```
Baum welch starting for 4 Gaussian(s), iteration: 3 (1 of 1)
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
Normalization for iteration: 3
Current Overall Likelihood Per Frame = 12.9858644286942
Convergence Ratio = 0.153864444461992
Baum welch starting for 4 Gaussian(s), iteration: 4 (1 of 1)
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100% Normalization for
iteration: 4
```

Figure.5. 8 : La phase d'apprentissage avec l'algorithme Baum Welch.

Après la phase de training, SphinxTrain nous donne des informations sur la base de données utilisée dans l'apprentissage :

- Nombre de phonèmes
- Nombre de mots
- La valeur du GMMs (Gaussian Mixture Models)
- Nombre de segments
- la durée totale de la base de données d'apprentissage etc

Le déroulement interne de l'apprentissage où il y a plusieurs nouveaux répertoires numérotés de manière séquentielle à partir de 00 * à 99 * ont été créés. Chaque répertoire contient un répertoire nommé slave*.pl ou un fichier avec l'extension «.pl », comme ci-dessous:

perl scripts_pl/000.comp_feat/slave_feat.pl

perl scripts_pl/05.vector_quantize/slave.VQ.pl

perl scripts_pl/20.ci_hmm/slave_convg.pl

perl scripts_pl/30.cd_hmm_untied/slave_convg.pl

perl scripts_pl/40.buildtrees/slave.treebuilder.pl

perl scripts_pl/45.prunetree/slave-state-tying.pl

perl scripts_pl/50.cd_hmm_tied/slave_convg.pl

Il est à signaler que dans le processus de passage par les scripts de 00 à 99 *, on aura produit plusieurs suites de modèles acoustiques, chacune d'entre elles pourraient être chargées pour la reconnaissance. Ainsi que certaines des phases ne sont nécessaires que pour la création de modèles semi-continus. Si on exécute ces phases tout en génèrent des modèles continus, les scripts seront inutiles.

Sur l'étape 000.slave_feat les fonctionnalités « Feles » sont extraites. L'interaction entre le système et les signaux acoustiques se passe par le passage d'une séquence de vecteurs caractéristiques transformés et utilisés à la place des signaux acoustiques actuels.

Le script "Make_feats.pl" va calculer, pour chaque mot d'entraînement, une séquence de 13-vecteurs de dimension (vecteurs caractéristiques) composée avec les Mel-frequency cepstral coefficients (MFCC). Les MFCCs seront stocké automatiquement dans un répertoire appelé «feat».

On note que le type des vecteurs caractéristiques que nous déterminons à partir des signaux de parole pour l'entrainement et le test, ne se limite pas à MFCC. Nous pouvons utiliser n'importe quelle technique de paramétrage raisonnable au lieu de calculer les caractéristiques autres que MFCC. Une fois les travaux lancés à partir 20.ci_hmm et exécutés jusqu'à la fin, on aura entraîné un modèle acoustique Context-Independant (CI) pour les unités de sous-mots de notre dictionnaire.

6. Implémentation du système avec Sphinx 4

6.1. Installation Sphinx-4

Sphinx-4 peut être téléchargé de l'internet soit sous forme binaire soit sous forme source code. Il a été compilé et testé sur plusieurs versions de Linux et sur Windows. L'exécution de Sphinx-4 demande des logiciels supplémentaires qui sont :

- Java 2 SDK, Standard Edition 5.0.
- Java Runtime Environement (JRE).

- Les différentes librairies qui composent Sphinx-4.
- Ant: L'outil pour faciliter la compilation en automatisant les tâches répétitives.

6.2. Configuration de Sphinx-4

Un système de reconnaissance automatique de la parole comme Sphinx 4 utilise deux éléments dépendant de la langue : le modèle acoustique et le modèle de langue. Dans notre application nous avons procédé à la modification de ces deux modèles comme il est décrit précédemment. Le sphinx 4 doit être configuré en utilisant un fichier xml.

Après l'apprentissage et la création du modèle acoustique, le modèle devrait avoir les fichiers suivants:

- mdef
- feat.params
- mixture_weights
- means
- noisedict
- transition_matrices
- variances

On utilise le même dictionnaire phonétique et le même modèle que celui utilisé pour l'apprentissage et le test initial. Ils sont situés dans le dossier amdigits/etc/ et ont des noms comme Amdigits.dic et Amdigits.lm.DMP. Il faut Faire les changements suivants dans le modèle et la configuration du dictionnaire, il suffit de pointer vers les fichiers:

<component name="trigramModel"
type="edu.cmu.sphinx.linguist.language.ngram.large.LargeTrigramModel">

<property name="unigramWeight" value="0.7"/>

<property name="maxDepth" value="3"/>

<property name="logMath" value="logMath"/>

<property name="dictionary" value="dictionary"/> <Nom de la propriété = "emplacement" value = "le nom du fichier de modèle de langage par exemple amdigits/etc /amdigits. lm.DMP "/>

</ Component>

<component name="dictionary" type="edu.cmu.sphinx.linguist.dictionary.FastDictionary">

<Nom de la propriété = "dictionaryPath" value = "le nom du fichier de dictionnaire par exemple amdigits/etc /amdigits. dic "/>

<Nom de la propriété = "fillerPath" value = "le nom du fichier de remplissage par exemple amdigits/etc /amdigits filler "/>

<property name="addSilEndingPronunciation" value="false"/>

<property name="allowMissingWords" value="false"/>

<property name="unitManager" value="unitManager"/>

</ Component>

Le modèle situé dans amdigits/model_parameters/amdigits.ci_cont. Ce dossier doit comprendre plusieurs fichiers, comme les moyennes, variances, feat.params, MDEF. Il y aura également des dossiers pour un nombre différent de gaussiennes comme 4 -8-16-32-64. Ils sont ceux intermédiaires et nous n'avons pas besoin d'eux.

Encore une fois, nous allons définir un modèle dans le fichier de configuration :

<component name="amdigit"

<type="edu.cmu.sphinx.model.acoustic.amdigit.Model">

<property name="loader" value="sphinx3Loader"/>

<property name="unitManager" value="unitManager"/>

</component>

<property name="logMath" value="logMath"/>

<property name="unitManager" value="unitManager"/>

</component>

6.3. Création de fichier JAR

Nous pouvons emballer les modèles dans un fichier JAR. L'avantage de la création d'un fichier JAR est que ce fichier peut être inclus dans le chemin de classes (classpath) et référencée dans le fichier de configuration pour qu'il puisse être utilisé dans une application Sphinx4. Une fois que nous avons fait, il ne faut pas oublier d'inclure le JAR dans le "classpath". Pour configurer le chargement depuis les fichiers jars, Sphinx4 permet aux URls (Uniform Resource Locator, littéralement « localisateur uniforme de ressource ») de contenir la ressource: < chemin d'acoustique ou langage modèle > ce qui permet aux fichiers de configuration XML de référencer facilement des modèles dans des fichiers JAR.

6.4. Extraction des caractéristiques

Sphinx-4 utilisait Front-End, qui est une classe d'emballage pour la chaîne de processeurs frontaux. Il fournit des méthodes de manipulation et de navigation des processeurs. Le front-end est modélisé comme une série de processeurs de données (voir la figure 5.9), chacun effectuant une fonction de traitement du signal spécifique. Par exemple, un processeur effectue une transformation rapide de Fourier (FFT) sur des données d'entrée, un autre processeur effectue un filtrage passe-haut. Les données d'entrée au front-end sont généralement des données audio, mais ce front-end autorise tout type d'entrée. De même, les données de sortie sont des caractéristiques typiques, mais cette interface autorise tout type de sortie. Vous pouvez configurer le serveur frontal pour accepter n'importe quel type d'entrée et renvoyer n'importe quel type de sortie. Le client doit être configuré via le fichier de propriétés du Sphinx. Les frontaux actuels génèrent des fonctionnalités qui contiennent MFCC. Pour spécifier un tel frontal (appelé «pipeline») dans Sphinx-4, nous insérons les lignes suivantes dans le fichier de configuration du Sphinx-4.

Figure.5. 9 : La configuration du front-end.

```
<component name="epFrontEnd" type="edu.cmu.sphinx.frontend.FrontEnd">
    <propertylist name="pipeline">
        <item>audioFileDataSource </item>
        <item>dataBlocker </item>
        <item>speechClassifier </item>
        <item>speechMarker </item>
        <item>nonSpeechDataFilter </item>
        <item>preemphasizer </item>
        <item>windower </item>
        <item>fft </item>
        <item>melFilterBank </item>
        <item>dct </item>
        <item>liveCMN </item>
        <item>featureExtraction </item>
    </propertylist>
</component>
```

Les filtres de traitement de signal appliqués au signal audio d'enregistrement sont mentionnés comme suit:

1- preemphasizer >> Filtre de pré-accentuation

2- windower >> Éolienne en cosinus surélevée

3- fft >> Transformée de Fourier discrète

4- melFilterBank >> Banque de filtres MelFrequency

5- Dct >> Transformée en cosinus discrète

6- liveCMN >> fonctionnalité liveCMN

7- featureExtractor >> Extracteur de fonctionnalités Deltas

Ensuite, dans le sphinx 4, de nombreux traitements de données pourraient être utilisés comme:

1- SpeechClassifier - classifie les morceaux d'audio en parole et en non parole. Il a la propriété 'seuil' pour contrôler le degré de sensibilité du point d'extrémité. Il est déterminé empiriquement que la valeur de 13 est optimale pour la plupart des environnements. Un seuil inférieur rendra le point final plus sensible, c'est-à-dire qu'il marquera plus audio que parole. Un seuil plus élevé rend le point final moins sensible, c'est-à-dire qu'il marque moins l'audio que parole.

2- SpeechMarker - marque le flux audio dans les régions vocales et non vocales en donnant des «zones amorties» autour de ces régions.

3- NonSpeechDataFilter - supprime les régions non vocales de l'audio.

4- LiveCMN: Soustrait la moyenne de toutes les entrées si loin des objets de données.

Le frontal du Sphinx-4 est connecté au reste du système via le marqueur. Nous montrerons comment le buteur obtiendra le début du score. Dans le fichier de configuration, le marqueur doit être spécifié comme présente dans la figure 5.10:

```
<component name="threadedScorer"
           type="edu.cmu.sphinx.decoder.scorer.ThreadedAcousticScorer">
    <property name="frontend" value="${frontend}"/>
</component>
```

Figure.5. 10 : La configuration du score.

7. Conclusion

Dans ce chapitre, nous avons présenté le corpus de parole utilisé dans le cadre de cette thèse et qui nous a permis de construire le modèle acoustique qui passe par plusieurs étapes. Parmi ces étapes, il y a l'étape d'apprentissage qui est la plus importante. Pour construire un modèle acoustique pour sphinx 4 on utilise Sphinxtrain qui se compose d'un ensemble de script PERL et des fichiers de configuration modifiable selon nos besoins. L'utilisation de notre modèle acoustique nécessite la création d'un fichier jar pour l'intégré dans Sphinx 4. Puis, on peut tester notre système de reconnaissance en évaluant leur taux d'erreurs automatiquement avec un script ou manuellement en testant la reconnaissance.

Chapitre 6 : Le système de reconnaissance de la parole et analyse des formants pour détecter les anomalies de la voix

1. Introduction

L’intégration de reconnaissance vocale dans les domaines de la santé est l'un des sujets sensibles au centre de nombreuses études. L’objectif de ce chapitre est l’utilisation de RAP pour détecter les anomalies liées au tabagisme ou aux maladies chez les locuteurs et l'analyse fréquentielle de la voix des fumeurs.

2. Diagnostic de la parole de Fumeurs

2.1. Vue générale

Le tabagisme a un impact négatif sur le larynx humain. Il irrite et sèche les cordes vocales, les gonflant et les empêchant de fonctionner correctement et il provoque une infiltration inflammatoire des muqueuses des cordes vocales. Cela entraîne des modifications du volume de la voix, de la cinétique des cordes vocales, de la hauteur, de la qualité du son et du geste phonatoire (Mckeating et al. 1988). Les auteurs (González J., 2004a; Tafiadis et al. 2017) montrent qu'une courte durée de tabagisme (moins d'une décennie) a un effet évident sur les paramètres de la voix tels que la fréquence fondamentale. Le nombre des cigarettes fumées par jour a eu un effet linéaire sur ces paramètres en fréquence fondamentale, pitch, amplitude, etc. (Verdonck-de Leeuw et al. 2004).

Dans cette section, nous décrivons notre expérience pour la réalisation et l'implémentation d’un système de reconnaissance automatique de la parole pour la langue Amazighe. Ce système est capable de distinguer la voix entre fumeurs et non-fumeurs en fonction de la parole produite, ainsi que l’analyse des données acoustiques en calculant les paramètres prosodiques telles que la fréquence fondamentale, les formants, Jitter et Shimmer pour bien déterminer les effets de la cigarette sur la corde vocale.

2.2. Système de Reconnaissance de la Parole pour les Fumeurs

2.2.1. Base de données de test

Afin de tester notre système automatique de reconnaissance de la parole Amazigh indépendant du locuteur pour les fumeurs, un ensemble d’enregistrements vocaux riches et équilibrés est nécessaire. Les corpus de parole utilisés lors de la phase de test sont collectés auprès de 10 locuteurs fumeurs de nationalité marocaine parlant Tarifit. Pour les enregistrements, on adopte les mêmes normes de base de données d'apprentissage. Ainsi, le

corpus de test est constitué de 1000 jetons. D'autre part, les mêmes étapes sont suivies pour l'enregistrement de dix locuteurs non-fumeurs pour présenter la comparaison entre les voix.

2.2.2. Expérience

Les expériences de ce travail ont été menées sur deux groupes de locuteurs, les non-fumeurs et les fumeurs, en utilisant un petit identificateur de parole isolé et indépendant du locuteur pour la langue Amazighe. Afin de déterminer la meilleure combinaison de paramètres pouvant être utilisée pour concevoir un système de reconnaissance vocale préférentiel capable de distinguer les voix des fumeurs des non-fumeurs, deux expériences ont été menées. Dans la première expérience, le système a été formé et testé en utilisant uniquement la voix des non-fumeurs, tandis que dans la seconde expérience, on a utilisé la voix des non-fumeurs pour la formation et la voix des fumeurs pour les tests. Le tableau 6.1 donne plus de détails sur les données de formation et de test utilisées pour chaque expérience.

		Données d'apprentissage		Données de Test	
ID d'expériences	Locuteur	Nombre de locuteur	Nombre de Jetons	Nombre de locuteur	Nombre de jetons
Expérience 1	Non-fumeurs	30	3000	10	1000
	Fumeurs	–	–	–	–
Expérience 2	Non- fumeurs	30	3000	–	–
	Fumeurs	–	–	10	1000

Tableau.6. 1 : Données de formation et de test pour chaque expérience.

2.2.3. Résultats

Pour les deux expériences, différents ensembles de paramètres d'apprentissage et de test ont été utilisés pour identifier la meilleure combinaison pouvant être utilisée pour concevoir un système RAP efficace, capable de distinguer les voix des fumeurs et des non-fumeurs. Nous avons formé et testé le système en utilisant différentes valeurs gaussiennes allant de 4 à 32, ainsi que 3 et 5 états par HMM, pour tous les dix chiffres. Le tableau 6.2 montre les résultats de la première expérience où le système a été formé et testé en utilisant la voix des non-fumeurs pour 3 et 5 HMM et le mélange gaussien allant de 4 à 32. Dans le cas de 3 HMM, le taux de reconnaissance était de 87,98, 89,09, 90,04 et 90,13% respectivement pour 4, 8, 16 et 32 GMM, alors que le résultat correspondant à 5 HMM était de 88,95, 89,25, 89,50 et 89,47%.

Le meilleur résultat est 90,13% a été trouvé avec 3 HMM en utilisant 32 GMM. Dans la deuxième expérience, les mêmes configurations que la première expérience sont répétées sauf que la base de données de tests a été remplacée par des personnes fumeuses.

Le tableau 6.3 montre le résultat de la deuxième expérience. Dans le cas de 3 HMM, les performances du système sont de 43,79, 44,38, 45,94 et 45,89% en utilisant des distributions de mélange de 4, 8, 16 et 32 gaussiennes, respectivement. Alors que dans 5 HMM, les taux corrects du système étaient de 44,89, 45,50, 45,91 et 45,83% correspondent à 4, 8, 16 et 32 GMM un à un. La performance la plus élevée est de 45,93%. Elle a été réalisée avec 3 HMM et 16 MGM. Les résultats que nous avons trouvés lors d'expériences montrent qu'il existe une grande différence en termes de reconnaissance de la parole pour les deux catégories.

Chiffres Amazigh	Taux de reconnaissance							
	3 HMM				5 HMM			
	4 GMM	8 GMM	16 GMM	32 GMM	4 GMM	8 GMM	16 GMM	32 GMM
AMYA	88.80	89.40	90.40	90.40	89.50	89.50	89.80	89.80
YEN	88.10	88.70	89.60	89.80	88.90	89.10	88.90	88.80
SIN	87.20	87.70	88.80	88.90	88.10	88.20	88.40	88.40
KRAD	91.10	92.30	93.00	93.10	92.70	92.70	93.00	93.00
KOZ	85.80	86.70	87.80	87.80	86.60	86.80	86.90	86.90
SMMUS	88.70	90.20	91.40	91.40	89.60	90.40	91.30	91.30
SDES	88.80	91.10	92.30	92.30	89.90	91.10	91.40	91.40
SA	86.60	87.50	88.40	88.70	87.30	87.50	87.90	87.90
TAM	87.40	88.90	89.60	89.80	88.60	88.70	88.80	88.70
TZA	87.30	88.40	89.10	89.10	88.30	88.50	88.60	88.50
Total	87.98%	89.09%	90.04%	90.13%	88.95%	89.25%	89.50%	89.47%

Tableau.6. 2 : Taux de reconnaissance des chiffres Amazighs non-fumeurs (%) pour différents GMMs et HMMs.

Afin de choisir les meilleurs mots à utiliser pour faire une distinction entre fumeurs et non-fumeurs, nous proposons de calculer la différence de taux de reconnaissance la plus élevée pour tous les mots utilisés dans notre étude. La figure 6.1 et la figure 6.2 montrent la différence de taux de reconnaissance calculée entre fumeur et non-fumeur dans le cas de 3 HMM et de 5 HMM, respectivement. La différence était comprise entre 42,50 et 46,20% et la différence la plus élevée a été observée avec les chiffres de Krad et la différence minimale avec Koz. Sur la base des résultats obtenus à partir des expériences, nous pouvons constater que notre système est capable de détecter la différence entre les voix des fumeurs et des non-fumeurs et de confirmer que le tabagisme affecte la voix humaine.

Chiffres Amazigh	Taux de reconnaissance							
	3 HMM				5 HMM			
	4 GMM	8 GMM	16 GMM	32 GMM	4 GMM	8 GMM	16 GMM	32 GMM
AMYA	44.10	44.40	46.10	46.00	45.10	45.90	46.10	46.10
YEN	43.60	44.10	45.50	45.40	44.90	45.20	45.60	45.40
SIN	43.10	43.80	45.20	45.20	44.50	44.80	45.20	45.10
KRAD	45.80	46.10	47.90	47.90	46.70	47.50	47.90	47.90
KOZ	42.30	43.10	44.70	44.50	43.20	44.10	44.40	44.20
SMMUS	44.60	45.20	47.10	47.10	45.60	46.40	46.90	46.90
SDES	44.80	45.50	47.30	47.30	45.90	47.10	47.50	47.50
SA	42.90	43.40	44.70	44.70	43.70	44.30	44.70	44.70
TAM	43.50	44.30	45.60	45.60	44.80	44.90	45.50	45.30
TZA	43.20	43.90	45.30	45.20	44.50	44.80	45.30	45.20
Total	43.79%	44.38%	45.94%	45.89%	44.89%	45.50%	45.91%	45.83%

Tableau.6. 3 : Taux de reconnaissance des chiffres Amazighs fumeurs (%).

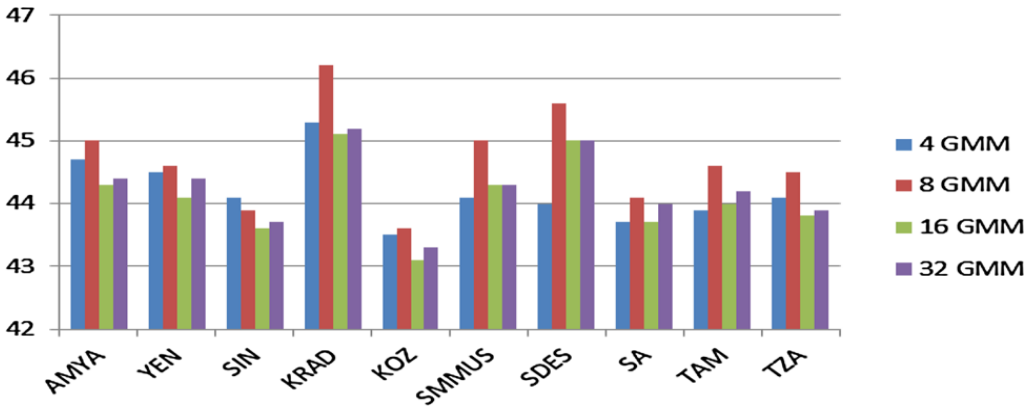


Figure.6. 1 : La différence entre les taux de fumeur et non-fumeur avec 3 HMM.

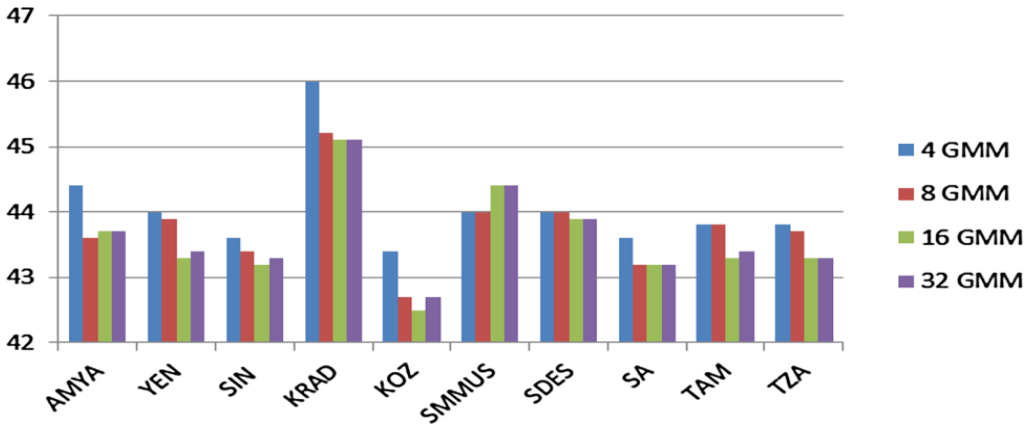


Figure.6. 2 : Différence de taux de reconnaissance entre fumeur et non-fumeur dans le cas de 5 HMM.

2.3. Analyse des paramètres vocaux du fumeur

Dans cette section, nous étudions les paramètres de la voix humaine chez les fumeurs et les adultes non-fumeurs en fonction de trois voyelles de la langue Amazighe. Le but de cette étude est d'examiner l'influence des habitudes du tabagisme sur le comportement vocal dans une perspective plus large en fonction des quatre : pitch, fréquence des formants, Shimmer et Jitter.

2.3.1. Préparation du corpus

Quarante adultes (20 fumeurs et 20 non-fumeurs) ont été sélectionnés dans la base de données vocale, collectée auprès de locuteurs Amazighs d'origine marocaine appartenant à différentes régions sans répartition géographique particulière. Les sujets ont été appariés par âge. L'âge des non-fumeurs était compris entre 26 et 50 ans avec une moyenne de 38 ans. L'âge du fumeur est compris entre 28 et 50 ans, avec un âge moyen de 39 ans. La majorité fume depuis au moins 13 ans. Notre objectif est d'analyser les 3 voyelles prononcées en étudiant les paramètres vocaux importants: pitch, formants, shimmer et jitter. Les voyelles A, I et U ont été extraites manuellement des spectrogrammes Krad, Sin et Kuz, respectivement. La procédure à été répétée dix fois successivement pour chaque chiffre. La figure 6.3 montre la voyelle A extraction du chiffre de Krad.

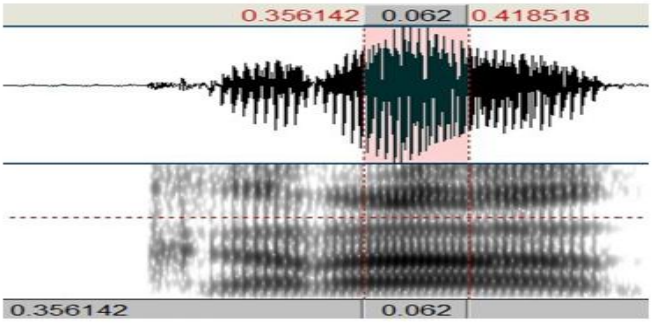


Figure.6. 3 : Sélection manuelle de la voyelle A à partir de la forme d'onde et du spectrogramme des chiffres de Krad.

2.3.2. Description des matériaux

Dans cette étude, nous utilisons un microphone et un ordinateur portable avec 4 Go de RAM et un processeur Intel Core i3 d'une vitesse de 1,2 GHz. Outre le système d'exploitation utilisé dans notre expérience, Ubuntu 14.04 LTS. Le microphone était placé à une distance de 4 à 10 cm de la bouche de la personne qui se trouvait dans une pièce calme. Pour enregistrer le fichier wav, nous utilisons le taux d'échantillonnage de 16 kHz avec une résolution de 16

bits. La figure 6.4 montre les spectrogrammes de la voyelle A pour les fumeurs (à gauche) et les non-fumeurs (à droite).

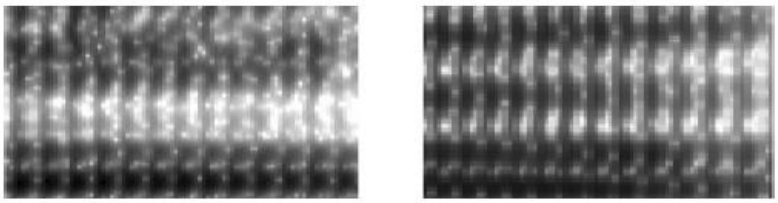


Figure.6. 4 : Spectrogramme de la voyelle A d'un fumeur à gauche et de non-fumeur à droite.

2.3.3. Résultat et discussion

Toutes nos expériences et analyses ont été réalisées à l'aide du logiciel statistique SPSS (IBM SPSS Statistics V.19) et la signification statistique a été fixée à $p < 0,05$.

Les valeurs moyennes de pitch pour les fumeurs et les non-fumeurs sont présentées dans le tableau 6.4, qui montre la différence entre les deux catégories. La différence et la valeur p pour trois voyelles A, I et U sont respectivement (25 Hz, $p = 0,008$), (19 Hz, $p = 0,009$) et (17 Hz, $p = 0,011$). Sur la base du test t, tous les résultats sont statistiquement significatifs et les valeurs moyennes de pitch pour trois voyelles (A, I, U) sont inférieures pour les fumeurs par rapport à celles des non-fumeurs. Sur la base de ces résultats, nous pouvons dire que le tabagisme élargit les cordes vocales en les rendant plus épaisses et en augmentant les sécrétions laryngées, ce qui entraîne une réduction de la fréquence tonale pour les fumeurs dans toutes les tâches de la parole.

Notre découverte révèle que les valeurs de pitch présentées par le locuteur masculin pour fumeurs et non-fumeurs sont plus élevées que celles publiées dans les travaux précédents (Guimarães et al. 2005; González J., 2004b). Cela peut s'expliquer par l'originalité des phonèmes Amazighs caractérisés par une plus grande énergie articulatoire et une durée plus longue.

	Non-fumeurs	**Fumeurs**
A	168 Hz	143 Hz
I	159 Hz	140 Hz
U	164 Hz	147 Hz

Tableau.6. 4 : Valeurs pitch pour trois voyelles (A, I, U) en Hz (fumeurs et non-fumeurs).

Les voyelles considérées dans cet ouvrage (A, I, U) sont liées à trois cavités: les cavités pharyngées, buccales et nasales.

Quatre dimensions permettent de modifier la forme ou l'accès à ces cavités: (1) ces degrés d'ouverture de la mandibule; (2) la position de la langue; (3) la position des lèvres; (4) la position du vélum (permettant ou non le passage de l'air dans les fosses). La moyenne des premiers, deuxièmes, troisièmes et quatrièmes formants des trois voyelles de la langue Amazighe pour les fumeurs et les non-fumeurs est calculée. Les résultats de mesure globaux sont résumés à la figure 6.5.

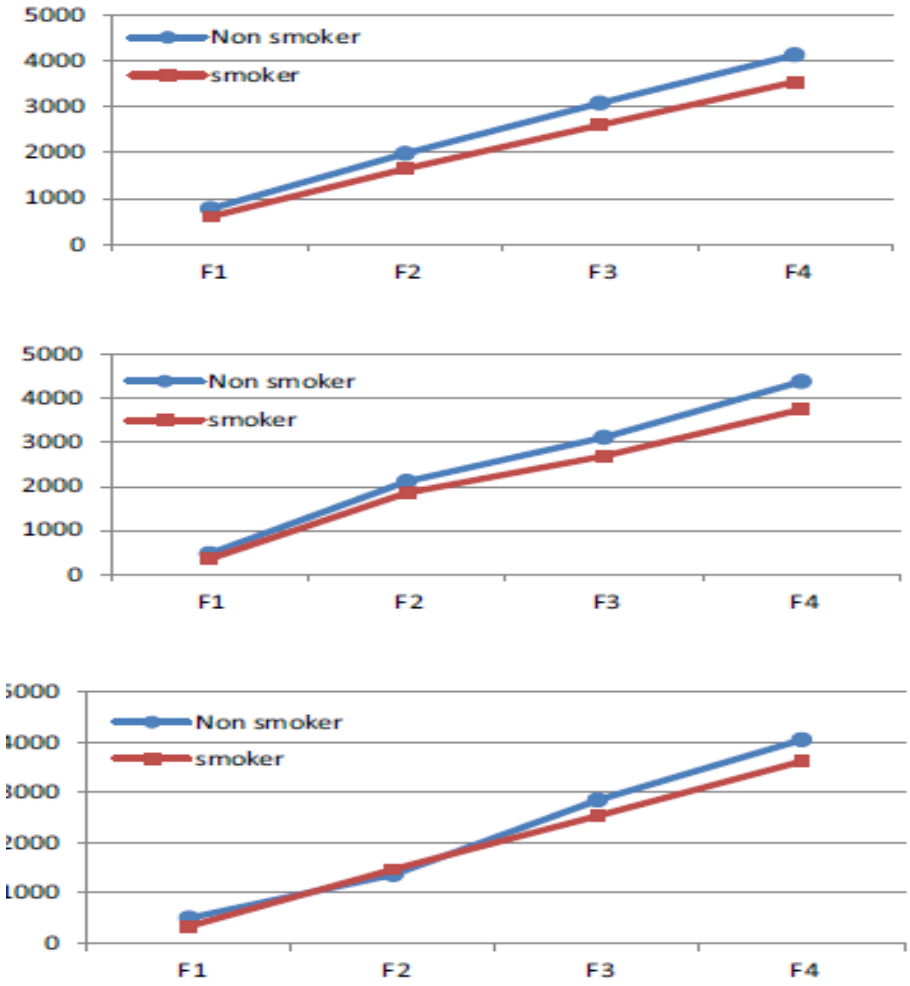


Figure.6. 5 : Fréquences de formant en Hz pour les voyelles A / I / U (fumeurs et non-fumeurs).

Les formants F1 pour les voyelles / A /, / I / et / U / étaient proches pour les deux, avec une légère augmentation pour les non-fumeurs. Le second formant, F2, les valeurs des non-fumeurs sont un peu plus élevées que celles qui fument, à l'exception de la / u / voyelle. Les formants F3 et F4 sont très élevés pour les non-fumeurs par rapport aux fumeurs. Notre expérience montre que ces formants sont les plus affectés par l'effet de fumer. Les cigarettes alternent la structure tissulaire de la cavité et des pistes vocales et contribuent à un

abaissement de la fréquence des formants. Cette tendance est similaire à l'effet de l'âge observé pour F1, F2 et F3 démontré par les autres de (Busby et al. 1995).

Le tableau 6.5 indique les moyennes de jitter locale, de jitter absolue locale, de jitter de rap, de jitter de ppq5, de scintillement local, de scintillement local en dB, de scintillement apq3 et apq5 obtenus avec Praat chez des sujets fumeurs et non-fumeurs. Des comparaisons entre fumeurs et non-fumeurs ont été effectuées à l'aide du test t (test unilatéral) et le niveau marginal de 0,05 a également été examiné.

Voyelles	**A**			**I**			**U**		
Paramètres	**Non-fumeurs (n=20)**	**Fumeurs (n=20)**	**P-value**	**Non-fumeurs (n=20)**	**Fumeurs (n=20)**	**P-value**	**Non-fumeurs (n=20)**	**Fumeurs (n=20)**	**P-value**
Jitter Relative (%)	0,509	0.956	0.0009	0.453	0.898	0.0007	0.450	1.019	0.0005
Jitter Absolut (µs)	39,710	53.035	0.0234	37.234	50..210	0.0287	34.024	52.746	0.0187
Jitter Rap (%)	0.178	0.263	0.0252	0.198	0.256	0.0398	0.190	0.227	0.0476
Jitter Ppq5 (%)	0.222	0.340	0.0082	0.241	0.337	0.0053	0.217	0.306	0.0068
Shimmer Relative (%)	4.021	6.357	0.0043	3.607	5.518	0.0032	4.574	6.161	0.0027
Shimmer DB	0.355	0.648	0.0073	0.379	0.510	0.0096	0.401	0.551	0.0091
Shimmer Apq3 (%)	1.448	1.909	0.0682	1.913	2.471	0.0561	1.878	2.350	0.0549
Shimmer Apq5 (%)	2.081	2.889	0.0185	2.261	3.237	0.0097	2.338	3.108	0.0134

Tableau.6. 5 : Pourcentage de Jitter et Shimmer (moyenne) chez les fumeurs et les non-fumeurs.

Jitter (paramètres de perturbation fondamentaux): Les résultats ont montré que les fumeurs présentent les valeurs de Jitter les plus élevées pour toutes les voyelles. L'analyse de la Jitter des fumeurs et des non-fumeurs révèle que la différence la plus importante est observée pour la voyelle / I / (0,447% pour la Jitter relative) et (13,325% pour la Jitter absolue). En utilisant également le Jitter rap et ppq5, la plus grande différence est observée pour la voyelle / A / qui a enregistré 0,085% pour le Jitter rap et 0,118% pour le Jitter ppq5. Tous les résultats sont

statistiquement significatifs pour les quatre paramètres de Jitter. Cela signifie qu'au moins un échantillon indépendant est différent des autres. Cela se voit car pour chaque paramètre, la valeur p est inférieure au niveau de signification de 0,05.

Shimmer (paramètres de perturbation d'amplitude): dans chaque tâche vocale, les valeurs de scintillement pour les non-fumeurs sont inférieures aux valeurs correspondantes pour les fumeurs. Une signification non statistique (p> 0,05) a été trouvée pour les mesures de Shimmer Apq3 en fonction de l'habitude de fumer dans trois voyelles, mais la même tendance était évidente, voyelle / A / (1,448% chez les non-fumeurs contre 1,909% chez les fumeurs, p = 0,0682), voyelle / I/(1,913% chez les non-fumeurs contre 2,471% chez les fumeurs, p = 0,0561), voyelle / U / (1,878 chez les non-fumeurs contre 2,350% chez les fumeurs, p = 0,0549).

Globalement, nos résultats acoustiques sont en accord avec la littérature. Plusieurs études ont mis en évidence une différence entre les mesures de perturbation chez les fumeurs et principalement la Jitter, bien que tous les résultats rapportés n'aient pas été significatifs (Chai et al. 2011; Vincent et al. 2012). Comme on le voit dans ce travail, les mesures de perturbation sont très variables lorsque les fumeurs testés prononcent des voyelles, ce qui peut contribuer à la signification limitée rapportée dans la littérature précédente. Le but de la présente étude était de déterminer les effets du tabagisme sur les paramètres vocaux et le changement de la voix survenant chez les fumeurs. Les résultats ont démontré l'association entre le tabagisme et les changements dans les caractéristiques vocales, où ont été révélés des changements dans les scores de pitch, de formants, de Shimmer et de Jitter moyens des fumeurs.

Dans cette partie, nous avons examiné l'effet du tabagisme sur les paramètres vocaux. Nous avons décrit le cadre expérimental qui démontre l'efficacité de l’approche proposée pour distinguer les données vocales des fumeurs et non-fumeurs et pour évaluer les performances et la pertinence de notre approche, deux différentes méthodes de recherche sont exploitées. Par conséquent de la première méthode, nous pouvons utiliser un système de reconnaissance vocale pour établir un diagnostic chez les fumeurs et confirmer qu'un locuteur est fumeur lorsque le taux de reconnaissance observé est inférieur à 50%. Ce système est base sur les HMMs et GMMs qui se focalisent sur l’amélioration de la performance du système de reconnaissance conçu. Dans la deuxième méthode, trois voyelles Amazigh (A, I, U) ont été utilisées pour analyser et comparer les pitch, fréquence des formants, Jitter et Shimmer des

fumeurs et des non-fumeurs. Les résultats obtenus montrent les différences de mesure des paramètres des voyelles et démontrent l'influence du tabac sur la voix.

3. L'évaluation de la pathologie vocale basée sur la technologie de RAP

Ces dernières années, la recherche sur les systèmes automatiques d'évaluation des troubles de la voix a fait l'objet d'une attention appréciable en raison de son objectivité et de son caractère non invasif. Le travail présenté dans cette partie vise à construire un système de reconnaissance automatique de la parole basé sur Sphinx4 permettant de détecter les personnes ayant des troubles de la voix. Ce projet de recherche utilise la langue Amazighe afin de différencier les voix normales des voix pathologiques. Les performances de notre système ont été mesurées à l'aide des combinaisons de HMM à 5 états avec 8 distributions de mélanges gaussiennes. Les résultats que nous avons obtenus sont très satisfaisants : une grande différence entre la précision des locuteurs normaux et pathologiques.

3.1. Parole pathologique

3.1.1. Généralité

La reconnaissance de la parole laryngée (pathologique) et son évaluation est un sujet sensible au centre de nombreuses études dans des domaines multidisciplinaires (Dibazar et al. 2006; Lachhab O., 2017). Le discours pathologique fait référence à la parole produite par des locuteurs qui souffrent du dysfonctionnement (changement de la voix laryngée) du son et de la parole. Les défauts phonologiques peuvent être évalués, soit par des jugements cognitifs, soit par une analyse objective.

L'analyse par des jugements de perception est la méthode la plus fondamentale utilisée en pratique clinique. Il consiste à caractériser la qualité vocale par une simple écoute attentive. Cependant, cette technique souffre de plusieurs inconvénients. Premièrement, ce jugement doit être porté par un jury d'experts pour accroître sa fiabilité. Deuxièmement, cette analyse est très coûteuse en temps et en ressources humaines et ne peut être planifiée régulièrement.

Actuellement, l'analyse objective (Wuyts et al. 2000; Yu et al. 2001) est de plus en plus exploité. Il est basé sur une analyse de mesures acoustiques, aérodynamiques et physiologiques. Ces mesures peuvent être extraites directement du signal vocal à l'aide d'un système informatique. Cette approche objective fournit des résultats acceptables mais encore insuffisants pour la reconnaissance et l'évaluation spontanées de la parole œsophagienne.

3.1.2. Les troubles de la voix

La parole pathologique provient de certaines perturbations phonologiques qui conduisent à une modification des paramètres acoustiques (changement objectif) ou / et du son (changement personnel) de la parole. Ce défaut de voix peut être temporaire ou permanent. En général, il existe trois grandes catégories de pathologies :

- Pathologie d'origine fonctionnelle : mauvaise utilisation des organes de la parole (conduit vocal), cause souvent associée à l'âge du patient (le locuteur). Parfois, nous constatons un changement dans le son de la cause psychologique, comme la dépression.
- Maladies d'origine organique : laryngite aiguë, lésions des cordes vocales, abcès, etc. Les principales causes de ces maladies sont l'effet phonémique et les infections virales ou bactériennes du larynx.
- Maladies cancéreuses : l'ablation partielle ou totale du larynx est une chirurgie causée par un cancer. Les principales causes sont la consommation d'alcool et le tabagisme.

Dans cette partie, nous étudierons les dysfonctionnements de la voix dus aux maladies cancérigènes.

3.1.3. Le cancer du larynx

Le larynx (figure 6.6) contient plusieurs organes. Il est situé au carrefour entre les voies respiratoires, le tube digestif, entre le pharynx et la trachée, et devant l'œsophage. Les cordes vocales sont des lèvres symétriques (fibreuses) placées à travers le larynx. L'air expiratoire passe des poumons lors de la connexion vocale en faisant vibrer la membrane muqueuse des cordes vocales proches, ce qui permet de produire un son vocal de haute qualité grâce à l'amplification du tractus vocal.

Le cancer du larynx est caractérisé par la présence d'une tumeur sous la forme d'une ulcération anormale de l'une des cordes vocales. Le traitement consiste ensuite en une radiothérapie et une chimiothérapie, ainsi que le retrait de la corde vocale affectée (cordectomie ombilicale). Cependant, en cas d'infection répétée ou lorsque le cancer auditif est très sévère et touchant presque tout l'organe, une ablation complète du larynx (laryngectomie totale) est nécessaire.

Le cancer de la gorge est une maladie tumorale relativement courante chez les hommes. Selon les dernières statistiques publiées par l'Institut français de surveillance de la santé publique.

Elle représente environ 25% des atteintes de cancer du système digestif supérieur et 15% de tous les cancers diagnostiqués en France. Au Maroc, selon le service d'épidémiologie de l'Institut national du cancer de Rabat entre 1985 et 2007, le cancer de la gorge représente 30,8% des cancers respiratoires et 9,2% de l'ensemble des cancers enregistrés. Les tranches d'âge les plus touchées pour les hommes sont celles de 50 à 54 ans, suivies de 55 à 59 ans. Cette condition affecte principalement les hommes avec 94% contre seulement 6% des femmes. Le tabagisme actif en est la principale cause, aggravée par la consommation combinée d'alcool et l'inhalation de cancérigènes tels que l'amiante.

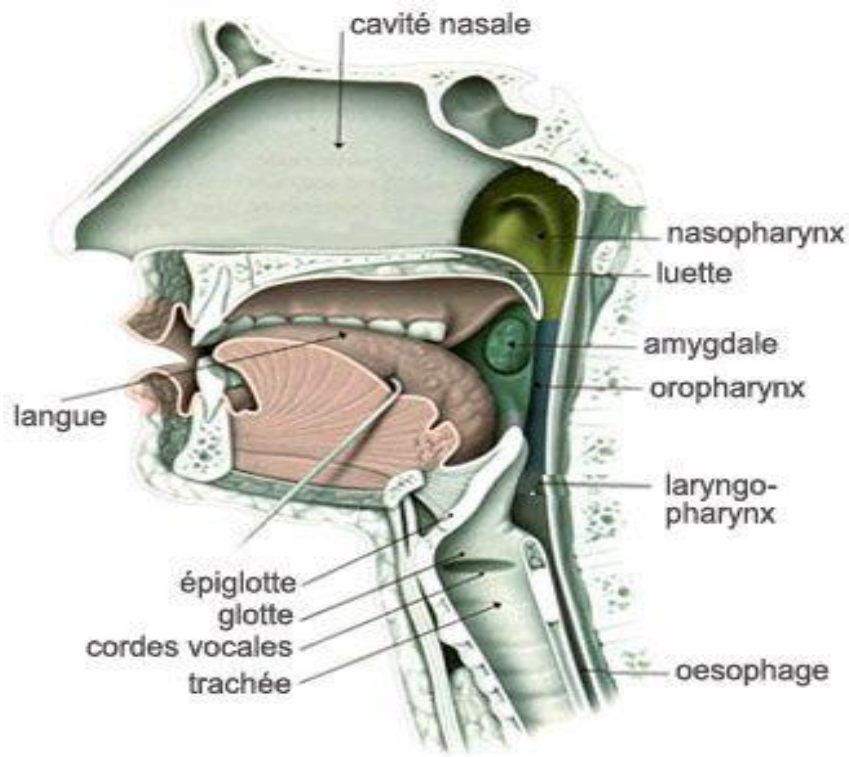


Figure.6. 6 : La schématisation de l'appareil vocal[4] .

3.2. Préparation de la base de données vocale

Cette phase consiste à enregistrer le signal de parole. Dans un premier temps, nous avons utilisé un microphone de bureau dans un environnement propre et un outil de surfeur d'ondes tout en maintenant une distance d'environ 5 à 10 cm entre la bouche du haut-parleur et le microphone. Le taux d'échantillonnage utilisé pour l'enregistrement est de 16 kHz, avec une résolution de 16 bits pour plus de détails sur le corpus. La base de données utilisée dans notre système comprend 24 locuteurs amazighs âgés entre 26 et 50 ans. Cette base de données

[4] Illustration extraite de: http://lecerveau.mcgill.ca

est divisée en deux catégories ; la première comprend 22 personnes normales et la seconde contient 2 locuteurs ayant une corde vocale troubles (l'enregistrement vocal avec des patients atteints de troubles a été approuvé par le comité d'éthique du CHU d'Oujda). La méthode suivie pour enregistrer la voix consiste à demander à ces locuteurs de prononcer les dix premiers chiffres amazighs (dix fois pour chaque chiffre) de manière séquentielle. Les enregistrements audio de chaque haut-parleur ont été sauvegardés dans dix fichiers «.wav», chaque fichier «.wav» comprend dix répétitions d'un même nombre. Ensuite, nous avons divisé chaque fichier en dix fichiers wav. Ainsi, ce corpus est composé de 2400 jetons.

3.3. Fonctionnement du système

La figure 6.7 présente notre architecture système. Dans le premier, l'utilisateur prononce les chiffres amazighs comme des mots isolés. Le système a coupé la voix capturée en différentes parties. Ensuite, il génère des vecteurs caractéristiques représentant les caractéristiques du signal de parole. Ensuite, le décodeur traite les informations reçues, les analyse et les compare avec la base de connaissances pour donner un résultat à l'application. Enfin, l'application de décision calcule l'exactitude du mot et donne la décision. Le mot exactitude est devenu le système de mesure standard pour évaluer les performances des systèmes de reconnaissance vocale. N est le nombre total des mots. R est le nombre des mots reconnus pour chaque chiffre. La précision des chiffres est R / N. La précision est généralement mesurée par un pourcentage.

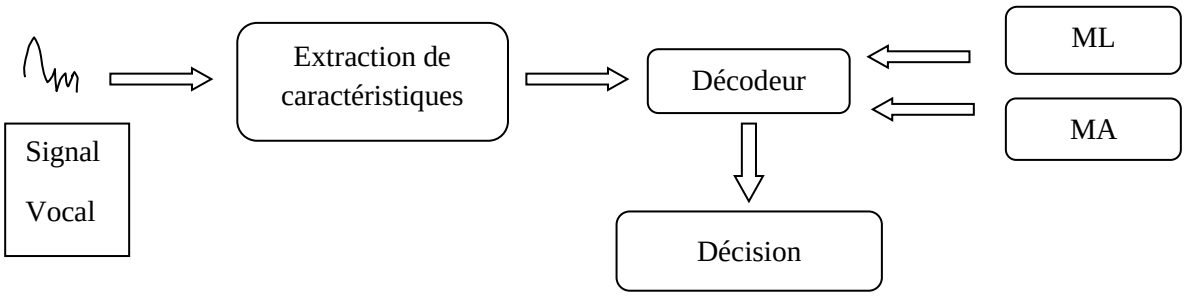


Figure.6. 7 : Processus du système de détection des troubles.

3.4. Résultats Expérimentaux

Afin d'évaluer les performances de nos systèmes, nous avons réalisé deux expériences principales axées sur une tâche de phonème connecté constituant dix premiers chiffres amazighs isolés. Chaque phonème a été modélisé par trois et cinq HMM. Le nombre de

mélanges dans le modèle de chaque état était de 16. Pour l'extraction des caractéristiques, des coefficients MFCC à 13 dimensions ont été utilisés. La première expérience concerne la formation et le test du système avec les enceintes normales (20 tests de formation 2). La deuxième expérience consiste à tester les performances du système avec les locuteurs qui ont des voix pathologiques (entraînement en utilisant la voix de 20 personnes normales et test par 2 voix pathologiques). Les 70% des fichiers wav en phase d'apprentissage et les 30% restants en phase de test ne sont pas respectés pour garantir l'aspect indépendant du locuteur de notre module RAP. Le tableau 6.6 montre une comparaison des performances du système pour les dix premiers chiffres amazighs à l'aide de 16 GMM et HMM.

Chiffres Amazigh	***Taux de Reconnaissance (%)***			
	Voix normal		***Voix pathologique***	
	16 GMM		***16 GMM***	
	3 HMM	***5 HMM***	***3 HMM***	***5 HMM***
AMYA	85,00	85,00	30,00	30,00
YEN	80,00	80,00	25,50	25,00
SIN	80,00	80,00	25,00	25,00
KRAD	90,80	90,00	35,00	35,00
KOZ	75,00	80,00	20,00	20,00
SMMUS	90,00	90,00	30,00	35,00
SDES	90,00	90,00	30,00	30,00
SA	80,00	80,00	20,00	20,00
TAM	80,00	85,00	25,00	30,00
TZA	80,00	80,00	25,00	25,00
Moyenne	83,08%	84,00%	26,55%	27,50%

Tableau.6. 6 : Précision de reconnaissance (%) des voix normales et pathologiques.

Comme présenté dans le tableau 6.6, la précision obtenue par les locuteurs normaux était très élevée par rapport aux locuteurs pathologiques où une perte significative de précision sur la reconnaissance vocale pour les échantillons de troubles de la voix est observée. Les performances globales du système sont de 83,08% et 84,00%. Ces derniers ont été trouvés

pour l'utilisation de 3 et 5 HMM respectivement pour des normales. Alors que le système corrige les taux. Dans le cas pathologique, ils étaient respectivement de 26,55% et 27,50% pour les 3 et 5 HMM. Cette différence entre le taux de reconnaissance des locuteurs normaux et des locuteurs souffrant de troubles de la voix est due au signal de parole d'un sujet présentant des troubles contenant une amplitude inférieure au signal de parole d'un sujet normal (Ali et al. 2016), en plus de l'altération des vibrations muqueuses. Sur la base des résultats obtenus à partir des expériences, nous pouvons voir que notre système est capable de faire la distinction entre les voix normales et pathologiques. Ces résultats sont en bon accord avec l'étude (Muhammad et al. 2011) qui a montré l'analyse des troubles de la voix. C'est un résultat très satisfaisant.

Dans cette partie, nous proposons une méthode permettant de montrer la différence entre les locuteurs normaux et pathologiques en utilisant la tâche de reconnaissance automatique de la parole. Ce système implémenté a été développé avec open source CMU Sphinx 4 dépend des dix premiers chiffres Amazighs. À notre connaissance, il s'agit de la première étude qui tente d'évaluer la précision de RAP dans le discours amazigh pour les personnes ayant des voix pathologiques. Dans nos travaux futurs, nous enregistrerons un plus grand nombre de locuteurs pathologiques et étudierons les performances du système proposé dans un discours continu pour analyser différents types de troubles des voies vocales.

4. Conclusion

Nous avons construit deux systèmes vocaux indépendants du locuteur où chacun d'eux comporte ses propres caractéristiques et méthodes de modélisation, d'apprentissage et de test. Nos systèmes sont formés par la voix des locuteurs sains et test pour la première fois avec la voix des fumeurs et pour la deuxième fois par la parole produite par des locuteurs atteints de dysfonctionnement de la voix et de la parole à cause du cancer du larynx. En outre, on a analysé la parole du fumeur par l'utilisation des paramètres vocale.

Le chapitre suivant présente une étude sur l'effet du bruit sur les performances d'un système de la reconnaissance automatique de la parole basé sur les modèles de Markov cachés (MMC) et les coefficients cepstraux de fréquence Mel (MFCC).

Chapitre 7 : Effet du bruit sur les chiffres Amazighs dans le système RAP

1. Introduction

La reconnaissance automatique de la parole (RAP) dans le discours amazigh, en particulier le discours accentué de Tarifit marocain, est un domaine moins étudié. Ce chapitre se concentre sur l'analyse et l'évaluation des dix premiers chiffres amazighs dans des conditions bruyantes d'un point de vue RAP basé sur le rapport signal sur bruit (SNR). Nos expériences de test ont été réalisées sous deux types de bruits et répétées avec un bruit environnemental supplémentaire avec différents rapports SNR pour chaque type allant de 5 dB à 45 dB. Plusieurs méthodes sont utilisées pour développer un système de reconnaissance vocale amazighe des mots isolés indépendants du locuteur comme le modèle de Markov caché (MMC), les modèles de mélange gaussien (GMM) et les coefficients cepstraux de fréquence de Mel (MFCC). Les résultats expérimentaux dans des conditions bruyantes montrent qu'une dégradation des performances a été observée pour tous les chiffres avec des degrés différents et que les taux dans un environnement bruyant de la voiture sont moins diminués que les conditions du broyeur.

2. RAP dans les environnements bruyants

Les performances des systèmes de reconnaissance vocale utilisés dans les environnements bruyants sont généralement en baisse. Ce phénomène est observé dans de nombreuses études (Benesty et al. 2007). Différentes techniques ont été étudiées pour développer la robustesse au bruit. Parmi eux se trouve l'utilisation des algorithmes d'amélioration de la parole. Dans ce processus, et avant la soumission du signal de parole au système RAP, il est soumis à un procédé de débruitage, par ex. par filtrage de Wiener ou soustraction spectrale, ou en utilisant une méthode différente pour développer de nouveaux modèles auditifs moins sensibles au bruit.

D'autres chercheurs suggèrent un traitement avancé des caractéristiques comme des techniques de normalisation cepstrale (par exemple, la normalisation moyenne cepstrale - CMN, normalisation moyenne cepstrale variable - VCMN), ou d'autres techniques qui tentent d'évaluer les paramètres cepstraux de la parole non déformée, étant donné les paramètres cepstraux de la parole bruyante. Ceci est parfois intégré à un apprentissage à conditions multiples, c'est-à-dire à des modèles acoustiques d'entraînement avec une parole déformée avec plusieurs types de bruit et des rapports signal sur bruit (SNR) [Hansen et al. 2001- Deng et al. 2001]. L'utilisation d'une classification basée sur une représentation clairsemée permet d'améliorer la robustesse même s'elle nécessite une grande puissance de traitement. Pour

certains types de bruit, l'utilisation des propriétés perceptives s'est avérée améliorer la précision du système RAP (Haque et al. 2009). Dans les méthodes traditionnelles de reconnaissance automatique de la parole résistante au bruit, les modèles acoustiques sont généralement entraînés à l'aide de la parole claire ou à l'aide de données à conditions multiples qui sont traitées par le même algorithme d'amélioration des fonctionnalités qui devrait être utilisé dans le décodage.

3. Préparation du corpus Amazigh

La base de données amazighe digits a été créée dans le cadre de ce travail et contient un corpus de discours et leur transcription de 40 locuteurs amazighs marocains. Le corpus se compose de 10 premiers chiffres prononcés en amazigh (0–9). Les fichiers audio ont été générés par des locuteurs prononçant les chiffres dans l'ordre numérique. Ainsi, la tâche d'étiquetage des signaux vocaux après segmentation est facile. La fréquence d'échantillonnage de l'enregistrement est de 16 kHz, avec une résolution de 16 bits, une fenêtre de Hamming de 25,6 ms avec des trames consécutives se chevauchant de 10 ms et des coefficients cépstraux Mel-Frequency (MFCC). Le tableau 7.1 présente plus de détails techniques sur le corpus vocal. Pendant les sessions d'enregistrement, les orateurs ont été invités à prononcer les 10 chiffres graduellement. Les enregistrements audio d'un seul haut-parleur ont été sauvegardés dans un seul fichier «.wav» et parfois jusqu'à quatre fichiers «.wav» selon le nombre des sessions passées par l'orateur pour terminer l'enregistrement. La sauvegarde de chaque enregistrement une fois prononcé prend du temps. Par conséquent, le corpus se compose de 10 répétitions de chaque chiffre produit par chaque locuteur. En fonction de cela, le corpus se compose de 4000 jetons. Pendant la session d'enregistrement, la forme d'onde de chaque énoncé a été visualisée pour s'assurer que le mot entier était inclus dans le signal enregistré.

Par conséquent, il était nécessaire de segmenter manuellement ces gros fichiers «.wav» en petites séquences, chacune ayant un enregistrement unique d'un seul mot et la classification manuelle de ces fichiers «.wav» dans les répertoires correspondants a été également effectuée. Les énoncés mal prononcés ont été ignorés et seuls les énoncés corrects sont conservés dans la base de données. Le logiciel utilisé pour la voix avec des haut-parleurs est WaveSurfer. Nos bases de données sur les bruits ont été enregistrées dans l'environnement progressif de faible jusqu’à un bruit élevé. De cette façon plusieurs ensembles d'enregistrements (types de voitures et de Broyeur) ont été obtenus, avec un SNR allant de +5 dB à + 45 dB, en fonction de la distance d'enregistrement entre la source bruyante et l'appareil d'enregistrement. Les

données de bruit d'origine ont été échantillonnées à 8 kHz et stockées sous forme d'entiers de 16 bits. Elles ont été préparées pour être utilisées dans cette base de données en augmentant l'échantillonnage à 16 kHz. L'augmentation a été appliquée pour permettre l'ajout du bruit sans modifier les autres paramètres.

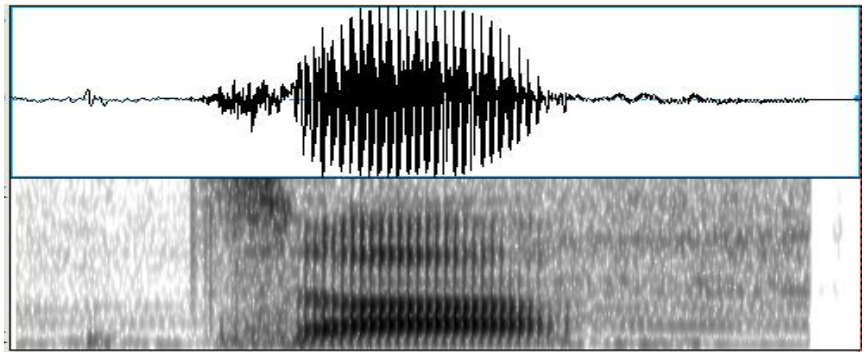
Paramètre	**Valeur**
Taux d'échantillonnage	16 kHz
Nombre de bits	16 bits
Nombre des canaux	1, Mono
Corpus	Amazigh_10 digits
Hamming	25.6 ms
Encadrement	10 ms
Types de bruit	Voiture – Broyeur

Tableau.7. 1 : Paramètres de système.

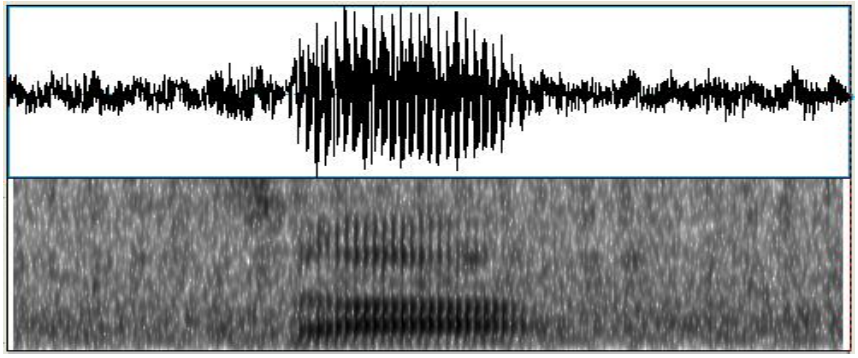
4. Test de reconnaissance en condition bruyante

Notre système RAP a d'abord été testé avec l'environnement bruyant de la voiture, puis il a été testé avec un environnement bruyant de broyeur. D'autre part, la base de données de formation pour les expériences indépendantes du locuteur comprend des mots isolés (chiffres) prononcés par 30 locuteurs (70% de la base de données) dans un environnement propre. Le reste des locuteurs (30% de la base de données) a été utilisé dans le test. Deux ensembles de bases de données de test ont été créés avec la parole dans le bruit à partir de 5 dB et avec une augmentation de 10 dB à chaque configuration jusqu'à atteindre 45 dB pour chaque type de bruit. La figure 3 présente les spectrogrammes de chiffres de Kuz utilisés sous bruit. On observe que le bruit de la voiture est un bruit stationnaire de bande inférieure, tandis que le broyeur a une bande passante beaucoup plus large et il contient des sons aigus soudains.

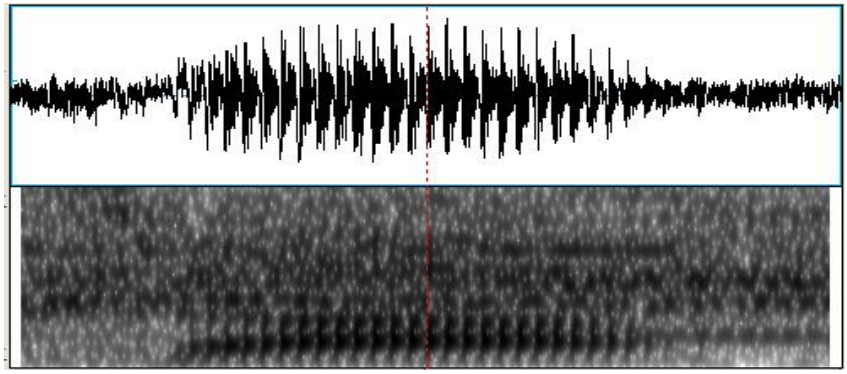
Un signal vocal qui est donné dans un environnement bruyant est moins intelligible que le même signal donné dans un environnement propre en raison du fait que la distance spectrale entre le signal vocal et le signal de bruit est réduite.



(a)



(b)



(c)

Figure.7. 1 : (a) Spectrogramme du chiffre Kuz dans un environnement normal. (b) Spectrogramme du chiffre kuz à 25 SNR bruit sous la voiture. (c) Spectrogramme du chiffre kuz à 25 SNR sous bruit de broyeur.

5. Performances du système vocal Amazigh

Les résultats fournis dans cet article dépendent principalement des résultats du système de reconnaissance des chiffres amazigh conçu dans des conditions bruyantes. Tous les tests d'environnements bruyants ont été réalisés sur la base d'un modèle acoustique indépendant du

locuteur. Pour l'environnement bruyant de la voiture, les performances globales sont respectivement de 72,92%, 56,50%, 31,75%, 4,83 et 0,17% pour un SNR = 5dB, 15dB, 25dB, 30dB et 35dB. Pour les conditions du broyeur, les performances globales sont de 70,08%, 51,42%, 23,33%, 3,58 et 0,00% pour les mêmes valeurs SNR respectivement. Le tableau 7.2 donne les taux de reconnaissance vocale des chiffres amazighs en utilisant les mêmes données de test dans des conditions de bruit de voiture et de broyeur pour divers niveaux de SNR.

Bruit	**5 dB**	**15 dB**	**25 dB**	**35 dB**	**45 dB**
Voiture	72,92	56,50	31,75	4,83	0,17
Broyeur	70,08	51,42	23,33	3,58	0,00

Tableau.7. 2 : Taux de reconnaissance globaux.

A titre de comparaison des résultats entre les deux types d'utilisation de bruit, une différence dans les taux de reconnaissance a été observée pour la même valeur SNR. Par exemple, la différence des performances globales est de 2,84%, 5,08%, 8,42%, 1,25 et 0,17 pour SNR = 5 dB, 15 dB, 25 dB, 35 dB et 45 dB, respectivement. Les niveaux de performance de la plupart des systèmes de reconnaissance vocale actuels D'après la figure 7.2, le chiffre de Krad a une précision de 79,17% à 5 dB et il s'est dégradé à 70,00%, 41,67%, 11,67% et 1,67% dans l'environnement bruyant à SNR 15 dB, 25 dB, 35 dB et 45 dB, respectivement. La confusion de Krad avec Kuz diminue progressivement avec l'augmentation du niveau de bruit. La situation similaire s'est également produite avec tous les chiffres utilisés. En outre, des taux inférieurs ont été observés pour Sin, Smmus, Sdes et Sa où ces chiffres ont des précisions plus faibles que les autres avec tous les SNR utilisés où l'influence bruyante a été clairement observée avec 25 dB. La figure 7.3 montre les taux de reconnaissance du système pour la parole bruyante du broyeur avec certaines valeurs SNR utilisées dans la première expérience. La grande précision obtenue grâce au chiffre Krad et les chiffres Amya, Kuz, Tam et Tza maintiennent les taux de reconnaissance à plus de 70% tandis que les autres chiffres atteignent une précision inférieure à 70% jusqu'à SNR 5 dB. Pour un SNR de 15 dB et plus, la reconnaissance diminue à nouveau pour tous les chiffres. Les chiffres étudiés ont une précision inférieure à 25 dB et une très faible précision a été obtenue à 35 dB. De plus, nous avons constaté que les chiffres qui contiennent l’alphabet S ne sont pas reconnus à 35 dB et que ces chiffres possèdent une très grande dissemblance par rapport à tous les autres chiffres

prononcés. Car le chiffre le plus résistant est Krad, en raison de ses consonnes fortes incluses et du nombre de syllabes.

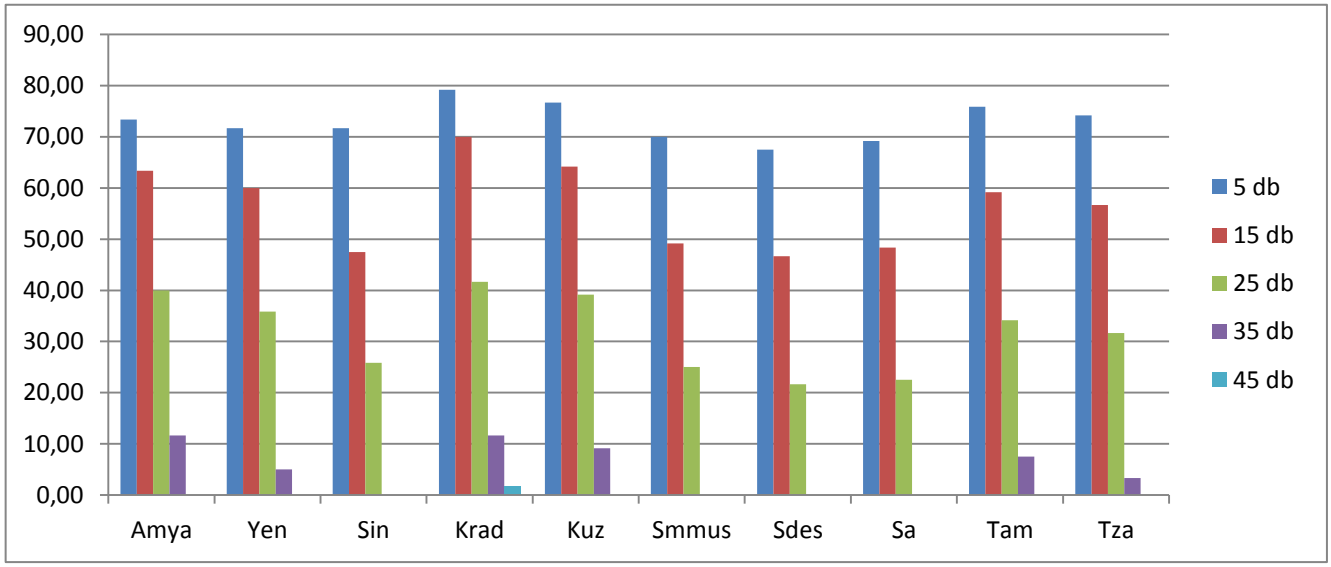


Figure.7. 2 : Taux de reconnaissance des chiffres dans des conditions de bruit de voiture.

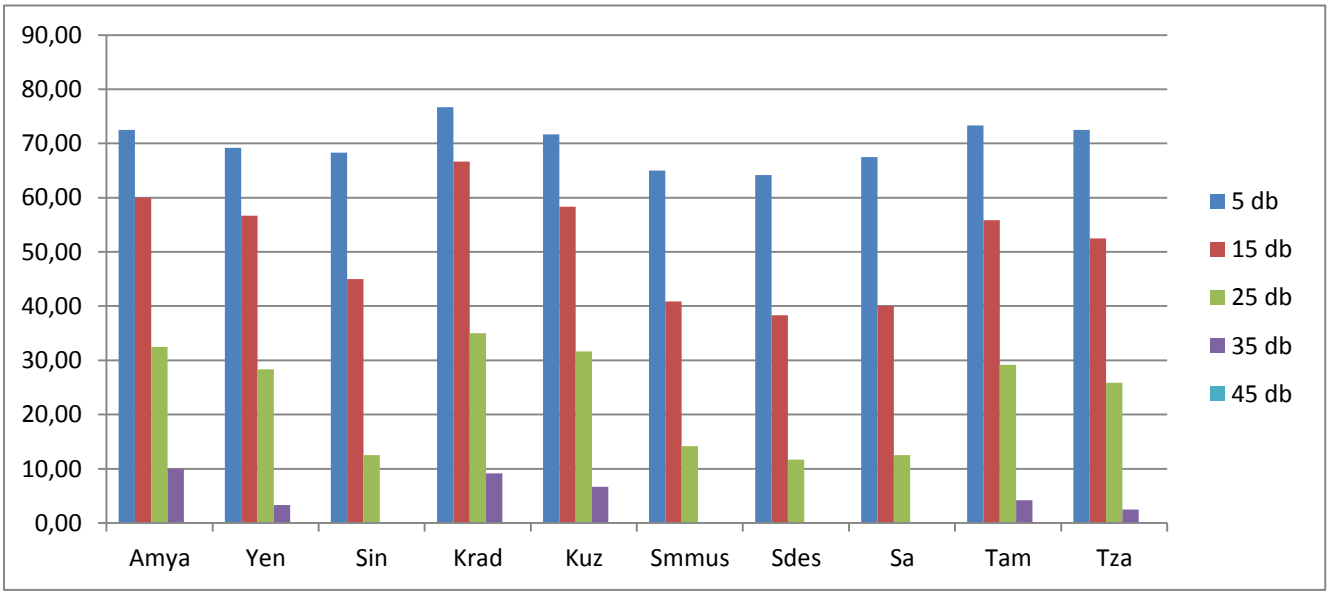


Figure.7. 3 : Taux de reconnaissance des chiffres dans des conditions bruyantes du broyeur.

6. Conclusion

Ce chapitre décrit nos expériences pour le système de reconnaissance vocale des chiffres Amazigh dans des conditions bruyantes. La conception et la mise en œuvre de modèles acoustiques et linguistiques basés sur les outils CMU Sphinx sont décrites. Des expériences de reconnaissance vocale dans des conditions bruyantes ont montré que la dégradation des performances était observée alors que la reconnaissance ne dépasse 5 dB et que le taux de reconnaissance était à peine affecté si le SNR dépassait 25 dB pour les deux types bruyants. Cependant, une dégradation majeure de la précision a été observée si le signal vocal était déformé par du bruit et que le SNR dépassait 35 dB. Dans cette enquête, nous avons constaté que les chiffres qui incluent l'alphabet S sont plus affectés que les autres chiffres pour différentes valeurs SNR.

Conclusion Générale

Le travail présent dans cette thèse s'intégré dans le cadre de la reconnaissance automatique de la parole, dans lequel nous avons fixé quatre objectifs.

Le premier objectif a visé la réalisation d'un système de reconnaissance vocale de dix premiers chiffres amazighe en utilisant le formalisme Markovien. Par ailleurs, nous avons adopté une procédure expérimentale systématique pour générer différent modelé acoustiques avec différents paramètres. Elle nous a permis de valider un certain nombre de choix sur la représentation du signal et les paramètres acoustiques les mieux adaptés pour la RAP de la langue Amazighe. Nous avons obtenu des performances de 90%.

Notre deuxième objectif consiste dans l'utilisation de la technologie de la reconnaissance automatique de la parole et l'analyse des formants pour caractériser la voix des personnes fumeurs. Nos résultats expérimentaux permettent de distinguer nettement qu'une personne est un fumeur.

Dans le troisième objectif, nous avons proposé une nouvelle approche pour détecter les locuteurs qui ont des troubles de la voix basée sur un système de reconnaissance automatique de la parole. A notre meilleure connaissance, c'est la première étude qui intègre la langue Amazighe dans une telle étude. Au cours de cette réalisation, en plus des difficultés techniques, nous avons rencontré plusieurs problèmes liés à la préparation du corpus, l'étiquetage manuel d'une grande quantité des données. Malgré ceci, nous espérons qu'elle deviendra une référence pour les chercheurs dans le domaine de RAP et son utilisation.

Le dernier objectif de cette thèse, c'est étudier la composition phonémique des chiffres amazighs les plus résistants et les plus affectés aux différents types de bruit dans différents niveaux. Les résultats expérimentaux montrent qu'une dégradation des performances pour tous les chiffres augmente le rapport signal bruits SNR et en particulier pour ceux se contenant le phonème « S ». Cette trouvaille nous permet de sélectionner les mots avec les listes des phonèmes adéquats pour certaines applications industrielles tels que les commandes vocales pour l'industrie automobile et mobile.

Le travail présenté dans cette thèse est une approche pour répondre au problème que nous avons posé. Les solutions proposées sont certes incomplètes mais offrent un aperçu sur de nombreuses perspectives.

Tout d'abord, nous projetons élargir la base de données dédiée à la reconnaissance automatique de la parole de l'Amazigh et d'autres dialectes marocains qui répondent aux normes internationales dans l'utilisation de la reconnaissance de la parole.

Nous cherchons à développer notre système de reconnaissance du fumeur pour qu'il soit capable d'identifier la durée de dépendance au tabac.

L'utilisation des modèles vocaux dépendent du contexte, pour la mise en place d'un système de classification pour différents types de maladies de la voix.

La réalisation d'un système de détection de l'anomalie vocale basant sur la combinaison de modèles Markov cachés et les modèles de mélange gaussiens en utilisant un mixage de dialectes Marocain.

Références

- Ali, Z., Alsulaiman, M., Elamvazuthi, I., Muhammad, G., Mesallam, T. A., Farahat, M., & Malki, K. H. (2016). Voice pathology detection based on the modified voice contour and SVM. Biologically Inspired Cognitive Architectures, 15, 10-18.
- Ameur, M., Bouhjar, A., & Boukhris, F. (2004). Initiation à la langue Amazigh. Institut Royal de la Culture Amazighe.
- Amini, M. R. (2001). Apprentissage automatique et recherche de l'information: application à l'extraction d'information de surface et au résumé de texte (Doctoral dissertation, Paris 6).
- Andries, P. (2008). Unicode 5.0 en pratique: Codage des caractères et internationalisation des logiciels et des documents. Dunod.
- Arbouz, C. (2016) Titre de l'œuvre : Écrire l'amazigh, Publié par : UPblisher, Nombre de pages : 157
- Arora, H. S., Singh, H., & Dhindaw, B. K. (2012). Some observations on microstructural changes in a Mg-based AE42 alloy subjected to friction stir processing. Metallurgical and Materials Transactions B, 43(1), 92-108.
- Ataa Allah FADOUA, et Boulaknadel SIHAM. Natural language processing for Amazigh language: Challenges and future directions. Language Technology for Normalisation of Less-Resourced Languages, 2012, vol. 19.
- Aupetit, S. (2005). Contributions aux modèles de Markov cachés: Métaheuristiques d'apprentissage, nouveaux modèles et visualisation de dissimilarité (Doctoral dissertation).
- Baggenstoss, P. M. (2001). A modified Baum-Welch algorithm for hidden Markov models with multiple observation spaces. IEEE Transactions on speech and audio processing, 9(4), 411-416.
- Bahl, L. R., Jelinek, F., & Mercer, R. L. (1983). A maximum likelihood approach to continuous speech recognition. IEEE transactions on pattern analysis and machine intelligence, (2), 179-190.
- Bahl, L., & Jelinek, F. (1975). Decoding for channels with insertions, deletions, and substitutions with applications to speech recognition. IEEE Transactions on Information Theory, 21(4), 404-411.
- Baker, J. K. (1975). Stochastic modeling as a means of automatic speech recognition. CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF COMPUTER SCIENCE.
- Baum, L. E., & Eagon, J. A. (1967). An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. Bulletin of the American Mathematical Society, 73(3), 360-363.
- Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. The annals of mathematical statistics, 41(1), 164-171.
- Benesty, J., Sondhi, M. M., & Huang, Y. (Eds.). (2007). Springer handbook of speech processing. Springer.
- Berbeche, K. (2014). Modèles de Markov Cachés: Application à La Reconnaissance Automatique de la Parole (Doctoral dissertation, Université Mouloud Mammeri).

- Boite, R. (2000). Traitement de la parole. PPUR presses polytechniques.
- Bouallegue, M. (2013, December). L'analyse factorielle pour la modélisation acoustique des systèmes de reconnaissance de la parole. Avignon.
- Boukhris F., Boumalk A., Elmoujahid E., Souifi H. (2008). La nouvelle grammaire de l'amazighe. Rabat, Maroc: IRCAM.
- Boukous, A. (1995). Société, langues et cultures au Maroc: Enjeux symboliques (No. 8). Faculté des lettres et des sciences humaines-Rabat.
- Bourlard, H., & Wellekens, C. J. (1990). Links between Markov models and multilayer perceptrons. IEEE Transactions on Pattern Analysis and Machine Intelligence, 12(12), 1167-1178.
- Busby, P. A., & Plant, G. L. (1995). Formant frequency values of vowels produced by preadolescent boys and girls. The Journal of the Acoustical Society of America, 97(4), 2603-2606.
- Caelen, J. (1979). Un modèle d'oreille (Doctoral dissertation, Toulouse).
- Calliope, L., & Fant, G. (1989). La parole et son traitement automatique. Paris: Masson.
- Celeux, G., Chauveau, D., & Diebolt, J. (1995). On stochastic versions of the EM algorithm.
- Chai, L., Sprecher, A. J., Zhang, Y., Liang, Y., Chen, H., & Jiang, J. J. (2011). Perturbation and nonlinear dynamic analysis of adult male smokers. Journal of voice, 25(3), 342-347.
- Chaker, S. (1992). Textes en linguistique berbère : introduction au domaine berbère. Paris : Editions d'Harmattan.
- Chaker, S. (1995). Linguistique berbère: études de syntaxe et de diachronie (Vol. 8). Peeters Publishers.
- Coleman, John, 2001, The vocal tract and larynx, Available from http://www.phon.ox.ac.uk/~jcoleman/phonation.htm.
- Coleman, R. O. (1971). Male and female voice quality and its relationship to vowel formant frequencies. Journal of speech and hearing research, 14(3), 565-577.
- Davis, K. H., Biddulph, R., & Balashek, S. (1952). Automatic recognition of spoken digits. The Journal of the Acoustical Society of America, 24(6), 637-642.
- Dénes, J. (1959). The representation of a permutation as the product of a minimal number of transpositions and its connection with the theory of graphs. Publ. Math. Inst. Hungar. Acad. Sci, 4, 63-70.
- Deng, L., Acero, A., Jiang, L., Droppo, J., & Huang, X. (2001, May). High-performance robust speech recognition using stereo training data. In 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221) (Vol. 1, pp. 301-304). IEEE.
- Devijver, P. A. (1985). Baum's forward-backward algorithm revisited. Pattern Recognition Letters, 3(6), 369-373.
- Dibazar, A. A., Berger, T. W., & Narayanan, S. S. (2006, August). Pathological voice assessment. In 2006 International Conference of the IEEE Engineering in Medicine and Biology Society (pp. 1669-1673). IEEE.
- Draper, M. H., Ladefoged, P., & Whitteridge, D. (1959). Respiratory muscles in

speech. Journal of Speech and Hearing Research, 2(1), 16-27.

- Dreyfus- Graf, J. (1950). Sonograph and sound mechanics. The Journal of the Acoustical Society of America, 22(6), 731-739.
- Durrand, J. (2009). L'alphabet phonétique international. Herrenschmidt, C., Mugnaioni, RM-Savelli, J. & Touratier, C. Le monde des écritures. Paris: Gallimard, 1-19.
- Dutrey, C. (2014). Analyse et détection automatique de disfluences dans la parole spontanée conversationnelle (Doctoral dissertation, Paris 11).
- El Yachi, R., Moro, K., Fakir, M., Bouikhalene, B., PETER, S. J., REDDI, K. K., ... & AL-ZHRANI, S. A. L. E. H. (2010). On the Recognition of Tifinaghe Scripts. Journal of Theoretical and Applied Information Technology, 20(2), 61-66.
- Farrús, M., Hernando, J., & Ejarque, P. (2007). Jitter and shimmer measurements for speaker recognition. In Eighth annual conference of the international speech communication association.
- Feller, W. (1958). An Introduction to probability theory and its applications, volume 1. John Willey, New York, 2nd edition.
- Forney, G. D. (1973). The viterbi algorithm. Proceedings of the IEEE, 61(3), 268-278.
- Freguson, J. D. (1980). Variable duration models for speech. In Proc. Symposium on the Application of Hidden Markov Models to Text and Speech, 1980.
- Furui, S. (2005). 50 years of progress in speech and speaker recognition research. ECTI Transactions on Computer and Information Technology (ECTI-CIT), 1(2), 64-74.
- Galand, L. (1988). Le berbère. Les langues dans le monde ancien et moderne, 207-242. Troisième partie : les langues chamito-sémitiques. Cohen & Perrot (eds.), Paris : Editions du CNRS.
- Ghio, A., & Pinto, S. (2007). Résonance sonore et cavités supralaryngées. Marchal, A. (1980). Les sons et la parole. Guérin.
- González, J. (2004). Formant frequencies and body size of speaker: a weak relationship in adult humans. Journal of phonetics, 32(2), 277-287.
- Gonzalez, J., & Carpi, A. (2004). Early effects of smoking on the voice: a multidimensional study. Medical Science Monitor, 10(12), CR649-CR656.
- Guimarães, I., & Abberton, E. (2005). Fundamental frequency in speakers of Portuguese for different voice samples. Journal of voice, 19(4), 592-606.
- Hacine-Gharbi, A. (2018). Sélection de paramètres acoustiques pertinents pour la reconnaissance de la parole (Doctoral dissertation).
- Hamani, H. (2015). Synthèse, caractérisation et étude du pouvoir inhibiteur de nouvelles molécules bases de Schiff. setif: Université Ferhat Abbas–Setif (Doctoral dissertation, These de doctorat).
- Hansen, J. H., Sarikaya, R., Yapanel, U., & Pellom, B. (2001). Robust speech recognition in noise: an evaluation using the spine corpus. In Seventh European Conference on Speech Communication and Technology.
- Haque, S., Togneri, R., & Zaknich, A. (2009). Perceptual features for automatic speech recognition in noisy environments. Speech communication, 51(1), 58-75.

- Hartley, H. O. (1958). Maximum likelihood estimation from incomplete data. Biometrics, 14(2), 174-194.
- Haton, J. P. (1992). Reconnaissance automatique de la parole et dialogue oral homme-machine.
- Haton, J. P., Cerisara, C., Fohr, D., Laprie, Y., & Smaïli, K. (2006). Reconnaissance automatique de la parole: Du Signal à son Interprétation. Dunod.
- Hermansky, H.: Perceptual linear predictive (PLP) analysis for speech. Journal of Acoustic Society of America, Vol. 87, pp. 1738-1752, 1990
- Hess, W., & O'Shaughnessy, D. (1984). Pitch Determination of Speech Signals: Algorithms and Devices by Wolfgang Hess.
- Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. Neural computation, 18(7), 1527-1554.
- Huang, X., Acero, A., Hon, H. W., & Reddy, R. (2001). Spoken language processing: A guide to theory, algorithm, and system development. Prentice hall PTR.
- Janicki, A. (2013). Non-linguistic vocalisation recognition based on hybrid GMM-SVM approach. In INTERSPEECH (pp. 153-157).
- Jiang, H. (2005). Confidence measures for speech recognition: A survey. Speech communication, 45(4), 455-470.
- Kalamani, M., Valarmathy, S., Poonkuzhali, C., & Catherine, J. N. (2014, January). Feature selection algorithms for automatic speech recognition. In 2014 International Conference on Computer Communication and Informatics (pp. 1-7). IEEE.
- Khelifa M., (2017). Contribution au développement des systèmes de reconnaissance automatique de la parole pour la langue Arabe classique à l'aide de l'approche Markovienne. Thèse de ENSIAS Rabat.
- Lachhab, O. (2017). Reconnaissance Statistique de la Parole Continue pour Voix Laryngée et Alaryngée (Doctoral dissertation).
- Lahouti, F., Fazel, A. R., Safavi-Naeini, A. H., & Khandani, A. K. (2006). Single and double frame coding of speech LPC parameters using a lattice-based quantization scheme. IEEE transactions on audio, speech, and language processing, 14(5), 1624-1632.
- Lee, K. F. (1988). Automatic speech recognition: the development of the SPHINX system (Vol. 62). Springer Science & Business Media.
- Lefèvre, F. (2000). Estimation de probabilité non-paramétrique pour la reconnaissance markovienne de la parole (Doctoral dissertation, ANRT).
- Léothaud, G. (2004). Théorie de la phonation.
- Markov, A. A. (1913). An example of statistical investigation in the text of "Eugene oneygin" illustrating coupling of "test" in chains. In Processings of Academic Scientific St. Petersburg, IV, pages 153 162.
- Martin, T. B., Nelson, A. L., & Zadell, H. J,(1964). Speech Recognition b Feature Abstraction Techniques , Tech.Report AL-TDR-64-176,Air Force Avionics Lab.
- McKeating, K., Bali, I. M., & Dundee, J. W. (1988). The effects of thiopentone and propofol on upper airway integrity. Anaesthesia, 43(8), 638-640

- Meo A.R, Righini G., (1965), “Riconoscitore istantaneo disuoni vocalici”, Alta Frequenza 34, 256-263.
- Muhammad, G., Mesallam, T. A., Malki, K. H., Farahat, M., Alsulaiman, M., & Bukhari, M. (2011). Formant analysis in dysphonic patients and automatic Arabic digit speech recognition. Biomedical engineering online, 10(1), 41.
- Nejme, F., Boulaknadel, S., & Aboutajdine, D. (2013). Analyse Automatique de la Morphologie Nominale Amazighe. In Actes de la conférence du Traitement Automatique du Langage Naturel (TALN).
- Nendaz, M., Charlin, B., Leblanc, V., & Bordage, G. (2005). Le raisonnement clinique: données issues de la recherche et implications pour l’enseignement. Pédagogie médicale, 6(4), 235-254
- Olson H.F, Belar H., (1956), “Phonetic Typerwiter”.J.Acoust. So Amer. 28, 1072- 1081.
- Ono, N., Miyamoto, K., Le Roux, J., Kameoka, H., & Sagayama, S. (2008, August). Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram. In 2008 16th European Signal Processing Conference (pp. 1-4). IEEE.Nicholas, F., & Carswell, I. (2007). U.S. Patent Application No. 11/465,735.
- Paulmann, S., & Pell, M. D. (2011). Is there an advantage for recognizing multi-modal emotional stimuli?. Motivation and Emotion, 35(2), 192-201.
- Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... & Silovsky, J. (2011). The Kaldi speech recognition toolkit. In IEEE 2011 workshop on automatic speech recognition and understanding (No. CONF). IEEE Signal Processing Society.
- Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2), 257-286.
- Rabiner, L. R., Levinson, S. E., & Sondhi, M. M. (1983). On the application of vector quantization and hidden Markov models to speaker- independent, isolated word recognition. Bell System Technical Journal, 62(4), 1075-1105.
- Ridouane, R. (2003). Suites de consonnes en berbère: phonétique et phonologie (Doctoral dissertation).
- Roach, P. (2010). English phonetics and phonology fourth edition: A practical course. Ernst Klett Sprachen.
- Sakai T., Doshita S., (1962), “Recognition of Japanese vowls”, Proc. IFIP Congress, Munich.
- Sakoe H. and Chiba S., (1978) Dynamic programming algorithm optimization for spoken word recognition ,IEEE Trans. Acoustics, Speech, Signal Proc., ASSP- 26(1).pp.43- 49.
- Samaria, F. S., & Harter, A. C. (1994, December). Parameterisation of a stochastic model for human face identification. In Proceedings of 1994 IEEE workshop on applications of computer vision (pp. 138-142). IEEE.
- Satori, H. (2009). Reconnaissance automatique de la langue Arabe en utilisant le système Sphinx.
- Satori, H., & Elhaoussi, F. (2014). Investigation Amazigh speech recognition using CMU tools. International Journal of Speech Technology, 17(3), 235-243.
- Satori, H. (2015). Contribution à la Reconnaissance Automatique de l’Amazighe à Base de

Modèles de Markov Cachés : l'Habilitation Universitaire présentée à la Faculté des sciences Dhar El Mahraz.

- Shah, D. H., & Shah, T. V. (2015). Speech Recognition: An Approach to Modernization. International Journal of Advanced Research in Computer Science and Software Engineering, 5(8), 1-6.
- Shannon, C. E. (1948). A mathematical theory of communication. The Bell system technical journal, 27(3), 379-423
- Silberztein, M. (2007, June). An alternative approach to tagging. In International Conference on Application of Natural Language to Information Systems (pp. 1-11). Springer, Berlin, Heidelberg.
- Skinner, T., Kloker, D., & Medress, M. (1976, April). A speech recognition system for connected word sequences. In ICASSP'76. IEEE International Conference on Acoustics, Speech, and Signal Processing (Vol. 1, pp. 434-437). IEEE
- Slimane, M. 2002. Les chaines de Markov cachés : définitions, algorithmes, architectures. Rapport interne n°260, Université François-Rabelais de Tours, Laboratoire d'Informatique, Tours, France.
- Strik, H., & Cucchiarini, C. (1999). Modeling pronunciation variation for ASR: A survey of the literature. Speech Communication, 29(2-4), 225-246.
- Tafiadis, D., Toki, E. I., Miller, K. J., & Ziavra, N. (2017). Effects of early smoking habits on young adult female voices in Greece. Journal of Voice, 31(6), 728-732
- Teixeira, J. P., Oliveira, C., & Lopes, C. (2013). Vocal acoustic analysis–jitter, shimmer and hnr parameters. Procedia Technology, 9, 1112-1122.
- Verdonck-de Leeuw, I. M., & Mahieu, H. F. (2004). Vocal aging and the impact on daily life: a longitudinal study. Journal of voice, 18(2), 193-202.
- Vincent, I., & Gilbert, H. R. The effects of cigarette smoking on the female voice. Logopedics Phoniatrics Vocology, 2012, 37(1), 22-32.
- Vintsyuk, T. K. (1968). Speech discrimination by dynamic programming. Cybernetics, 4(1), 52-57.
- Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE transactions on Information Theory, 13(2), 260-269.
- Von Békésy, G., & Wever, E. G. (1960). Experiments in hearing (Vol. 8). New York: McGraw-Hill.
- Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., ... & Woelfel, J. (2004). Sphinx-4: A flexible open source framework for speech recognition.
- Wertzner, H. F., Schreiber, S., & Amaro, L. (2005). Analysis of fundamental frequency, jitter, shimmer and vocal intensity in children with phonological disorders. Brazilian journal of otorhinolaryngology, 71(5), 582-588.
- Wester, M. (2003). Pronunciation modeling for ASR–knowledge-based and data-derived methods. Computer Speech & Language, 17(1), 69-85.
- Woodland, P. C., Leggetter, C. J., Odell, J. J., Valtchev, V., & Young, S. J. (1995, May). The 1994 HTK large vocabulary speech recognition system. In Acoustics, Speech, and Signal Processing, International Conference on. IEEE, vol. 1, pp. 73-76.

- Wuyts, F. L., Bodt, M. S. D., Molenberghs, G., Remacle, M., Heylen, L., Millet, B., ... & Heyning, P. H. V. D. (2000). The dysphonia severity index: an objective measure of vocal quality based on a multiparameter approach. Journal of speech, language, and hearing research, 43(3), 796-809.
- Yu, P., Ouaknine, M., Revis, J., & Giovanni, A. (2001). Objective voice analysis for dysphonic patients: a multiparametric protocol including acoustic and aerodynamic measurements. Journal of voice, 15(4), 529-542.
- Zenkouar, L. 2004. L'écriture Amazighe Tifinaghe et Unicode. Revue Etudes et Documents Berbères n°22, pp. 175--192.

Annexe A

Liste des Publications

Journaux internationaux :

Satori, H., Zealouk, O., Satori, K., & Elhaoussi, F. (2017). Voice comparison between smokers and non-smokers using HMM speech recognition system. *International Journal of Speech Technology*, *20*(4), 771-777.

Zealouk, O., Satori, H., Hamidi, M., Laaidi, N., & Satori, K. (2018). Vocal parameters analysis of smoker using Amazigh language. International Journal of Speech Technology, 21(1), 85-91.

Zealouk, O., Satori, H., Hamidi, M., & Satori, K. (2019). Speech recognition for Moroccan dialects: feature extraction and classification methods. Journal of Advanced Research in Dynamical and Control Systems, 11(2), 1401-1408.

Zealouk, O., Satori, H., Laaidi, N., Hamidi, M., & Satori, K. (2020). Noise Effect on Amazigh Digits in Speech Recognition system. International Journal of Speech Technology, 23, 885–892.

Conférences internationales:

Zealouk, O., Satori, H., Hamidi, M., & Satori, K. (2018, October). Voice pathology assessment based on automatic speech recognition using Amazigh digits. In Proceedings of the 2nd International Conference on Smart Digital Environment (pp. 100-105).

Zealouk, O., Satori, H., Hamidi, M., & Satori, K. (2018). Automatic Speech Recognition for Moroccan Dialects: A Review. The 1st International Conference of Computer Science and Renewable Energies « ICCSRE'2018 ».

Zealouk, O., Satori, H., Hamidi, M., & Satori, K. (2020). Pathological Detection Using HMM Speech Recognition-Based Amazigh Digits. In Embedded Systems and Artificial Intelligence (pp. 281-289). Springer, Singapore.

Zealouk, O., Hamidi, M., Satori, H., & Satori, K. (2020). Amazigh Digits Speech Recognition System Under Noise Car Environment. In Embedded Systems and Artificial Intelligence (pp. 421-428). Springer, Singapore.

Zealouk, O., Satori, H., Hamidi, M., & Satori, K. Pathological voice detection using automatic speech recognition based on Amazigh language. International Conference on Information and Communication Technologies for Amazigh « TICAM 2018 »- Rabat

Annexe B

Algorithme d'apprentissage et de décodage

Pour réaliser un apprentissage étiqueté, aussi comme il est défini dans littérature par l'apprentissage de Viterbi, on dispose de deux informations : la séquence d'observations $O= (O^1, O^2,.., O^T)$ et la séquence d'états cachés $Q= (Q^1, Q^2,.., Q^T)$ qui a engendré la séquence précédente. Le modèle que l'on cherche à maximiser est $P (V=O, S=Q/\lambda)$. Pour faire la maximisation, on doit calculer différentes transitions du système entre les états et la probabilité. En général, avec cette méthode d'apprentissage on considère plusieurs séquences d'observations au même temps. Dans ce cas, on utilise le comptage des différentes longueurs des séquences de transitions $T= (T^1, T^2 \ldots, T^K)$ du système avec la considération de toutes les séquences simultanément de manière indistincte. L'algorithme d'apprentissage étiqueté est présenté par l'algorithme ci-dessous (Algorithme B.1). Sa complexité est $O (N^2 + NM + T)$ en désignant par T, la longueur totale des séquences d'observations considérées.

$\forall i = 1....N,\ x_i = 0$

$\forall i = 1....N, j = 1....N,\ y_{i,j} = 0$

$\forall i = 1....N, j = 1....M,\ z_{i,j} = 0$

Pour $K=1$ à K Faire

  Incrémenter $x_{q_1^K}$

  Pour $t=1$ à T^k Faire

  Incrémenter $y_{q_1^K}, o_t^k$

  Si $t < T^k$ *Alors*

  Incrémenter $z_{q_1^K}, q_{t+1}^k$

  Fin Si

  Fin Pour

Fin Pour

$$\forall i = 1....N,\ \pi_i = \frac{x_i}{\sum_r^N x_r}$$

$$\forall i = 1....N, j = 1....N,\ a_{i,j} = \frac{y_{i,j}}{\sum_{r=1}^N y_{i,r}}$$

$$\forall i = 1....N, j = 1....M,\ b_i(j) = = \frac{z_{i,j}}{\sum_{r=1}^M z_{i,r}}$$

Algorithme B.1 : Algorithme d'apprentissage étiqueté

Dans ce type de L'apprentissage, la performance n'est pas efficace dans le cas où le nombre de séquences d'observations ou les séquences d'états cachés ou le nombre d'apparitions d'un ou plusieurs motifs est trop réduit, car le modèle n'arrive pas à généraliser ce qu'il doit reconnaître. Pour trouver des solutions à ce type de problème il est suffisant d'insérer un coefficient de lissage lors de l'estimation des probabilités. En notant c >0 le coefficient de lissage, l'algorithme est donné par l'algorithme B.2. Dans cet algorithme, le coefficient de lissage est identique pour toutes les probabilités, mais rien n'empêche de le choisir différent pour chacune d'elles, afin d'inclure des connaissances expertes dans l'apprentissage.

$\forall i = 1....N,\ x_i = 0$

$\forall i = 1....N, j = 1....N,\ y_{i,j} = 0$

$\forall i = 1....N, j = 1....M,\ z_{i,j} = 0$

Pour $K=1$ à K Faire

Incrémenter $x_{q_1^K}$

Pour $t=1$ à T^k Faire

Incrémenter $y_{q_1^K}$, o_t^k

Si $t < T^k$ *Alors*

Incrémenter $z_{q_1^K}$, q_{t+1}^k

Fin Si

Fin Pour

Fin Pour

$$\forall i = 1....N,\ \pi_i = \frac{c+x_i}{N_c+\sum_r^N x_r}$$

$$\forall i = 1....N, j = 1....N,\ a_{i,j} = \frac{c+y_{i,j}}{N_c+\sum_{r=1}^N y_{i,r}}$$

$$\forall i = 1....N, j = 1....M,\ b_i(j) = = \frac{c+z_{i,j}}{N_c+\sum_{r=1}^M z_{i,r}}$$

Algorithme B.2 : Algorithme d'apprentissage étiqueté avec lissage

Exemple simple de l'algorithme Viterbi :

- Une personne en vacances envoie une carte postale mentionnant les activités suivantes : jour 1: plage ; jour 2 : magasinage ; jour 3 : sieste.
- On veut en déduire la séquence météorologique sous-jacente probable sachant que :

- Les conditions météorologiques suivent une chaîne de Markov à 2 états : Pluie et soleil
- On possède des statistiques sur le comportement des touristes selon les états

Transition d'état

A=

0,7	0,3
0,4	0,6

Emission de symboles par les états

B=

0,4	0,3
0,1	0,6
0,5	0,1

État initial

π=

0,6
0,4

Σ = {Pluie=1, Soleil=2}, Ω= {magasinage=1, plage=2, sieste=3}

- Séquence d'observations : *O = 2, 1, 3*

  Calcule :

- Étape 1

  $\alpha_1(1) = \pi_1 * b_1(2) = 0.6*0.1 = 0.06,$

  $\alpha_1(2) = \pi_2 * b_2(2) = 0.4*0.6 = 0.24,$

  $\Psi_1(1) = \Psi_1(2)=0$

- Étape 2
  - t = 2
    - $\alpha_2(1) = \max_j (\alpha_1(j)*a_{j1})*b_1(1)$

      $= \max \{0.06*0.7, 0.24*0.4\}*0.4 = 0.0384$

      $\Rightarrow \Psi_2(1) = \arg\max_j (\alpha_1(j)*a_{j1}) = 2$

    - $\alpha_2(2) = \max_j (\alpha_1(j)*a_{j2})*b_2(1)$

      $= \max \{0.06*0.3, 0.24*0.6\}*0.3 = 0.0432$

=> $\Psi_2(2) = 2$

- t = 3
  - $\alpha_3(1) = \max_j (\alpha_2(j)*a_{j1})*b_1(3)$

    = max {0.0384*0.7, 0.0432*0.4}*0.5 = 0.01344

    => $\Psi_3(1) = 1$
  - $\alpha_3(2) = \max_j (\alpha_2(j)*a_{j2})*b_2(3)$

    = max {0.0384*0.3, 0.0432*0.6}*0.1 = 0.002592

    => $\Psi_3(2) = 2$

- Étape 3 :

  $S(3) = \arg\max_j \{\alpha_3(1), \alpha_3(2)\} = 1$

- Étape 4 :

  $S(2) = \Psi_3(S(3)) = 1$ $S(1) = \Psi_2(S(2)) = \mathbf{2}$