ROYAUME DU MAROC

Université Sultan Moulay Slimane

Faculté des Sciences et Techniques

Département d'informatique

Béni Mellal

N° d'ordre : 69/2015

Centre d'Etudes Doctorales « Sciences et Techniques »

Formation Doctorale « Mathématiques et Physique Appliquées »

THÈSE

Présentée par

Ahmed EL GHAZI

Pour obtention du grade de

Docteur

Spécialité : Informatique

Reconnaissance automatique de la parole : application aux dialectes marocains

Soutenue publiquement le 14 mars 2015 devant les membres du jury :

Pr. A. ZEGHAL	Faculté des Sciences et Techniques, Béni Mellal	**Président**
Pr. A. HAQIQ	Faculté des Sciences et Techniques, Settat	**Rapporteur**
Pr. A. HAIR	Faculté des Sciences et Techniques, Béni Mellal	**Rapporteur**
Pr. A. MERBOUHA	Faculté des Sciences et Techniques, Béni Mellal	**Rapporteur**
Pr. C. DAOUI	Faculté des Sciences et Techniques, Béni Mellal	**Directeur de thèse**
Pr. N. IDRISSI	Faculté des Sciences et Techniques, Béni Mellal	**Co-directrice de thèse**
Pr. M. FAKIR	Faculté des Sciences et Techniques, Béni Mellal	**Examinateur**

Introduction générale

1. Problématique générale et contributions de la thèse.

Actuellement l'existence des dialectes constitue en général un défi pour le traitement automatique des langues, car ils ajoutent une autre série de variation de dimensions à partir d'une norme connue. Sur le plan pratique, les services liés aux technologies pour ces dialectes sont inexistants. Le problème est particulièrement intéressant en langue Arabe et ses différents dialectes qui restent presque absents de la plupart des systèmes de dialogue homme-machines [1], ce qui reflète la pauvreté de la recherche dans le domaine de traitement de ses dialectes.

Dans ce mémoire de thèse, nous nous concentrons sur les dialectes populaires qui ont peu de ressources informatiques pour les implémenter dans la technologie en langue naturelle. Certains de ces dialectes sont des langues majoritaires des pays en voie de développement car elles peuvent être parlées par un grand nombre de locuteurs. Dans ce cadre, nous avons considéré le Darija et le Tamazight qui sont les deux dialectes les plus dominants au Maroc, ils sont notamment parlés par vingt millions de personnes. Pour ces dialectes, très peu de ressources électroniques utilisables sont disponibles. Ainsi, nous proposons de nouveaux systèmes de traitement automatique des dialectes en question. Exactement, on s'intéresse à la reconnaissance automatique de la parole et la vérification automatique du locuteur, pour ces deux variantes, dans le but de créer un système de sécurité basé sur les empreintes vocales. Ce système suscite beaucoup d'intérêt et peut être utilisé dans de nombreuses applications, entre autres les téléphones portables, les guichets automatiques, l'accès aux bases de données, la surveillance des appels téléphoniques et le contrôle de l'identité personnelle via des appels téléphoniques.

L'approche d'identification vocale proposée vise à améliorer la fiabilité et la sécurité de tous les systèmes utilisant le traitement automatique de la voix comme méthode d'identification personnelle. La première partie de ce mémoire repose sur la reconnaissance automatique de la parole, elle est consacrée à la création de quelques modèles acoustiques

DEDICACES

A ma famille

A mes amis

A tous ceux qui m'ont aidé de près ou de loin

destinés aux deux dialectes marocains : Tamazight et Darija. Nous avons utilisé une base de données audio développée au sein du Laboratoire de Traitement de l'Information et Aide à la Décision de la faculté des sciences et techniques de Béni Mellal. Cette optique qui vise la création d'un système de vérification du mot de passe, reste insuffisante pour créer un système de sécurité fiable suite à l'instabilité liée au signal de la parole, ce dernier peut être changé ou imité par une autre personne. De ce fait, dans la deuxième partie de ce mémoire, on ajoute une procédure de vérification du locuteur et nous proposons ainsi un nouveau système de sécurité fondé sur la reconnaissance vocale et l'identification du locuteur.

La première partie de ce travail (chapitre II) est consacrée à la reconnaissance automatique des mots isolés des dialectes marocains en se basant sur le modèle de Markov caché (MMC) et en faisant une étude comparative des résultats obtenus avec ceux de la méthode de programmation dynamique et les réseaux de neurones multicouches (PMC). Dans ce sens, on constate que les résultats obtenus par le modèle de Markov caché sont les meilleurs, ce qui démontre la robustesse de la modélisation stochastique dans la reconnaissance automatique de la parole. Ensuite, nous avons considéré le cas des mots enchainés, en prenant comme exemple les chiffres enchainés en Tamazight, les résultats obtenus sont meilleurs en comparaison avec le cas des chiffres isolés. L'objectif principal de cette partie du travail est la mise en place d'un système de reconnaissance de mots de passe, en effet un client est invité à prononcer son mot de passe, ensuite le système RAP envoie la série des chiffres reconnus en paramètres à un système de vérification. Ce dernier permet de décider sur la validation de mot de passe en comparant celui-ci avec les données personnelles dans une base de données. Si le mot de passe est validé le client est redirigé vers l'étape suivante qui consiste à la vérification de son identité (étape de reconnaissance du locuteur).

La reconnaissance automatique du locuteur consiste à reconnaitre l'identité d'une personne par l'analyse de sa voix. Néanmoins, d'autres éléments autres que la voix peuvent être utilisés pour authentifier une personne, tels que les empreintes digitales ou les empreintes génétiques. Contrairement à la voix, ces derniers éléments sont des composantes du corps humain, ils ne varient pas (ou très peu) dans le temps et ne peuvent pas être modifiés sciemment par un individu. De par ces propriétés, le terme de *biométrie* est souvent employé pour les définir et souligne leur très grande fiabilité. Les gestes de parole ne sont, en aucune façon, un élément du corps humain et ne sont pas reproductibles à l'identique dans le temps. Dans ce sens, les dénominations de *biométrie* ou d'*empreintes* vocales ne sont pas appropriées pour caractériser la voix. Cependant, la voix reste pour certaines applications (services accessibles par réseau téléphonique) le seul élément disponible pour authentifier l'utilisateur.

REMERCIEMENTS

Finalement, ces cinq années de thèse sont passées extrêmement vite. Cinq ans avant, je n'aurai pas imaginé me lancer dans ce long et passionnant travail.

Mon intérêt pour l'informatique s'est principalement développé au cours de mes années d'études, au Master et au Doctorat. Mes balbutiements en reconnaissance automatique de la parole se sont développés au cours de la première année de thèse sous la direction de Monsieur Cherki DAOUI. Ce domaine qui me paraissait si mystique m'a immédiatement passionné car j'ai eu la chance de travailler avec une équipe sérieuse, qui m'a beaucoup apporté tant dans la connaissance que sur le plan humain. Je tiens particulièrement à remercier Monsieur Cherki DAOUI mon directeur de thèse et Madame Najlae IDRISSI ma Co-encadrante qui m'ont soutenu tout au long de ce travail et merci encore une fois pour leurs conseils d'or. M. Cherki DAOUI a dirigé ma thèse et m'a aidé dans toutes les démarches relatives à celle-ci. Il m'a encadré au long des cinq années, il a su habilement me remotiver dans les passages à vide, et m'a fait partager énormément de connaissances et d'expériences. Je remercie également Monsieurs Mohamed FAKIR, Belaid BOUIKHALENE, qui, au sein du laboratoire, ont été patients vis-à-vis de mes nombreuses interrogations.

Je remercie Monsieurs A. HAQIQ, A. HAIR et A. MERBOUHA d'avoir accepté d'être les rapporteurs de ma thèse, mais également pour leurs corrections et remarques pertinentes vis-à-vis de mon document. Je remercie également le président de mon jury Monsieur Ahmed ZEGHAL .

Un grand merci également à l'ensemble des personnels de la Faculté des Sciences et Techniques Béni Mellal avec qui j'ai partagé ces cinq années. Je tiens aussi à remercier mes proches Cécile, Mon frère, ma mère, qui m'ont apporté leur soutien au cours de ces cinq dernières années et qui m'ont supporté dans les moments de stress.

Finalement, je tiens à remercier sincèrement tous les personnes qui m'ont aidé et qui ont participé de près ou de loin au succès de cette thèse. Je remercie encore une fois tous les personnels administratif de la faculté qui nous ont aidé et qui nous donnent une précieuse aide durant les années de préparation de thèse.

La deuxième partie de ce travail (chapitre III et chapitre IV) se focalise à la reconnaissance automatique du locuteur (RAL), ce qui a permis d'achever notre nouveau système de sécurité basé sur la validation d'un mot de passe et la vérification de l'identité du locuteur qui l'a prononcé. On a commencé à donner un aperçu sur les différentes approches utilisées en RAL, puis un aperçu sur les réseaux de neurones multicouches qui sont utilisés pour modéliser les différents locuteurs. Ainsi, nous nous sommes amenés à créer un système de sécurité rassemblant la reconnaissance vocale et l'identification automatique du locuteur. Ce système de sécurité se compose de deux phases consécutives :

- ✓ La première phase permettant la validation d'un mot de passe. Elle est basée sur les systèmes de reconnaissance automatique des dialectes marocains présentés dans le troisième chapitre.
- ✓ La deuxième phase s'intéresse à la vérification du locuteur. Celle-ci se base sur un modèle référence de chaque utilisateur du système afin de vérifier l'identité de la personne qui a prononcé le mot de passe.

2. Structure du document

Ce document est organisé comme suit :

Chapitre 1 : Consacré à l'état de l'art de la reconnaissance automatique de la parole et aux différentes approches théoriques utilisées dans ce sens. Dans ce chapitre, nous allons décrire précisément la paramétrisation du signal vocale et les outils de sa modélisation, à savoir le modèle de Markov caché (MMC), la programmation dynamique (DTW) et les réseaux de neurones multicouches (PMC).

Chapitre 2 : Focalisé aux résultats obtenus, la première partie de ce chapitre présente deux systèmes de reconnaissance proposée pour les mots isolés du Darija et Tamazight et un troisième système rassemblant ces deux dialectes. La deuxième partie propose un nouveau système RAP pour les chiffres enchainés en Tamazight.

Chapitre 3 : Consacré à la description des techniques utilisées pour la reconnaissance automatique du locuteur (RAL). Dans le premier volet de ce chapitre, on expose ces quatre grandes approches à savoir le modèle mélange de gaussiennes, la programmation dynamique, le modèle de Markov Caché et les réseaux de neurones artificiels.

TABLE DES MATIERES

Chapitre 4 : Destiné à l'architecture du système de sécurité proposé et les résultats expérimentaux. Dans ce sens, nous donnons les résultats liés à la reconnaissance automatique du locuteur et les résultats d'évaluation du système de sécurité proposé.

Chapitre 5 : Dédié aux conclusions et perspectives, en résumant nos contributions et nos principaux résultats ainsi qu'en ouvrant des perspectives pour des futurs travaux de recherche.

Chapitre 2 : RECONNAISSANCE AUTOMATIQUE DES DIALECTES MAROCAINS

Chapitre 1

Reconnaissance automatique de la parole

1. Introduction

La reconnaissance automatique de la parole est un domaine actif depuis le début des années 50. Les avancées réalisées dans ce domaine permettent aujourd'hui de fournir des systèmes de reconnaissance plus performants, le but est d'obtenir une transcription parfaite de la parole. Les enjeux actuels sont nombreux pour ce domaine, ces systèmes sont plus en plus utilisés par de nombreuses applications (dialogue homme-machines, traduction automatique, indexation des documents audio, recherche d'information dans un flux audio,…).

Afin d'obtenir un système RAP plus performant, il faut bien gérer les spécificités de la parole. La première difficulté réside dans le fait que le document audio à transcrire peut contenir les segments de la parole pour des personnes différentes : le signal audio peut varier selon le sexe (homme/femme), son âge, son mode d'élocution, son niveau de stress ou encore son accent. Selon le contexte ou le locuteur, la prononciation des mots peut considérablement changer. A ces difficultés peuvent s'ajouter les conditions d'enregistrement du signal audio (microphone, environnement d'enregistrement) et le bruit extérieur qui parasitent le signal. Enfin, le découpage du signal en phrase et en mots n'est pas simple pour un système RAP, car ce flux est continu et aucune frontière n'est clairement définie.

Dans ce chapitre, nous allons s'intéresser au principe général d'un système RAP, en décrivant toutes les étapes indispensables au système pour transformer un signal audio en une transcription, mais aussi la façon dont les systèmes statistiques gèrent les différentes difficultés énoncées précédemment. Dans ce sens, nous présentons la méthode dont ils sont extraits les paramètres acoustiques du signal de la parole puis nous donnons la théorie principale de ces systèmes statistiques.

2. Signal de la parole

La parole est un signal réel et continu, d'énergie finie, non stationnaire. Sa structure est complexe et variable dans le temps (il est quasi-stationnaire pour les sons voisés, aléatoire pour les sons fricatifs et impulsionnels pour les sons occlusifs). La complexité du signal vocal provient de plusieurs facteurs, nous donnons par exemple la variabilité du signal de la parole, la redondance du signal acoustique et les effets de coarticulation dans la parole continue qui doivent être pris en compte lors de la construction d'un système de reconnaissance vocale.

Chapitre 3 : RECONNAISSANCE ET VERIFICATION DU LOCUTEUR

La production du signal de la parole fait intervenir plusieurs organes. La source de la parole provient des poumons qui émettent un flux d'air. Ce flux d'air va traverser le larynx pour faire vibrer ou non les cordes vocales. Il va ensuite traverser le conduit vocal (cavité nasale et buccale) et les articulateurs tels que les lèvres et la langue (figure 1.1). Cet ensemble agit comme un filtre, considéré comme linéaire, dont la réponse impulsionnelle comporte des fréquences de résonance caractérisées par des pics, appelés formants, dans le spectre du signal de sortie. Le signal résultant est globalement non stationnaire mais peut être considéré comme stationnaire sur de très courtes périodes, de l'ordre de 20ms (signal pseudo-stationnaire). Sur un segment de parole de cette longueur la voix est habituellement et schématiquement séparée en deux classes distinctes :

1. Voisée lorsqu'il y a vibration des cordes vocales, le signal est alors quasi-périodique [2,3].
2. Non voisée dans le cas d'un simple soufflement, le signal est alors considéré comme aléatoire [2,3].

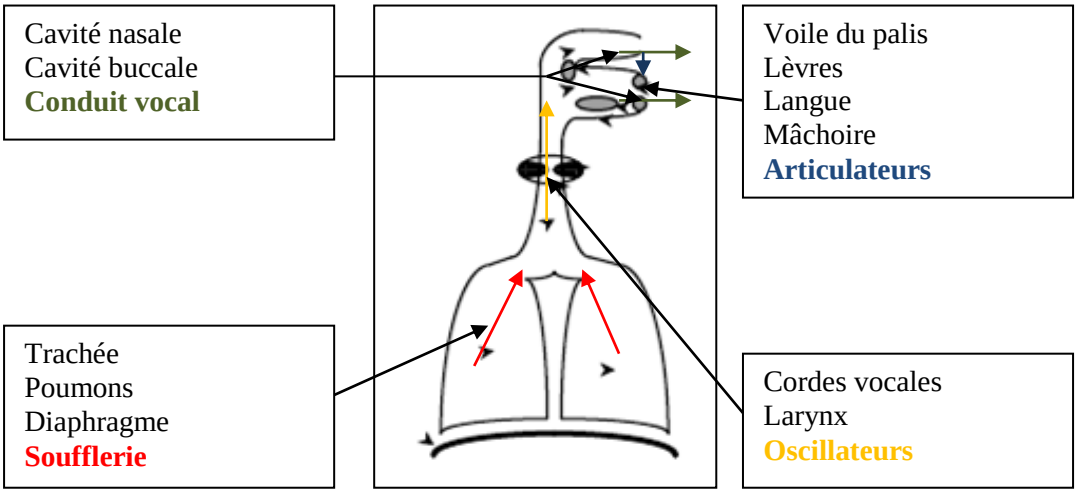


Figure 1.1 : Modèle physiologique de la production de la parole

Dans le premier cas, la source d'excitation est modélisée par un train d'impulsions périodiques, de fréquence fondamentale F_0, qui correspond à la fréquence de vibration des cordes vocales. Dans le second cas, la source est modélisée par un bruit blanc. Cette représentation binaire de la production de la parole a été introduite par [4,5]. Elle est reprise sur la figure 1.2.

Chapitre 4 : TRAITEMENT DE LA PAROLE ET SECURITE

Chapitre 5 : CONCLUSIONS ET PERSPECTIVES

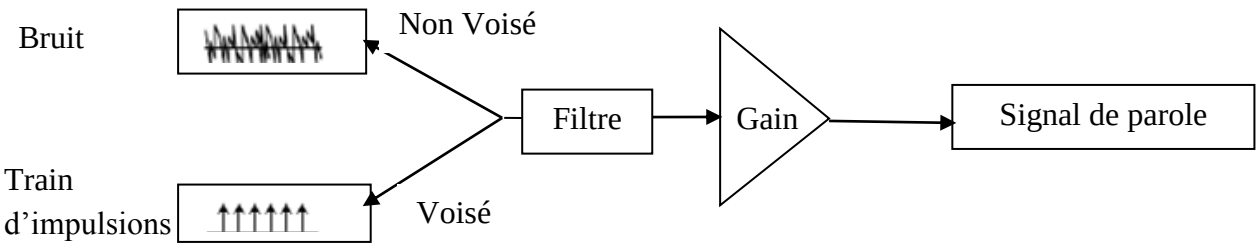


Figure 1.2 ***:*** *Modèle de production de la parole.*

3. Difficultés liées au signal de la parole

3.1. Variabilité du signal de la parole

Le signal de la parole n'est pas un signal ordinaire, ses propriétés varient au cours du temps, ces variations sont dus à :

- ✓ **La variabilité intra-locuteur** : la voix d'un locuteur dépend de plusieurs facteurs, psychologiques et physiologiques, on ne peut pas trouver un locuteur qui peut prononcer deux fois un mot de la même façon. Parmi les causes de la variabilité intra-locuteur on cite la vitesse d'élocution, l'état du locuteur (malade, sain, fatigue, triste, joyeux,…), l'âge du locuteur…
- ✓ **La variabilité interlocuteur** : la prononciation d'un mot diffère d'un locuteur à un autre et cela dépend de plusieurs paramètres, comme la forme de la cavité nasale, le sexe du locuteur, l'accent du locuteur, …
- ✓ **L'influence contextuelle** : la prononciation d'un mot diffère en fonction des sons qui l'entourent, de même nous remarquons que les sons des extrémités d'un mot peuvent être modifiés ou supprimés en fonction des sons qui sont adjacents.
- ✓ **Facteurs technologiques** : le signal de la parole dépend aussi de l'acoustique du milieu d'enregistrement (salle sourde ou environnement bruité), qualité du microphone, qualité du réseau téléphonique (le cas d'enregistrement à partir de la ligne téléphonique).

3.2 . Redondance du signal de la parole

Le signal de la parole est caractérisé par sa redondance, il faut un traitement au préalable pour éliminer les informations inutiles. Prenons par exemple un signal échantillonné à 16Khz sur 16 bits, le débit des informations est de 256 Kbits/s. Ceci nécessite un temps très élevé pour faire certains traitements sur ce signal, d'où la recherche d'une représentation plus compacte et moins redondante du signal s'avère nécessaire.

3.3 . Influence du contexte

Au plan articulatoire, la coarticulation peut se définir comme l'influence qui s'exerce entre deux contigus. Ce phénomène produit des interférences dans le signal de la parole, ce qui peut entrainer la disparition de certains sons dans la phrase prononcée. La prononciation

Résumé

Le traitement automatique de la parole suscite actuellement un grand intérêt vu le besoin de communiquer avec les machines en utilisant la parole spontanée. Ce domaine est riche d'applications potentielles allant de la transcription automatique des signaux de la parole à l'indexation des documents audiovisuels. Aussi, le traitement des différents dialectes mondiaux constitue une branche importante de la reconnaissance automatique de la parole et permet par la suite la généralisation des systèmes de dialogue homme-machines.

Cette thèse s'inscrit dans le cadre de traitement automatique de la parole qui se focalise particulièrement sur la Reconnaissance Automatique de la Parole (RAP) et la Vérification Automatique du Locuteur (VAL). En particulier, on s'intéresse aux deux dialectes marocains Darija et Tamazight qui constituent les " langues " les plus parlées au Maroc. L'objectif principal est de réaliser un système de sécurité composé de deux parties et exploitable au niveau national. La première phase de ce système consiste à vérifier des mots de passe en se basant sur la reconnaissance vocale. Néanmoins, cette approche n'est pas suffisamment robuste notamment pour l'accès aux données plus sensibles. Pour tenter de résoudre ce problème, une deuxième phase est nécessaire et consiste à ajouter une deuxième couche permettant l'identification automatique du locuteur. Dans le domaine de sécurité un tel système permet d'améliorer l'usage de l'identification à base des empreintes vocales.

Pour tester les performances de l'ensemble de ces systèmes, nous avons utilisé une base de données orale pour les deux dialectes. Celle-ci a été créée au sein de notre laboratoire (laboratoire de Traitement de l'Information et Aide à la Décision). L'évaluation des résultats obtenus a été réalisée en se basant sur le calcul du taux de reconnaissance. Les meilleurs résultats ont été obtenus par les méthodes basées sur les modèles de Markov cachés (MMC) et les réseaux de neurones (RN).

Mot clefs : Modèle de Markov Caché (MMC), MFCC, Tifinaghe, Dialectes marocains, Tamazight, Darija, Vérification du locuteur, Réseaux de neurones, Sécurité.

d'un son diffère en fonction de son emplacement dans le mot, un ***d*** dans le mot ***deux*** ne se prononce pas de la même façon que dans le mot ***date***. De même les sons des extrémités d'un mot peuvent subir des modifications très importantes en fonction du mot qui suit et du mot qui précède. Nous parlons dans ce cas des effets de coarticulation [5].

4. Caractéristique d'un système de reconnaissance automatique de la parole

4.1. Principe du système de reconnaissance

Un système de reconnaissance automatique de la parole (RAP) se base sur l'approche statistique fondée sur la théorie de l'information [6]. Un système de RAP a pour but d'associer une séquence de mots à une séquence d'observations acoustiques. Ainsi, à partir de la séquence d'observations acoustiques $X = x_1 x_2 \dots x_n$, un système de RAP recherche la séquence de mots $\widehat{W} = w_1 w_2 \dots w_k$ qui maximise la probabilité P(W|X), qui est la probabilité d'émission de W sachant X. La séquence de mots $\widehat{W}$ doit alors maximiser l'équation 1.1.

$$\widehat{W} = \underset{W}{argmax}\ P(W|X) \qquad \textbf{1.1}$$

En appliquant la règle de Bayes, on obtient la formule :

$$\widehat{W} = \underset{W}{argmax}\ \frac{P(X|W)P(W)}{P(X)} \qquad \textbf{1.2}$$

Comme la séquence d'observations acoustiques X est fixée, P(X) peut être considérée comme une valeur constante inutile dans l'équation 2.2. On obtient alors :

$$\widehat{W} = \widehat{argmax}_W (P(X|W)P(W) \qquad \textbf{1.3}$$

Deux types de modèles probabilistes sont utilisés pour la recherche de la séquence de mots la plus probable : un modèle acoustique qui fournit la valeur de P(X|W) et un modèle de langage qui fournit la valeur P(W). P(X|W) peut se concevoir comme la probabilité d'observer *X* lorsque W est prononcé, alors que P(W) se réfère à la probabilité que W soit prononcé dans un langage donné. Pour obtenir un système de RAP performant, il est essentiel de définir les modèles les plus pertinents possibles pour le calcul de P(W) et P(X|W). La figure 1.3 présente une schématisation du fonctionnement d'un système de RAP.

ABSTRACT

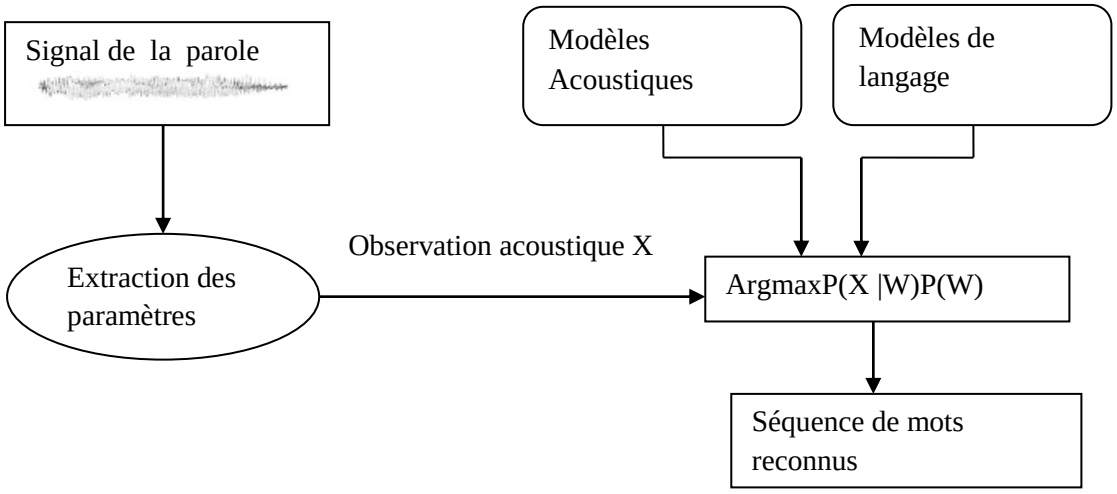
View the need to communicate with machine using spontaneous speech, Speech processing arouse currently a great interest. This field is rich of potential applications from automatic transcription of speech signals to indexing audiovisuals documents. Thus, the processing of a different worldwide dialects constitute an important branch for the generalization of humane-machines systems dialogue.

This thesis belongs to automatic speech processing that focuses particularly on automatic speech recognition (ASR) and speaker verification (SV); In particular, we are interested into Moroccan dialects Tamazight and Darija that constitute the most popular "languages" in Morocco. The principal object is to realize a security system that is composed from two parts and can be exploitable in national level. The first step consists on password verification based on automatic speech recognition. However, this approach is not sufficiently robust especially in acceding to sensitive data. To try to resolve this problem, a second step is necessary and consists to add a second layer that permits the speaker identification. In the field of security, such a system permits to ameliorate the use of the identification based on voice prints.

To test the performances of all these systems, we used an oral database for the two dialects. This one was created in our laboratory (Laboratory of modeling and calculation). The obtained results are evaluated on the basis of recognition rate calculation. The best results are obtained with a method based on the hidden Markov model (HMM) and neural network (NN).

Title : Automatic speech processing: application to Moroccan dialects

Keywords : Hidden Markov Model (HMM), MFCC, Tifinaghe, Darija, Moroccan dialects, Tamazight, Speaker verification, Neural network.



Figue 1.3 : fonctionnement d'un système RAP

4.2. Caractéristiques d'un système RAP

Un système de reconnaissance automatique de la parole est caractérisé par son mode de fonctionnement, le mode d'élocution, le type de vocabulaire et la syntaxe de langage.

4.2.1. Système dépendant ou indépendant de locuteur

Les systèmes de reconnaissance automatique de la parole peuvent être regroupés en trois classes suivant le nombre de locuteurs qui utilisent ce système :

- ✓ **Mono-locuteur :** le système de reconnaissance est adapté à un seul locuteur.
- ✓ **Multi-locuteur :** le système est utilisé par un groupe de personnes qui sont apprises par le système dans la phase d'apprentissage.
- ✓ **Indépendant du locuteur :** tout le monde peut utiliser le système.

4.2.2. Mode d'élocution

Les systèmes de reconnaissance diffèrent selon le mode d'élocution utilisé, on trouve trois types de systèmes :

- ✓ **Système de mots isolés :** chaque mot est prononcé isolement en marquant des pauses entre les mots, ce type de système ne prend pas en compte le contexte du mot.
- ✓ **Système de reconnaissance de mots connectés :** le système peut reconnaitre une suite de mots sans marquer des pauses entre les mots (exemple de reconnaissance des chiffres enchainés).
- ✓ **Système de reconnaissance de la parole continue :** la parole continue est le discours usuel, dans ce cas nous sommes amenés à introduire le modèle de langage, ce dernier permet la prise en compte du contexte du mot dans la phrase prononcée.

ملخص

نظرا للحاجة إلى الحوار مع الآلات باستعمال الكلام العفوي، تثير المعالجة التلقائية للكلام حاليا أهمية كبيرة ٠ هذا المجال غني بتطبيقات قطبية من النسخ التلقائي للكلام الى فهرسة الوثائق السمعية البصرية٠ايضا ،تشكل معالجة مختلف اللهجات العالمية شعبة مهمة لتعميم انظمة الحوار انسان آلة .

هذه الاطروحة تسجل في نطاق المعالجة التلقائية للكلام الذي ينطوي خصوصا على التعرف الآلي على الكلام والتحقق الآلي من المتكلم ٠على وجه التخصيص، نهتم باللهجتين المغربيتن الدرجة والأمازيغية اللاتي تشكل اللهجات الأكثر تكلما في المغرب٠ الهدف الرئيسي هو انشاء نظام امن مركب من طبقتين ومشغل على الصعيد الوطني ٠الجزء الأول من هذا النظام يرتكز على التحقق من كلمة المرور بالإعتماد على التعرف الآلي على الكلام ٠لكن، هذه المقاربة ليس قويا كفاية للإعتماد عليه خصوصا في الولوج الى معطيات جد حساسة٠ لمحاولة حل هذا المشكل، يجب اضافة مرحلة ثانية التي تعتمد على التحقق من المتكلم٠ في المجال الأمني مثل هذا النظام يسمح بتطوير استعمال كشف الهوية بالإعتماد على البصمات الصوتية٠

لإختبار قدرات مجموعة هذه الأنظمة ، استعملنا قاعدة بيانات شفوية بالنسبة للهجتين٠هذه القاعدة تم انشاؤها في المختبر٠ يتم تقييم النتائج المحصل عليها بالإعتماد على حساب معدل التعرف ٠ افضل النتائج تم الحصول عليها بواسطة الطرق التي تعتمد على نموذج ماركوف الخفي وشبكة نورون.

العنوان : المعالجة التلقائية للكلام : تطبيق على اللهجات المغربية

كلمات مفتاح : نموذج ماركوف الخفي، المعايير الصوتية ،تيفناغ،الامازيغية،الدرجة.

4.2.3. Vocabulaire

Le vocabulaire est l'ensemble de mots ou corpus que le système est capable de reconnaitre, il est caractérisé par :

- ✓ Sa taille qui peut varier de quelques mots à plusieurs dizaines de mots.
- ✓ Sa nature qui correspond aux types des mots choisis par exemple les mots qui sont phonétiquement proches.

4.2.4. Syntaxe du langage

La syntaxe spécifie les contraintes imposées sur la suite de mots prononcés. Le but de la syntaxe est de faciliter la tâche du système pendant la reconnaissance, en limitant le nombre de candidats, par exemple après la phrase trente et on est obligé de mettre le mot un.

5. Différentes étapes du Système de reconnaissance automatique de la parole

Le but de RAP est d'identifier à partir d'un signal vocale le message linguistique le plus semblable à ce signal, cette identification se fait selon trois grandes phases (figure 1.4) :

- ✓ Une phase d'analyse acoustique ou paramétrisation, son rôle est d'extraire du signal les informations pertinentes et d'éliminer la redondance.
- ✓ Une phase d'apprentissage, elle permet d'ajuster les paramètres acoustiques afin de stabiliser le modèle de chaque mot, cette phase s'adapte pour les modèles stochastiques par exemple le modèle de Markov caché (MMC).
- ✓ Une phase de reconnaissance, son rôle est la comparaison des informations issues de la première phase, aux données de référence.

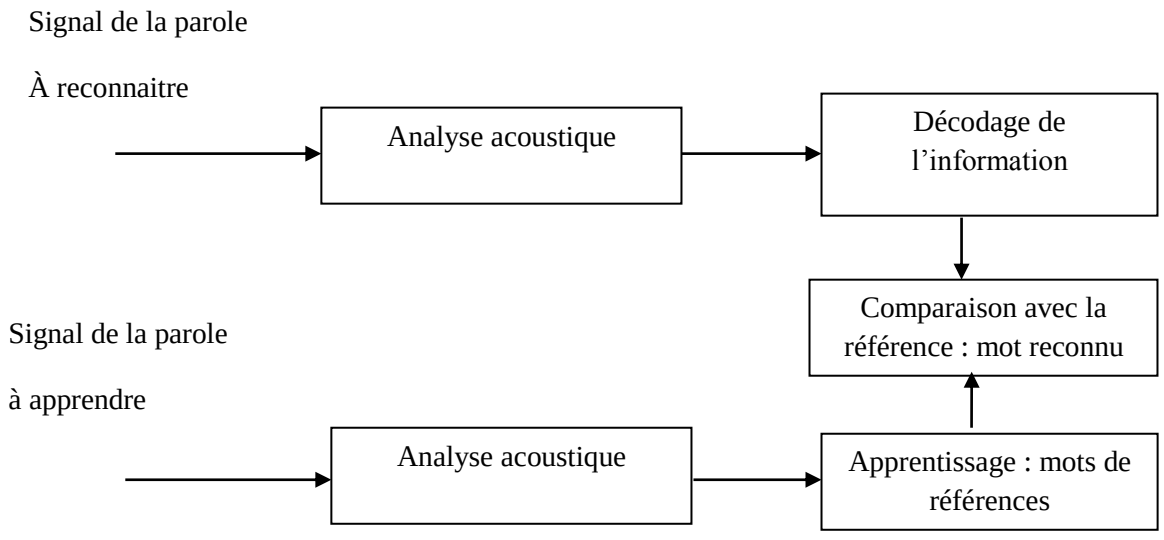


Figure 1.4 : les principales tâches d'un système de reconnaissance de la parole

ABRÉVIATIONS

MFCC	Mel Frequency Cepstral Coefficients
RAP	Reconnaissance Automatique de la Parole
SRAP	Système de reconnaissance Automatique de la Parole
MMC	Modèle de Markov Caché
GMM	Gaussians Mixture Model
EM	Expectation Maximisation
EMV	Estimation du Maximum de Vraisemblance
LPC	Linear Predictive Coding
DTW	Dynamic Time Warpping
LFCC	linear frequency cepstral coefficients
LFSC	Linear Frequency Spectral Coefficients (LFSC)
CFSM	Coefficients Fréquentiels Spectrales de Mel (MFCS)
QV	Quantification Vectorielle (VQ)
RAL	Reconnaissance Automatique du Locuteur
VAL	Vérification Automatique du Locuteur
IAL	Identification Automatique du Locuteur
PMC	Perceptron Multi Couches (MLP)
MMSO	Méthodes Statistiques du Seconde Ordre (SOSM)
HF	Hamming Framming
MV	Maximum de Vraisemblance (ML)
VA	Vecteurs Acoustiques (AV)

5.1. Analyse acoustique

En acoustique, un son se définit classiquement au moyen de son amplitude, de sa durée et de son timbre. Le traitement du signal vocal a pour but de quantifier ces trois grandeurs pour faire correspondre à l'onde sonore une description multidimensionnelle. En particulier, l'analyse acoustique du signal est utilisée pour résoudre le problème lié à la redondance du signal de parole et pour diminuer la quantité de calculs. Cette analyse permet de représenter le signal par des vecteurs de coefficients qui sont calculés sur des intervalles de temps.

5.1.1. Mise en forme du signal

Avant de commencer les calculs des coefficients acoustiques, il est nécessaire de faire le calcul préalable suivant :

- ✓ **Filtrage analogique en sortie du microphone :** notons que les informations acoustiques du signal se situent dans la bande fréquentielle [50Hz, 8Khz], le but principal de ce filtrage est d'éliminer toute information hors de cette bande passante.
- ✓ **Conversion analogique numérique :** afin d'utiliser ou de traiter les signaux continus, sortant d'un microphone ou d'un appareil électronique, par des calculateurs, il est nécessaire de numériser ou de discrétiser ce signal. Cette opération de discrétisation s'appelle l'échantillonnage du signal et l'opération inverse s'appelle l'interpolation (figure 1.5) [7]. Si on note x(t) un signal continu, l'échantillonnage de x(t) est l'application qui fait correspondre au signal x(t) un signal discret (x_1, x_2,…, x_n) avec :

$$x_n = x(t_n) \qquad 1.4$$

Lorsque $t_n - t_{n-1} = T$ est constante, pour tout n, on note par : $f_e = \frac{1}{T}$

t_n : l'échantillon numéro *n*.

T : le pas d'échantillonnage du signal x(t).

f_e : fréquence d'échantillonnage.

Pour avoir une perte d'information presque nulle entre le signal continu et le signal échantillonné suivant une fréquence d'échantillonnage f_e, il faut et il suffit que f_e soit au moins supérieur au double de la fréquence la plus élevée f_m de ce signal (théorème de Shannon) [7], c'est-à-dire :

$$f_m \leq \frac{f_e}{2} \qquad (f_m \text{ est déterminée à partir du signal}) \qquad 1.5$$

Si les enregistrements sont effectués à travers les lignes téléphoniques on a f_m=3.3Khz ce qui implique que : $f_e \geq 6.6Khz$

Si les enregistrements sont effectués dans le laboratoire dans ce cas f_m=8Khz, ce qui donne : $f_e \geq 16Khz$. Dans ce travail on fixe : $f_e = 16Khz$.

NOTATIONS

$p(a\|b)$	Probabilité de l'événement a sachant l'événement b
$p(a)$	Probabilité à priori de l'événement a
$(x_k)_{1\leq k\leq N}$	Suite d'observations de taille N.
N	Taille des vecteurs acoustiques, nombre de paramètres des coefficients utilisés.
m	Nombre d'états du modèle MMC.
$(a_{ij})_{1\leq i\leq m,1\leq j\leq m}$	Matrice de transition du modèle MMC.
$(b)_{1\leq i\leq m,1\leq j\leq m}$	Matrice de probabilités d'observations.
Λ	Modèle de Markov Caché défini par certains paramètres.
μ_i	Moyenne liée au vecteur acoustique x.
$\mathbf{C_i}$	Matrice de covariance liée à l'état i.
$\mathbf{Y_N}$	Suite de coefficients issue de paramétrisation du signal de la parole.
$\mathbf{fe}$	Fréquence d'échantillonnage.
$\mathbf{f_m}$	Fréquence maximale du signal.
T	Pas d'échantillonnage (période).
$\boldsymbol{\lambda}^*$	Modèle de Markov optimal (probabilité d'observation maximale).
$\mathbf{w}^*$	Graphe du mot reconnu selon le chemin optimal.
$\mathbf{TF}(\mathbf{x_n})$	Transformée de Fourier pour un vecteur de données x_n.
$(\boldsymbol{\pi_i})_{1\leq i\leq N}$	Vecteur de probabilités initiales.
$\mathbf{e_n}$	Énergie d'un échantillon de signal n.
$\mathbf{N}_{(\mu_k,\Sigma_k)}$	Gaussienne de moyenne μ_k et de matrice de covariance $\sum_k$.
$\boldsymbol{\lambda} = (\boldsymbol{\pi}, \mathbf{A}, \mathbf{B})$	Modèle de Markov défini par le vecteur de probabilité initial, la matrice de Transition A et la matrice de probabilités d'observations B.
$\mathbf{Pr}_{\boldsymbol{\lambda}}(\mathbf{Y_i})$	Probabilité d'observation d'un vecteur Y_i sachant un modèle λ.
$\boldsymbol{\alpha_k}$	Poids de sous gaussienne k.

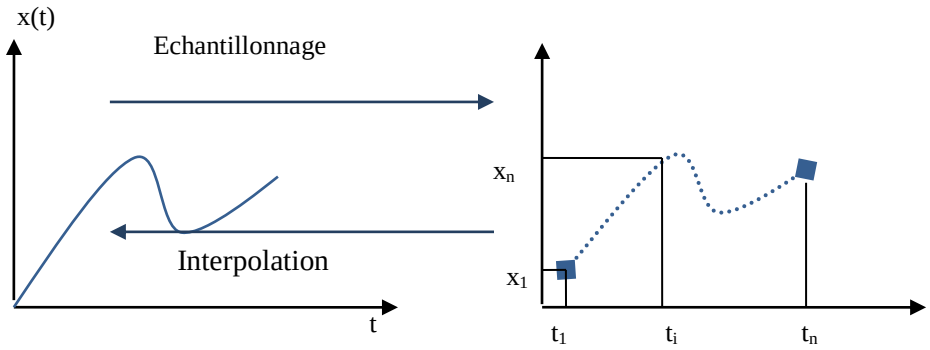


Figure 1.5 : l'échantillonnage et l'interpolation d'un signal de la parole

- ✓ **Préaccentuation du signal :** Cette opération est effectuée afin de relever les hautes fréquences qui sont moins énergétique que les basses fréquences. Si on note x_n le signal échantillonné, alors la préaccentuation de x_n est :

$$x_n = x_n - \alpha x_{n-1} \text{ avec } 0{,}9 \leq \propto \leq 1 \qquad \textbf{1.6}$$

- ✓ **Segmentation du signal :** La plupart des méthodes d'analyse acoustique utilisent l'hypothèse de la stationnarité du signal, ce qui n'est pas vrai pour le cas de la parole, une analyse sur des segments à court terme sur lesquels le signal est supposé quasi-stationnaire est nécessaire.

Deux types de segmentation sont utilisés, soit une segmentation de signal en trames de longueur variable et qui s'appuie sur un algorithme de segmentation automatique, qui isole les zones homogènes du signal, soit une segmentation du signal en trames de longueur fixe qui se recouvrent entre eux. La longueur de la trame varie entre 20ms et 40ms. Si cette longueur est égale à 32ms et f_e=16Khz, le nombre d'échantillons par trame est de 512 échantillons.

$$f_e = \frac{1}{T} \Leftrightarrow T = \frac{1}{f_e} = \frac{1}{1610^3} = 0.0625\ ms$$

$$nbre_{echantillon} = \frac{32}{0.0625} = 512$$

- ✓ **Fenêtrage du signal :**

Pour appliquer la segmentation du signal nous sommes obligés de multiplier le signal par une fenêtre rectangulaire *w(n)*. L'application de ce type de fenêtrage crée des oscillations importantes dans le domaine (fréquence spectre), de plus il produit une discontinuité aux frontières des trames. Pour réduire ces effets, on multiplie le signal par une fenêtre w(n). La transformée de Fourier de cette fenêtre s'approche d'une impulsion de Dirac [8]. Le nouveau signal devient : $x_n = x_n . w(n)$.

TABLE DES FIGURES

Il existe plusieurs types de fenêtrage, nous citons par exemple :

- ✓ Fenêtrage de Hamming [8] dont l'équation est : $w(n) = 0.54 - 0.46.\cos(\frac{2\pi.n}{N})$
- ✓ Fenêtrage de Hanning [9] défini par : $w(n) = 0.5(1 - \cos\left(\frac{2.\pi.n}{N}\right)))$
- ✓ Fenêtrage de Blackman [10] dont l'équation est :

$$w(n) = 0.42 - 0.5.\cos\left(\frac{2.\pi.n}{N}\right) + 0.08.\cos(\frac{4.\pi.n}{N})$$

Dans notre analyse nous avons choisi le fenêtrage de Hamming. Ce dernier est le plus utilisé en littérature du traitement du signal de la parole, il permet d'estimer le signal sur une proportion jugée stationnaire (10 à 20ms) en réduisant les effets de bord et la discontinuité du signal vocal.

Les étapes de la mise en forme du signal de la parole sont données dans la figure 1.6 :

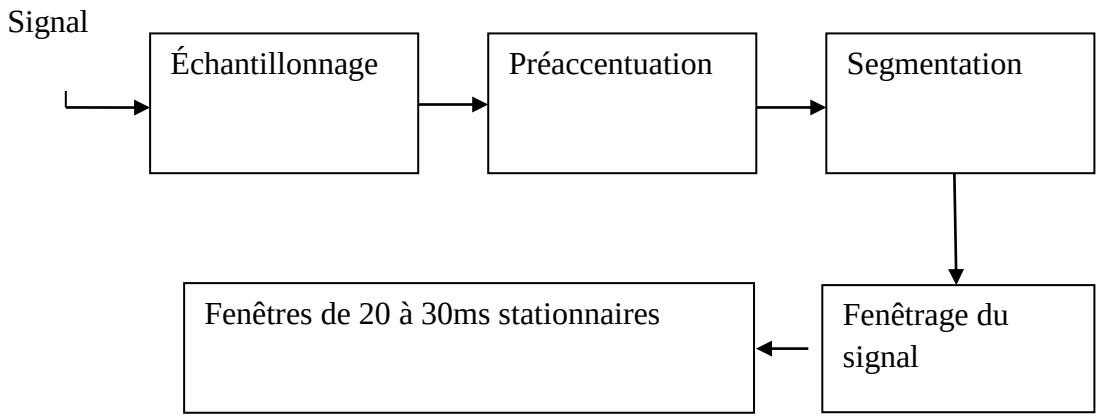


Figure 1.6 : la mise en forme du signal de la parole

5.1.2. Calcul des coefficients acoustiques

Une fois que le signal a subi ces transformations, les méthodes de calcul des coefficients le traitent par bloc. Parmi les méthodes les plus utilisées en analyse acoustique nous citons :

- ✓ La méthode d'analyse spectrale qui donne les coefficients MFCC (Mel Frequency Cepstral Coefficients) [11].
- ✓ La méthode de prédiction linéaire qui donne les coefficients LPCC (Linear Prediction Ceptsral Coefficients) [12].

5.1.2.1. Analyse spectrale (Coefficients MFCC)

Les coefficients MFCC sont les plus connus en reconnaissance automatique de la parole, ils sont introduits pour la première fois en 1980 par Davis et Mermelstein [13]. Pour calculer ces coefficients, nous appliquons sur le signal les transformations suivantes :

✓ Transformée de Fourier :

L'analyse spectrale présente l'intérêt de séparer la contribution de la source de celle du conduit vocal. Cette séparation est réalisée par un homomorphisme qui transforme le produit de convolution entre la source glottique g_n et la réponse impulsionnelle du conduit vocal h_n en une addition dans le domaine cepstral pour se ramener au cas linéaire [8, 10,11]. Si x_n est le signal de la parole alors :

$$x_n = g_n * h_n \qquad \textbf{1.7}$$

En appliquant la transformée de Fourier sur le produit de convolution nous obtenons :

$$X_w = TF(x_n) = TF(g_n) * TF(h_n) \qquad \textbf{1.8}$$

Avec :

$|X_w|$: le spectre d'énergie de x_n.

TF(x_n) est la transformée de Fourier de x_n. Pour réduire le taux de calcul, il est préférable d'utiliser l'algorithme FFT (Fast Fourier Transform: Transformée de Fourier Rapide) [13].

La FFT est appliquée au bloc $\{x_{i1}, x_{i2}, \dots, x_{in}\}$ pour obtenir le spectre $\{|X|_{i1}, |X|_{i2}, \dots, |X|_{in}\}$.

✓ Filtrage triangulaire :

Nous remarquons que la période fondamentale des sons voisés produit de nombreuses harmoniques sur le spectre obtenu par la FFT, pour diminuer ces phénomènes nous effectuons des lissages sur ces spectres en appliquant une suite de filtres triangulaires, répartis sur la bande passante [100Hz,7.5Khz] suivant une échelle de Bark ou de Mel [9,10] et cela pour se rapprocher de l'oreille humaine (Figure 1.7).

La relation de l'échelle de Bark [10] est : $B = 6Arcsinh(\frac{F}{600})$ (F: la fréquence)

La relation de l'échelle de Mel [9] est : $M = \frac{1000}{log2} Log(1 + \frac{F}{1000})$

Ensuite les bornes de ces filtres sont exprimées en échelle Mel.

On note par F_n et F_{n+1} les bornes inférieures des deux filtres triangulaires numéro n et n+1.

L'équation du $n^{ème}$ filtre est donnée par :

$$I_n = \begin{cases} \frac{f - F_n}{F_{n+1} - F_n} \; si \; F_n \leq f \leq F_{n+1} \\ 1 - \frac{f - F_n}{F_{n+1} - F_n} \; si \; F_{n+1} \leq f \leq F_{n+2} \end{cases} \qquad \textbf{1.9}$$

Nous appliquons ce filtrage sur le bloc $\{|X_{i1}|, \dots, |X_{in}|\}$ obtenu par la FFT et qui correspond à

la suite des fréquences $\{f_{i1}, \dots, f_{in}\}$. Ensuite l'énergie du $n^{ème}$ filtre est donnée par :

$$e_n = \sum_{j=1}^{N} |X_j|^2 I_n \qquad \textbf{1.10}$$

A la fin le bloc de signal *i*, la suite $\{|X_{i1}|, \dots, |X_{in}|\}$ va être représenté seulement par les F énergies $\{e_1, \dots., e_F\}$, F est le nombre de filtres triangulaires.

✓ Transformée de Fourier inverse ou transformée en cosinus discret

Nous appliquons ensuite la transformée de Fourier inverse ou la transformée en cosinus discret sur le bloc$\{e_1, \dots., e_F\}$ et nous ne prenons que la partie réelle de cette transformation. La formule de calcul de cette transformée donne les coefficients spectraux C_k et qui sont notés MFCC.

$$C_k = \sum_{i=1}^{F} Log(e_i) \cos\left(\frac{\pi k(i-0.5)}{F}\right) \; k = 1, \dots, d \qquad \textbf{1.11}$$

d : le nombre de coefficients cepstraux.

Une dizaine de ces coefficients est généralement jugée suffisante pour présenter une trame du signal. Dans notre travail nous avons fixé d à 13.

Chaque trame du signal est représentée par un vecteur de coefficients :

$$X = (C_1, \dots, C_d)$$

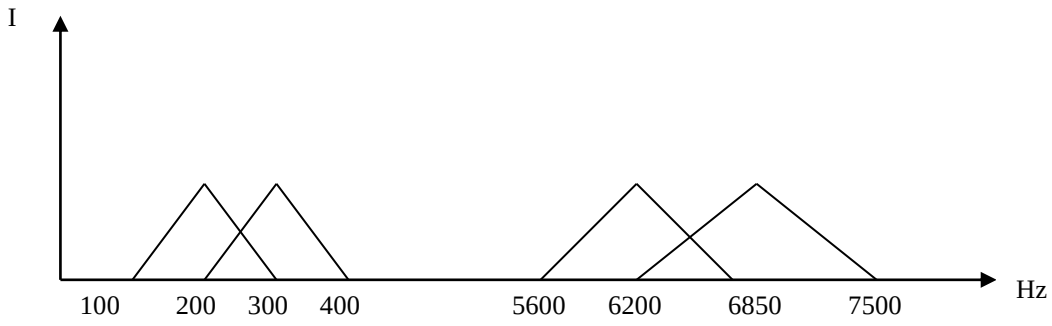


Figure 1.7 : répartition des filtres triangulaires

Les étapes de paramétrisation du signal sont données sur la figure 1.8.

FIGURES DES TABLES

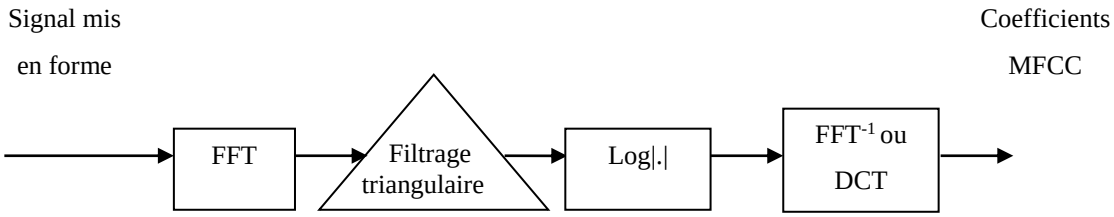


Figure 1.8 : calcul des coefficients MFCC

5.1.2.2. Analyse par la prédiction linéaire

Après l'étape de la mise en forme du signal, on calcule les coefficients spectraux du signal de la parole, cette analyse suppose que le signal vocal comme un signal autorégressif décrit par le modèle :

$$x_n - \sum_{i=1}^{T} a_i x_{n-1} = e_n \; n = 1, \dots, T$$

e_n est un bruit blanc gaussien de variance σ^2.

Pour calculer les a_i et σ^2 on est amené à résoudre le problème de minimisation donné par la méthode des moindres carrées :

$$\min_{a_1,\dots,a_p} \sum_{n=1}^{T} (x_n - \sum_{i=1}^{p} a_i x_{n-i})^2 \qquad \textbf{1.12}$$

Après les calculs on trouve :

$$\begin{cases} \sum_{i=0}^{p} a_i R_{ij} = 0 \\ \sum_{i=0}^{p} -a_i R_{i0} = \sigma^2 \end{cases} \qquad \textbf{1.13}$$

Où $a_0 = -1$ et $R_{ij} = \sum_{n=1}^{T} (x_{in-1} . x_{in-j})$.

La démonstration de ces formules est donnée en annexe 1.

Après le calcul des coefficients a_i, on calcule les coefficients cepstraux C_k, qui sont nommés LPCC (Linear Prediction Cepstral Coefficient). Pour k = 1,…, d on a :

$$C_k = a_k - \sum_{i=1}^{k-1} \frac{i}{k} . a_{k-1} . C_i \qquad \textbf{1.14}$$

5.1.2.3. Décodage de l'information acoustique

Pour décoder les informations issues de l'analyse acoustique (MFCC), nous utilisons l'une des trois approches suivantes : l'approche analytique, l'approche globale ou l'approche statistique.

✓ Approche analytique :

Cette approche est composée par les étapes suivantes [14] :

- ✓ Segmentation du signal obtenu par l'analyse acoustique, en unités de taille phonétique (phonème ou syllabe,…). Cette segmentation est faite à l'aide de différents critères (énergie, stabilité,…).
- ✓ Identification phonétique des segments par comparaison avec des formes de références.
- ✓ Exploitation de la suite phonétique identifiée par certains analyseurs (lexical, syntaxique) pour déterminer le mot ou la phrase prononcée.

Cette approche est restée toujours au stade expérimental, à cause de sa faiblesse qui provient du processus de décision trop précoce (identification au préalable sans prise en compte des niveaux linguistiques).

✓ Approche globale :

Cette approche permet de prendre un mot ou une phrase comme des entités élémentaires en effectuant des comparaisons avec des références déjà enregistrées. Cette approche pourra être réalisée en utilisant le principe de déformation temporelle dynamique (Dynamic Time Warping DTW), cette déformation utilise le principe d'optimalité de Bellman [15].

Si on note $V = \{w_1, \dots, w_l\}$ le corpus vocabulaire du système, où chacun des mots w_i est représenté par une ou plusieurs formes acoustiques de références notée R_{wi} (par exemple les paramètres spectraux de l'analyse acoustique et qui sont calculés de manière périodique).

Notons O_w la suite de formes associée au mot à reconnaitre, l'identification du mot w est réalisée suivant le critère :

$$w^* = \arg\min_{w_i \in V}(D(O_w, R_{wi})) \qquad \textbf{1.15}$$

Avec D la distance euclidienne définie par la formule 1.19.

Le problème de calcul de cette distance est que la durée du mot w_i est différente de celle de w. La solution est de calculer cette distance par un alignement temporel qui rapproche le mieux les deux formes R_{wi} et O_w. La construction de cet alignement est réalisée récursivement sur l'indice temporel en exploitant le fait que le chemin optimal est l'extension d'un chemin partiel optimal. Cette approche a donné des meilleurs résultats pour les systèmes mono locuteur à petit vocabulaire et en mots isolés.

✓ Approche statistique :

En 1976, Jelinek a proposé une formalisation statistique simple issue de la théorie de l'information et qui consiste à décomposer le problème de la reconnaissance automatique de la parole [16]. Etant donnée une suite d'observations $Y_1, \dots, Y_T$ associée à une suite de mots prononcés w, l'approche statistique consiste à trouver la suite de mots w^* la plus probable connaissant la suite d'observations $Y_1, \dots, Y_T$.

LISTE DES ALGORITHMES

$$w^* = argmax_{W \epsilon V} \Pr(w/Y_1, \dots, Y_T) \qquad \textbf{1.16}$$

La règle de Bayes nous donne :

$$\Pr(w|Y_1, \dots, Y_T) = \frac{\Pr(Y_1,\dots,Y_T|w).\Pr(w)}{\Pr(Y_1,\dots,Y_T)} \qquad \textbf{1.17}$$

- ✓ $\Pr(Y_1, \dots, Y_T|w)$ représente la probabilité d'observer la suite $Y_1, \dots, Y_T$ sachant que la suite de mots prononcés w, cette probabilité est estimée par la modélisation acoustique.
- ✓ $\Pr(w)$ la probabilité a priori que la suite de mots w soit prononcée. Elle est estimée par un modèle de langage. Dans le cas de la reconnaissance de mots isolés on suppose que tous les mots ont la même probabilité d'être prononcés.

Puisque $\Pr(Y_1, \dots, Y_T)$ ne dépend pas de w, l'équation 2.16 devient :

$$w^* = argmax_{W \epsilon V} \Pr(Y_1, \dots, Y_T/w).P(w) \qquad \textbf{1.18}$$

L'approche statistique permet aussi d'intégrer les niveaux acoustiques et linguistiques dans un seul processus de décision (ce qui est impossible dans l'approche analytique). Les unités acoustiques modélisées peuvent être des mots comme dans le cas de l'approche globale, comme elles peuvent être des unités plus courtes comme les phonèmes (le cas de l'approche analytique).

6. Modèles utilisés en reconnaissance automatique de la parole

6.1. Programmation dynamique

Dans la reconnaissance vocale, il est impossible de comparer deux spectres (ou cepstres) directement, tout simplement parce qu'une même personne ne peut prononcer deux fois le même mot sur la même durée, le même rythme, la même intensité. Il est donc nécessaire de développer une méthode de comparaison. Il en existe plusieurs, on se contentera d'une comparaison spectre à spectre par l'algorithme de comparaison dynamique détaillé ci-après.

✓ L'algorithme de déformation temporelle dynamique DTW

Cet algorithme permet la comparaison dynamique de deux spectres. A cause des variations inévitables entre deux prononciations du même mot, on ne peut pas comparer directement ces deux signaux car ils n'ont pas la même durée. C'est pour cela que la comparaison dynamique a été développée (DTW : Dynamic Time Warping), aussi appelée normalisation temporelle [14]. Cet algorithme est une forme particulière de l'algorithme de programmation dynamique [15]. On peut distinguer deux sources de variation de l'échelle temporelle : la variation de la vitesse de prononciation et la variation du rythme de prononciation [14].

L'algorithme décrit ici est celui de Vintsjuk proposé en 1968 [17].

Soient A et B deux images acoustiques de longueur I et J respectivement. La distance entre l'évènement $i \in [1, I]$ de A et l'évènement $j \in [1, J]$ de B se calcule avec une simple distance euclidienne :

$$d(i,j) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2} \text{ Avec } i = \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_N \end{bmatrix} \text{ et } j = \begin{bmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_N \end{bmatrix} \quad \textbf{1.19}$$

Cela suppose bien sûr que l'on considère la même plage de fréquence pour les deux signaux (entre 0 et N) [18]. On crée donc un chemin $\{C(k) = (n(k), m(k)), k\epsilon[1, k]\}$. Il est nécessaire que les fonctions n(k) et m(k) soient croissantes et doivent satisfaire certaines contraintes : les seuls chemins valides arrivant au point (i , j) sont ceux provenant des points (i-1, j), (i , j-1) et (i-1, j-1). De plus on prendra k tel que $C(k) = (I, J)$. On pose $C(1) = (1,1)$ [19].

La méthode consiste à choisir le chemin qui passe par les distances $d(i,j)$ les plus petites, de sorte que la distance cumulée le long de ce chemin soit la plus petite possible.

On définit $g(i,j)$ la distance cumulée au point (i , j) comme [18] :

$$g(i,j) = min \begin{cases} g(i-1,j) + d(i,j) \\ g(i-1,j-1) + 2.d(i,j) \\ g(i,j-1) + d(i,j) \end{cases} \quad \textbf{1.20}$$

On remplit ensuite la matrice I*J (le plan du chemin) avec en $i^{ème}$ et $j^{ème}$ colonnes le résultat de g(i , j).

Enfin, on définit la distance normalisée entre les deux prononciations du mot :

$$G = \frac{g(I,J)}{I+J} \quad \textbf{1.21}$$

On obtient bien une distance entre deux spectres. On effectue ce travail entre le mot à reconnaitre et tous les mots du dictionnaire. On prend ensuite le mot du dictionnaire qui a la plus petite distance spectrale avec le mot à reconnaitre (figure 1.9 et figure 1.10).

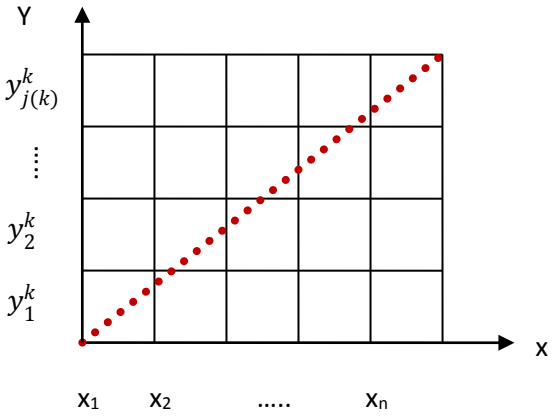


Figure 1.9 : chemin idéal entre deux spectres

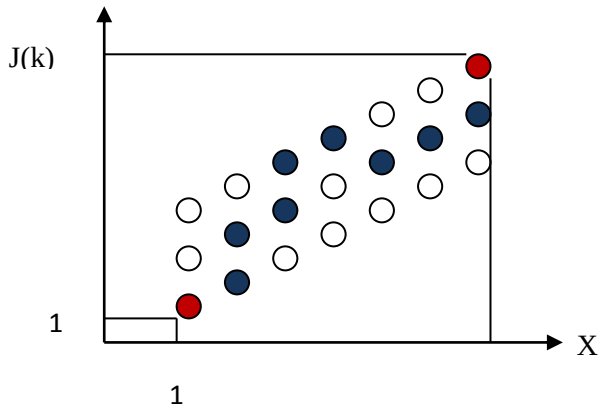


Figure 1.10 : Représentation du chemin entre deux spectres. Les différences entre deux chemins tordent le chemin idéal (diagonal)

6.2. Modèle connexionniste

La première tentative de définition et d'utilisation des modèles connexionnistes remonte aux années 40-50, ces modèles sont basés sur les réseaux de neurones. L'absence des calculateurs performants au niveau de la vitesse et de la capacité mémoire a conduit les chercheurs à la fin des années 60 à abandonner ces modèles. Mais nous remarquons ces dix dernières années que ces modèles ont commencé à avoir plus d'intérêt dans le domaine de la reconnaissance automatique de la parole.

Actuellement ces modèles sont utilisés dans distinctes étapes du processus de reconnaissance de la parole comme : l'analyse acoustique du signal (paramétrisation), décodage acoustico-phonétique, reconnaissance de mots isolés.

Il existe plusieurs types de réseaux de neurones utilisés en reconnaissance automatique de la parole [20,21], nous citons à titre d'exemple :

- ✓ Les réseaux de neurones multicouches perceptrons (Multi-Layer Perceptron MLP), plusieurs travaux ont montré l'intérêt de ce type de réseaux dans la reconnaissance automatique de la parole [22,23,24].
- ✓ Les réseaux de neurones récurrents (Recurrent Neural Network). Il existe des auteurs qui ont montré que ce type de réseaux peut être efficace dans la RAP [25].
- ✓ Les réseaux de neurones à délais temporels (Time Delay Neural Network TDNN) [26].
- ✓ Les réseaux prédictifs (Predective Neural Network PNP) [27], ces réseaux sont très peu utilisés en RAP.

6.3. Modèle de Markov Caché.

Les systèmes les plus utilisés en reconnaissance automatique de la parole ces dernières années sont basés sur les modèles de Markov Cachés MMC [28]. Une présentation détaillée de ces modèles avec les techniques de calcul utilisées dans ce cadre sera l'objectif du paragraphe 7. Nous donnerons aussi dans ce paragraphe un aperçu des systèmes de reconnaissance de la parole utilisant ce modèle.

6.4. Modèle mixte ou modèle hybride

Au cours de ces dernières années, une nouvelle modélisation en parole a vu le jour. Cette modélisation est connue sous le nom du modèle mixte, appelé aussi modèle hybride. Ces modèles combinent les modèles probabilistes et les modèles connexionnistes. Ces modèles tentent de diminuer le nombre d'hypothèses nécessaires à l'utilisation des modèles MMC (pour plus d'informations sur ces modèles voir [28]).

7. Modèle de Markov Caché

Depuis leur utilisation en traitement de la parole en 1975 par les chercheurs de CMU (Carnergie Mellon University) et d'IBM [9,29], les MMC sont devenus la base de la majorité des systèmes de traitement de la parole. Nous citons par exemple : le système SPHINX de CMU [30], BYBLOS de BBN (Bolt Beranek and Newman) [31], AT & T [32], CENT [16].

De nombreuses présentations théoriques des MMC existent dans la littérature. Pour une présentation détaillée, on peut consulter par exemple [33,34].

7.1. Définition du MMC

Un modèle de Markov caché est un double processus stochastique $(X_t, Y_t)_{t\geq 1}$ avec X_t est une chaine de Markov d'ordre 1 à valeur dans un ensemble d'états fini $Q = \{q_1, \dots, q_N\}$, X_t vérifie la propriété de Markov :

$$Pr\left(x_{t+1} = q_j|_{X_1=q_{i1},\dots,X_t=q_{in}}\right) = Pr\left(X_{t+1} = q_j|_{X_t=q_{in}}\right) = a_{ij} \qquad \textbf{1.22}$$

Et le vecteur de probabilité initial $(\Pi_i)_{i\epsilon E}$:

$$\Pi_i = P(X_1 = q_i)_{i\epsilon E} \qquad \textbf{1.23}$$

Y_t est un processus observable à valeurs dans un ensemble mesurable Y, Y_t vérifie :

$$\Pr(Y_t = y_t|X_t = q_i, \dots, , X_1 = q_1, Y_{t-1} = y_{t-1}, \dots, Y_1 = y_1) = \Pr(Y_t = y_t|X_t = q_i)$$

$$= b_i(y_t) \qquad \textbf{1.24}$$

$b_i(y_t)$ est la probabilité d'émission de l'observation y_t à partir de l'état q_i à l'instant *t*.

On peut encore associer les lois de probabilité des observations acoustiques aux transitions et on note par $b_i(y_t)$ la loi de probabilité des observations associées à la transition

$q_i \rightarrow q_j$:

$$b_{ij}(y_t) = \Pr(Y_t = y_t | X_t = q_j, X_{t-1} = q_i)$$

L'état q_i du processus X_t n'est pas directement observable, on dit qu'il est caché, mais le processus X_t émet après chaque changement d'état une observation y_t qui est une réalisation du processus Y_t.

Les observations acoustiques sont supposées indépendantes les unes des autres conditionnellement à la suite d'états. Ces observations peuvent être discrètes ou continues (figure 1.11).

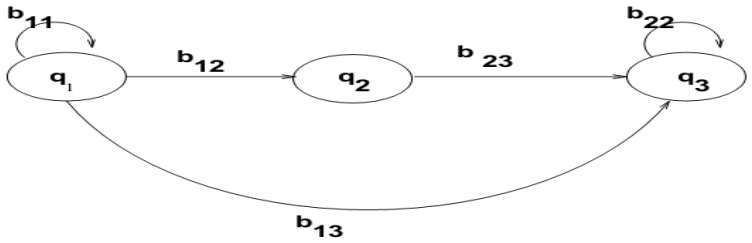


Figure 1.11 : Exemple de MMC à trois états

7.1.1. Observations discrètes

Les vecteurs de coefficients calculés sur les trames du signal sont des points d'un espace E multidimensionnel continu. Il est possible de représenter ce dernier par un espace discret au moyen de la quantification vectorielle (QV), en partitionnant l'espace E en classes et en choisissant ensuite pour chaque classe un représentant. L'ensemble de ces représentants constituent ce qu'on appelle un dictionnaire noté $D = \{d_1, \dots, d_n\}$, d_i est un vecteur acoustique.

Chaque vecteur $x \in E$ est quantifié par un élément d_i de D qui est le plus proche au sens d'une distance définie dans l'espace E et qui vérifie :

$$i = \arg\left(\min_{1 \le k \le n_D} d(x, d_k)\right) \qquad \textbf{1.25}$$

Plusieurs algorithmes ont été proposés pour calculer le dictionnaire D. Ces algorithmes sont basés sur la classification hiérarchique descendante ou sur les nuées dynamiques [35] (figure 1.12).

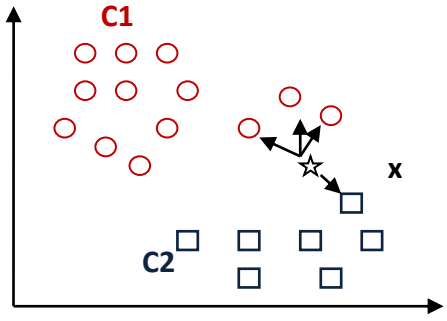


Figure 1.12 : Quantification vectorielle, dictionnaire à deux centroides C1 et C2

7.1.2. Observations continues

Dans le cas où l'espace E est supposé continu, on prend les probabilités d'émission des observations $y_1, \dots, y_T$ comme des densités de probabilités continues [6,10,13]. Pour modéliser la probabilité d'émission de ces observations, on utilise souvent un mélange de gaussiennes :

$$b_j(y_t) = \sum_{k=1}^{G} \alpha_{jk} . N_{(\mu_{jk}, \Sigma_{jk})}(y_t) \qquad \textbf{1.26}$$

- ✓ $b_j(y_t)$ Probabilité de l'émission de l'observation y_t à partir de l'état q_j.
- ✓ μ_k et $\sum_k$ sont respectivement, le vecteur des moyennes et la matrice de covariance de la loi normale $N_{(\mu_k, \Sigma_k)}$.
- ✓ α_k la pondération affectée à $N_{(\mu_k, \Sigma_k)}$, elle reflète l'importance qui peut être accordée à cette loi.
- ✓ G le nombre de gaussiennes.

On suppose souvent que les composantes du vecteur aléatoire Y_t sont indépendantes, ce qui diminue le nombre de paramètres à estimer (la matrice $\sum_k$ est diagonale).

Dans notre étude, on a utilisé quatre composantes gaussiennes.

Le modèle de Markov caché [10] est un modèle caractérisé par :

- ✓ L'ensemble des états :

$$Q = \{q_1, \dots, q_N\}$$

- ✓ La matrice des probabilités de transitions entre les états q_i, elle est notée :

$$A = (a_{ij})_{1 \le i,j \le N}$$

- ✓ Le vecteur des probabilités initiales :

$$\pi = (\pi_i)_{i \le i \le N}$$

- ✓ Le vecteur des lois de probabilité des observations :

$$B = (b_i(.))_{1 \le i \le N}$$

Dans la suite on note le modèle MMC par $\lambda = (\pi, A, B)$.

7.2. Vraisemblance à partir d'un MMC

L'utilisation des modèles MMC en reconnaissance automatique de la parole s'effectue suivant deux étapes successives :

- ✓ L'apprentissage du modèle MMC pour obtenir un réseau ou un modèle optimal λ^*.

- ✓ La reconnaissance des mots ou la recherche d'un chemin optimal $£^*$ dans le réseau optimal.

Avant de voir en détail ces deux étapes, nous présentons quelques définitions.

7.2.1. Probabilité d'émission d'une suite d'observations le long d'un chemin

Soit $\{y_1, \dots, y_T\}$ une suite d'observations associées à une prononciation inconnue (suite de vecteurs acoustiques d'un mot prononcé par une personne donnée). Le problème qui se pose en reconnaissance automatique de la parole, est de reconnaitre le mot prononcé sachant cette suite d'observations.

La solution de ce problème est équivalente à chercher un chemin dans le réseau, sur lequel on peut émettre ces observations avec une probabilité maximale.

Soit $£ = q_{i_1}, \dots, q_{i_T}$ un chemin de longueur T et $\Pr(y_1, \dots, y_T, q_{i1}, \dots, q_{iT}|\lambda)$ la probabilité de la suite d'observations $(y_1, \dots, y_T)$ suivant le chemin £ sachant le modèle λ. Cette probabilité est donnée par la formule :

$$\Pr(y_1, \dots, y_T, q_{i1}, \dots, q_{iT}|\lambda) = \pi_{i1} b_{i1}(y_1) \prod_{t=2}^{T} a_{i(t-1)it} b_{it}(y_t) \qquad \textbf{1.27}$$

7.2.2. Probabilité d'émission d'une suite d'observations

La probabilité d'émission des observations $y_1, \dots, y_T$ pour un modèle λ est la somme des probabilités d'émission des observations $y_1, \dots, y_T$ selon tous les chemins possibles £ de longueur T :

$$\Pr(y_1, \dots, y_T|\lambda) = \sum_{£} \Pr(y_1, \dots, y_T, £|\lambda) \qquad \textbf{1.28}$$

- ✓ **Méthode de calcul de cette probabilité**

Notons par $\alpha(t, q_j)$ la variable avant (Forward) qui représente la probabilité d'observer $y_1, \dots, y_t$ et d'aboutir à l'état q_j à l'instant t, cette variable est donnée par :

$$\alpha(t, q_j) = \Pr(y_1, \dots, y_t, x_t = q_j|\lambda) \qquad \textbf{1.29}$$

La variable arrière (Backward) est la probabilité d'observer $y_{t+1}, \dots, y_T$ sachant que l'on part de l'état q_j à l'instant t, elle est définie par :

$$\beta(t, q_j) = \Pr(y_{t+1}, \dots, y_T|x_t = q_j, \lambda) \qquad \textbf{1.30}$$

En développant ces formules, nous obtenons les relations récurrentes suivantes [6] :

$$\alpha(t+1, q_j) = \sum_{q_i} \alpha(t, q_i) a_{ij} b_j(y_{t+1}) \qquad \textbf{1.31}$$

$\beta(t+1, q_j) = \sum_{q_i} \beta(t+2, q_i) a_{ij} b_j(y_{t+2})$Ces deux grandeurs sont initialisées par :

$$\begin{cases} \alpha(0, q_i) = b_i(y_1).\pi_i \\ \beta(T, q_i) = \text{probabilité que l}'\text{état } q_i \text{ soit atteint à l}'\text{instant T.} \end{cases}$$

Si on a un réseau avec un seul état initial noté q_I alors :

$$\alpha(0, q_i) = \begin{cases} b_i(y_1) \text{ si } q_i = q_I \\ 0 \text{ sinon.} \end{cases}$$

Si on a un réseau avec un seul état final noté q_F alors :

$$\beta(T, q_i) \begin{cases} 1 \; si \; q_i = q_F \\ 0 \; sinon \end{cases}$$

Dans ce cas la probabilité d'émission de la suite d'observation $y_1, \dots, y_T$ est :

$$\Pr(y_1, \dots, y_T) = \alpha(T, q_F) = \beta(0, q_I) \qquad \textbf{1.32}$$

Les détails de cet algorithme sont donnés en Annexe 2.

7.3. Procédure d'apprentissage

L'apprentissage est une opération nécessaire pour tout système de reconnaissance, il permet de déterminer les paramètres des modèles acoustiques associés à chaque unité phonétique. Un apprentissage incorrect ou insuffisant diminue la performance du système RAP. La procédure d'apprentissage s'effectue en deux étapes :

7.3.1. Première étape : Construction du réseau markovien

La première étape dans la construction du réseau est le choix de l'unité linguistique à modéliser (phonèmes, cyllabs,…), ce choix dépend du degré de précision désiré dans la modélisation et le type de l'application envisagée. Elle est appelée unité de base ou unité élémentaire, elle peut être un mot ou une unité de longueur inférieur au mot. Parmi les unités les plus utilisées on trouve les phonèmes qui sont la plus petit unité modélisant un mot, les diphones, les pseudo-diphones, les syllabes, … (voir l'annexe 3).

Ensuite on passe à la construction du réseau global qui se fait d'une manière hiérarchique. Le réseau global est obtenu en compilant l'ensemble de tous les modèles (Figure 2.13).

7.3.2. Deuxième étape : Apprentissage

Association à chaque modèle markovien des informations nécessaires apportées par l'ensemble d'apprentissage. Cette phase se déroule en deux étapes :

- ✓ Initialisation des paramètres du modèle.
- ✓ Ré-estimation de ces paramètres.

7.3.2.1. Initialisation des paramètres :

Une convergence rapide des méthodes d'estimation des paramètres est assurée par une bonne initialisation de ces paramètres. Pour l'initialisation des probabilités de transition, on suppose souvent équiprobable toutes les transitions partant d'un état. Pour la moyenne et la matrice de covariance de la loi d'observation acoustique, on peut utiliser par exemple les résultats des statistiques effectuées à partir d'une base de données étiquetée manuellement [36], ou utiliser la procédure itérative des k-means [34].

7.3.2.2. Ré-estimation des paramètres :

Après l'initialisation des paramètres du réseau, on effectue une ré-estimation de ces paramètres jusqu'à obtenir le réseau optimal. Plusieurs critères d'estimation sont utilisés on cite par exemple :

- ✓ L'estimation par maximum de vraisemblance (Maximum Likelihood Estimation MLE), elle est réalisée par l'algorithme de Baum-Welch ou Viterbi [37].
- ✓ L'estimation par maximum *a postériori* [37,38].
- ✓ L'estimation par maximum d'information mutuel [39].

7.3.3. Estimation par Maximum de Vraisemblance

L'estimation par maximum de vraisemblance est l'une des méthodes les plus utilisées dans l'apprentissage du modèle MMC. Cette méthode utilise la procédure de Baum-Welch, ou l'algorithme de Viterbi [38].

7.3.3.1. Méthode de Baum-Welch

L'algorithme de Baum-Welch permet une ré-estimation itérative des paramètres du modèle [40]. Si on prend un ensemble d'apprentissage $W = \{w_1, \dots, w_R\}$ constitué de R prononciations où à chaque prononciation w_i est associée une suite d'observations $Y_i = (y_1^i, y_2^i, \dots, y_{T_i}^i)$, la fonction de vraisemblance au point $(Y_1, \dots, Y_R)$ est :

$$L(Y_1, \dots, Y_R) = \prod_{i=1}^{R} Pr_\lambda(Y_i) \qquad \textbf{1.33}$$

$Pr_\lambda(Y_i)$ est la loi de probabilité associée à Y_i.

L'estimateur de maximum de vraisemblance λ^* est donné par la formule :

$$\lambda^* = \arg\max_{\lambda}(L(Y_1, \dots, Y_R)) \qquad \textbf{1.34}$$

Ce qui revient à :

$$\lambda^* = \arg\max_{\lambda} \prod_{i=1}^{R} Pr_\lambda(Y_i) \qquad \textbf{1.35}$$

Pour calculer ce maximum, on utilise la fonction auxiliaire Q(.,.) définie par [34] :

$$Q(\lambda_n, \lambda_{n+1}) = \sum_{i=1}^{R} \sum_{£} Pr_{\lambda_n}(£|Y_i) \ln(Pr_{\lambda_{n+1}}(Y_i, £))$$

£ est un chemin de réseau (figure 1.13).

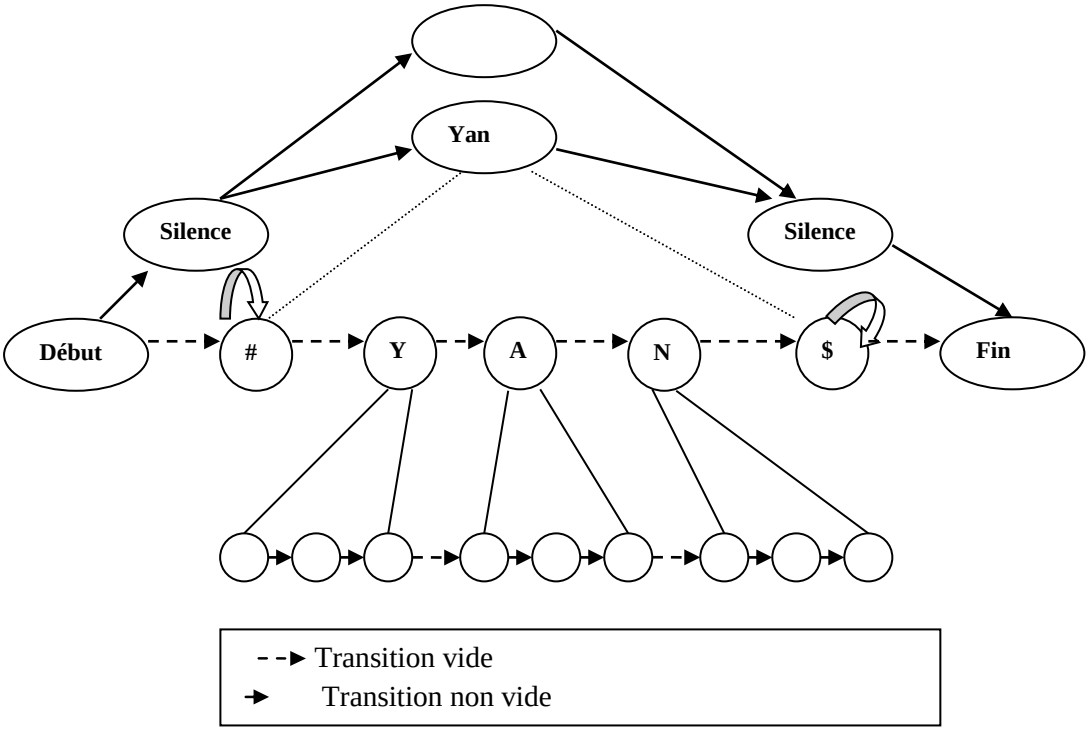


Figure 1.13 : Construction de réseau globale pour un système de reconnaissance de mots isolés avec unité de base syllabes

Le réseau sur la figure 2.13 est construit à partir des modèles des mots et par la suite modèle de phonème. Pour calculer la probabilité d'observation d'une séquence, on commence par l'état initial et on suit le chemin qui maximise cette probabilité.

Baum [41] a montré que si λ_{n+1} est un maximum de la fonction $Q(\lambda_n, .)$, alors λ_{n+1} est une estimation meilleure que λ_n car :

$$\prod_{i=1}^{R} Pr_{\lambda_{n+1}}(Y_i) \geq \prod_{i=1}^{R} Pr_{\lambda_n}(Y_i)$$

Avec cette idée on définit un processus itératif, tel qu'à chaque étape on obtient une estimation du modèle meilleur que la précédente [42,43].

La convergence de $\prod_{i=1}^{R} Pr_{\lambda_n}(Y_i)$ est locale, il faut donc bien choisir les valeurs initiales des paramètres des modèles acoustiques pour assurer une convergence correcte et rapide.

L'avantage de cette fonction auxiliaire, est qu'elle est décomposable en une somme de trois fonctions indépendantes et elle permet de faciliter l'estimation de la probabilité d'observation des vecteurs acoustiques :$Q(\lambda, \lambda') = Q_{\pi'}(\lambda, \lambda') + Q_{A'}(\lambda, \lambda') + Q_{B'}(\lambda, \lambda')$

Avec

$$Q_{\pi'}(\lambda,\lambda') = \sum_{i=1}^{R}\sum_{£} Pr_{\lambda}(£|Y_i)\ln(\pi'_{i1})$$

$$Q_{A'}(\lambda,\lambda') = \sum_{i=1}^{R}\sum_{£} Pr_{\lambda}(£|Y_i)\ln(a'_{ij})$$

$$Q_{B}(\lambda,\lambda') = \sum_{i=1}^{R}\sum_{£} Pr_{\lambda}(£|Y_i)\ln(\pi b'_{i}(y_i)) \qquad \textbf{1.36}$$

7.3.3.2. Méthode de Viterbi

Cette méthode consiste à maximiser la vraisemblance des observations conjointement aux chemins optimaux [4, 6].

Si on note par λ^* l'estimateur de λ on a :

$$\lambda^* = \underset{\lambda}{argmax}\prod_{i=1}^{R} Pr_{\lambda}(Y_i, £_i^*) \qquad \textbf{1.37}$$

$£_i^*$ est le chemin optimal associé à la suite d'observations Y_i.

Dans ce cas la fonction auxiliaire s'écrit comme suit :

$$Q(\lambda_n,\lambda_{n+1}) = \sum_{i=1}^{R}\sum_{£} \delta(£ - £_i^*) Pr_{\lambda_n}(£|Y_i)\ln\left(Pr_{\lambda_{n+1}}(Y_i,£)\right)$$

$$= \sum_{i=1}^{R} Pr_{\lambda_n}(£_i^*|Y_i)\ln(Pr_{\lambda_{n+1}}(Y_i,£_i^*)) \qquad \textbf{1.38}$$

Avec :

$$£_i^* = \underset{£}{argmax} Pr(Y_i,£) \text{ et } \delta_j(i) = \pi_i b_i(o_j)$$

L'estimation de la probabilité de transition a_{ij} entre les deux états q_i et q_j est donnée par la formule [6] :

$$a_{ij} = \frac{\sum_{n=1}^{R}\delta[\text{la transition } q_i q_j \text{ est dans le chemin } £_n^*]}{\sum_{i=n}^{R} le \text{ nombre de fois ou l'état } q_i \text{ est atteint le long du chemin } £_n^*} \qquad \textbf{1.39}$$

Les paramètres de la loi d'observation b_i associée à l'état q_i (elle est supposée gaussienne de moyenne m_i et la matrice de covariance diagonale $\sum_i$) sont donnés par [6] :

$$m_i = \frac{\sum_{n=1}^{R}\sum_{t=1}^{T}\delta(q_i \epsilon £_n^*, X_t = q_i) y_t^i}{\sum_{n=1}^{R}\sum_{t=1}^{T}\delta(q_i \epsilon £_n^*, X_t = q_i)} \qquad \textbf{1.40}$$

$$\sigma_i^2 = \frac{\sum_{n=1}^{R}\sum_{t=1}^{T}\delta(q_i \epsilon \pounds_n^*, X_t = q_i)(y_t^i - m_i)^2}{\sum_{n=1}^{R}\sum_{t=1}^{T}\delta(q_i \epsilon \pounds_n^*, X_t = q_i)} \qquad \textbf{1.41}$$

Avec :

$$\delta(x) = \begin{cases} 1 & \text{si l'évènement x est vrai} \\ 0 & \text{sinon} \end{cases}$$

7.3.4. Estimation par maximum *a postériori* (MAP)

La méthode de maximum de vraisemblance n'est pas toujours le meilleur choix pour estimer les paramètres du MMC, car elle nécessite d'une part, une bonne initialisation des paramètres et d'autre part un nombre assez important de données d'apprentissage pour assurer la convergence de la méthode.

Pour diminuer les effets de ces deux problèmes, on utilise l'estimation par maximum *a posteriori* (MAP) [44]. Cette méthode est utilisée aussi pour :

- ✓ Le lissage des paramètres acoustiques.
- ✓ La classification ou la segmentation des données.
- ✓ L'adaptation d'un système à un locuteur [45].

La différence qui existe entre l'estimation par maximum de vraisemblance et l'estimation par maximum *a postériori* réside dans la supposition que le vecteur de paramètres à estimer θ est non constant, mais la valeur d'une variable aléatoire θ qui possède une densité de probabilité $P_\theta(\theta)$ supposée *a priori*. $P_\theta(\theta)$ est appelé densité à priori.

L'estimation par maximum *a posteriori* (MAP) suppose que le modèle λ est une variable aléatoire, ce qui permet d'estimer $\Pr(\lambda)$ [38,46].

Le critère bayésien ou MAP (Maximum A Posteriori), consiste à maximiser la probabilité à posteriori :

$$\lambda^* = \underset{\lambda}{argmax}\,\Pr(\lambda|Y_1,\dots,Y_R) = \underset{\lambda}{argmax}\Pr(Y_1,\dots,Y_R|\lambda)\Pr(\lambda) \qquad \textbf{1.42}$$

Parmi les problèmes rencontrés lors de l'utilisation de cette estimation nous citons :

- ✓ Le choix d'une densité de probabilité *a priori* pour les paramètres du modèle.
- ✓ Le calcul de cette densité par apprentissage.
- ✓ L'estimation *a posteriori* du modèle.

7.3.5. Estimation par maximum d'information mutuelle

Le critère MMI (Maximum Mutuel Information) [47] tient compte de tous les modèles lors de la ré-estimation des paramètres, ce qui permet de décomposer la probabilité à priori

des observations en une somme de probabilités conjointes sur les modèles.

On note $E = \{\lambda_1, \dots, \lambda_r\}$ une famille de modèles MMC et $\theta = \{\theta_1, \dots, \theta_r\}$ l'ensemble de paramètres des modèles MMC (θ_i est le vecteur paramètres du modèle λ_i).

Etant donnée une suite d'observations acoustiques Y émise par un modèle λ, le critère MMI nous donne :

$$\lambda^* = argPr(\lambda|Y) = \underset{\lambda}{argmax} \frac{\Pr(Y|\lambda).\Pr(\lambda)}{\sum_{\lambda_i \epsilon E} \Pr(Y|\lambda_i).\Pr(\lambda_i)} \qquad \textbf{1.43}$$

Pour plus de détails, on peut consulter [48].

L'inconvénient de cette méthode est qu'elle est couteuse en temps de calcul d'apprentissage et la convergence n'est pas toujours assurée.

7.4. Application du MMC à la reconnaissance automatique de la parole

Dans ce paragraphe, nous allons voir le principe d'utilisation du MMC dans le cas de la reconnaissance des mots connectés et le cas de la parole continue.

7.4.1. Reconnaissance de mots connectés

En reconnaissance des mots connectés, on ne peut associer à chaque mot un modèle acoustique spécifique. Un tel choix nécessite une grande quantité de données d'apprentissage pour estimer ces modèles et un temps de calcul très élevé. Le choix d'unité de taille phonétique plus petite que le mot permet de réduire le nombre de paramètres à estimer et en particulier la quantité de données d'apprentissage.

Un modèle global pour un système de reconnaissance de mots connectés est construit d'une manière hiérarchique suivant trois niveaux (niveau syntaxique, niveau lexical, niveau acoustico-phonétique).

Au niveau syntaxique la phrase est décrite sous la forme d'une concaténation de mots, au niveau lexical chaque mot du vocabulaire est représenté par une suite d'unités élémentaires, au niveau acoustico-phonétique un modèle MMC est associé à chaque unité élémentaire.

Parmi les études faites dans ce sens, nous citons par exemple :

- ✓ [49,50], les auteurs utilisent un algorithme en une passe pour reconnaitre les mots enchainés.
- ✓ [51] a traité les mots avec syntaxe, il a utilisé l'algorithme level-building.

7.4.2. Reconnaissance de la parole continue (RPC)

La reconnaissance de la parole continue est l'un des domaines de recherche intéressant et ceci grâce à son utilisation très variée comme :

- ✓ La dictée vocale.

- ✓ L'accès vocal à des bases de données.
- ✓ Les systèmes de dialogue homme-machines.

La plupart des systèmes de reconnaissance de la parole continue qui existent actuellement sont basés sur le modèle de Markov caché. Cette large utilisation provient de leur simplicité topologique et théorique qui assure la convergence des modèles vers une solution acceptable à condition qu'il existe une quantité suffisante de données d'apprentissage.

Dans la reconnaissance de la parole continue le problème consiste à trouver une suite de mots qui vérifie :

$$w^* = \underset{\lambda}{argmax} \Pr(y_1, \dots, y_T|w)\Pr(w) \quad \textbf{1.44}$$

7.4.2.1. Modélisation acoustico-phonétique

La modélisation par mot dans les systèmes de reconnaissance de la parole continue n'est pas efficace à cause de la quantité énorme de données d'apprentissage. Le choix d'une unité de taille plus courte que le mot devient nécessaire. Les unités les plus utilisées et les plus répandues dans ces systèmes sont les allophones (ce sont des réalisations acoustiques particulières des phonèmes). Si ce modèle phonétique dépend du contexte phonétique gauche et droite, il est appelé triphone. Le problème qui se pose pour ces unités est le nombre très élevé de paramètres à estimer, nous avons n^3 triphones pour n unités phonétiques [52].

Pour diminuer l'effet de ce problème Lee, Bahl et Ljoljie ont regroupé ces triphones en classes ayant des réalisations acoustiques similaires et les faire associer aux mêmes modèles acoustiques [38]. Une autre technique est utilisée pour limiter le nombre de paramètres de ces modèles. Elle consiste à chercher tous les états des MMC d'un même phonème et qui ont certaines similitudes entre eux, ensuite ces états sont liés à la même loi d'observation acoustiques [39,53,54].

7.4.2.2. Modélisation lexicale

Le dictionnaire de prononciation est le lien qui existe entre le modèle acoustique et les entrées lexicales du modèle de langage (chaque entrée lexicale est décrite par une suite d'unités sub-lexical).

Le choix du dictionnaire dépend de deux contraintes :

- ✓ Minimisation du nombre de mots hors vocabulaire (moins utilisé dans les données d'apprentissage)
- ✓ Détermination des prononciations possibles de chaque mot du vocabulaire.

7.4.2.3. Modélisation linguistique

Dans la reconnaissance automatique de la parole, on utilise souvent un modèle de langage qui caractérise la régularité de la langue en question et qui permet d'augmenter les performances des systèmes de reconnaissance.

Pour une prononciation constituée des mots $w = w_1, \dots, w_2$, nous avons :

$$\Pr(w) = \Pr(w_1) \cdot \prod_{i=2}^{n} \Pr(w_i | w_1, \dots, w_{i-1}) \qquad \textbf{1.45}$$

En pratique il est impossible de prendre en compte toute la suite des mots $w_1, \dots, w_{i-1}$. La solution est de se limiter à un nombre fini de mots, par exemple les k-1 derniers mots. Dans ce cas ces modèles sont appelés les modèles k-grammes et ils vérifient :

$$\Pr(w_i | w_1, \dots, w_{i-1}) \cong \Pr(w_i | w_{i-1}, \dots, w_{i-k+1}) \qquad \textbf{1.46}$$

Les modèles de langage les plus utilisés sont les modèles 2-grammes et les modèles 3-grammes.

7.4.2.4. Décodage des informations

En reconnaissance de la parole continue, la suite de mots recherchée est celle qui vérifie l'équation (2.44). Pour résoudre cette équation, il n'est pas possible de construire un modèle pour chacune des phrases pouvant être prononcées puis de comparer tous ces modèles avec la phrase à identifier (comme le cas de la reconnaissance des mots isolés), mais nous sommes amenés à construire un modèle unique qui regroupe toutes les phrases syntaxiquement correctes du langage.

L'identification de la phrase prononcée est déduite à partir du meilleur chemin dans le réseau global associé à ce modèle. Ce réseau est obtenu en compilant tous les modèles de mots du vocabulaire en respectant la syntaxe du langage (figure 1.14).

Plusieurs techniques de recherche du meilleur chemin ont été développées, nous citons par exemple :

- ✓ La recherche en faisceau trame-synchrone, cette technique est utilisée pour les petites et moyennes tailles de vocabulaire. Elle repose sur un algorithme de programmation dynamique [11,18].
- ✓ L'algorithme de Viterbi [37, 55,56].
- ✓ L'algorithme de recherche avant-arrière [56,57].
- ✓ L'algorithme du "Token passing" ou passage de jeton [54].

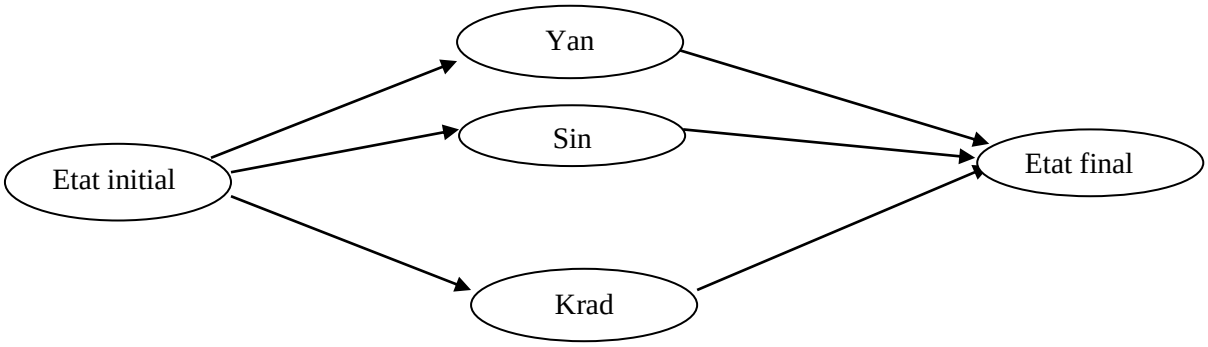


Figure 1.14 : Exemple de réseau pour l'enchainement de mots quelconques

8. Exemple d'un système RAP

Considérons les deux mots suivants en Tamazight :

- ✓ Le mot *YAN* : prononcé par une personne quelconque et qui contient la suite de phonèmes suivante : Y-A- N.
- ✓ Le mot *SA* : prononcé par la même personne et qui contient la suite S-A.

Pour créer un système de reconnaissance automatique de la parole (ici mono-locuteur) on aura besoin d'un ensemble de prononciations dans les situations différentes des deux chiffres ***yan*** et ***sa*** (yan=1 et sa=7). Ceci permet de suivre les différentes prononciations d'une personne afin de construire un système mono-locuteur. Ce système permet de reconnaitre la voix d'une seule personne.

La première étape consiste à paramétriser les signaux liés aux prononciations des deux chiffres. Cette étape donne une matrice de coefficients (ici MFCC) pour chaque prononciation. Si on utilise 13 coefficients (12 MFCC + énergie sans tenir compte la dérivée première et seconde) on obtient une matrice C de taille n*m (avec n=36 et m=13 pour le mot Y-A-N).

La matrice des coefficients MFCC doit être segmentée en trois parties, chacune présente un phonème :

$$Y= C(1 : 12,13),\ A=C(12 : 24,13),\ N=C(24 : 36,13)$$

Pour la création du MMC, on a deux possibilités :

- ✓ Création d'un MMC pour chaque mot sans tenir compte la décomposition phonétique. Ce type de modèle est généralement utilisé pour les bases de données à vocabulaire limité.
- ✓ Création du modèle de phonème : dans ce cas le modèle du mot est obtenu en rassemblant les modèles de phonèmes.

Prenant le deuxième cas, le phonème *Y* sera représenté par un MMC à trois états :

- ✓ Etat 1 : C(1 : 4,13) représenté par le vecteur des moyennes m_1 et la matrice de covariance Σ_1.
- ✓ Etat 2 : C(4 : 8,13) représenté par le vecteur des moyennes m_2 et la matrice de covariance Σ_2.
- ✓ Etat 3 : C(8 : 12,13) représenté par le vecteur des moyennes m_3 et la matrice de covariance Σ_3.

Avec :

$$m_1(1,j) = \frac{\sum_{i=1}^{4} C_{ij}}{4} \ (1 \leq j \leq 13)$$

$$\Sigma_1(k,i) = \frac{1}{4}\sum_{k=1}^{4}(C_{ki} - m_i)(C_{ki} - m_i)^T$$

L'apprentissage du modèle nécessite l'initialisation des paramètres du MMC (ici une seule gaussienne pour chaque état) :

- ✓ Matrice de transition de taille 3*3 est initialisée avec des valeurs quelconques en respectant le modèle gauche droite (remplir le diagonal et le diagonal supérieur).
- ✓ Matrice de probabilités initiales π est initialisée avec des valeurs proches de 1 pour le premier état et des valeurs non nulles pour les deux autres états.
- ✓ Vecteur de probabilité initiales est calculé en utilisant la relation :

$$b_j = \frac{1}{(2\pi)^{\frac{4}{2}}.(\prod_{i=1}^{4}\Sigma_1(j,i))^{1/2}} e^{-\sum_{i=1}^{4}\frac{((C_{ij}-m_j))^2}{2.\Sigma_1(j,i)^2}}.$$

L'apprentissage du modèle se fait en utilisant l'algorithme Expectation-maximisation. Durant l'apprentissage, on ré-estime les paramètres du modèle jusqu'à la convergence (stabilité des paramètres). A la fin de cette apprentissage, on obtient un modèle référence pour chaque phonème composant les deux mots *YAN* et *Sa*.

Pour la phase de reconnaissance on utilise l'algorithme de Viterbi. Cet algorithme permet de parcourir le chemin des états qui maximise la probabilité d'observation. Il cherche le chemin qui maximise la probabilité d'observation sur le réseau suivant :

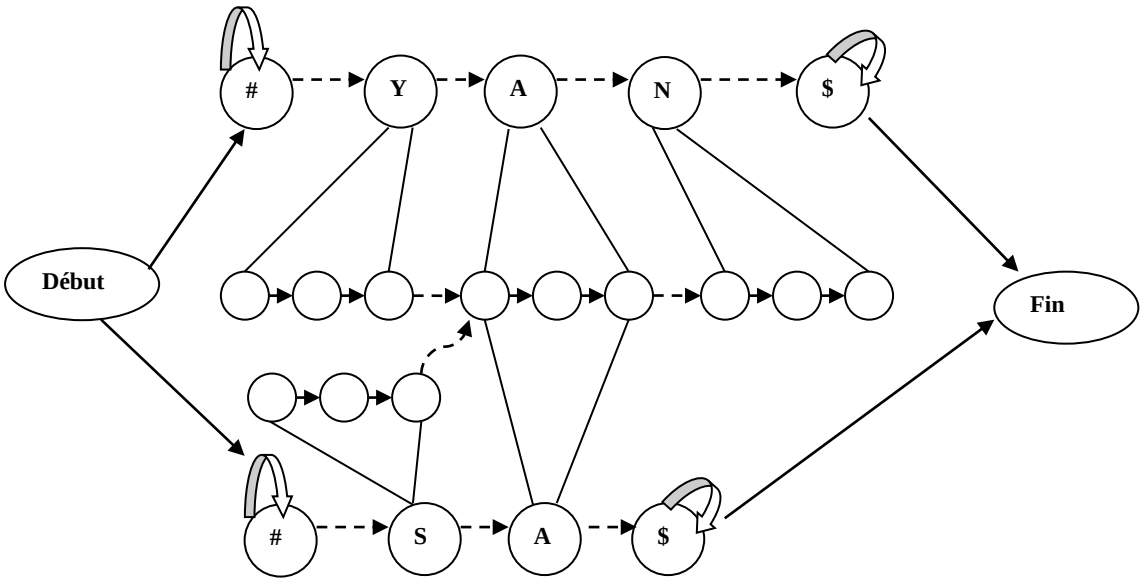


*Figure 1.15 : Réseau de reconnaissance pour un système de deux mots **YAN** et **SA***

Chapitre 2

Reconnaissance automatique des dialectes marocains

1. Introduction

Les dialectes marocains sont des langages populaires les plus communiqués dans le milieu social. Ils sont au nombre de trois : le Darija, le Tamazight et le Hassani.

Le Darija dérive de l'arabe classique en dépassent ses règles ce qui le facilite et élargit son utilisation. Malgré sa différence légère d'une région à l'autre, le Darija constitue le langage populaire le plus utilisé dans les établissements officiels. Dans ce mémoire, nous présenterons un système de reconnaissance de mots isolés du Darija.

Le Tamazight marocain est le dialecte concentré dans les régions montagneuses. On distingue trois familles : l'Amazighe de Souss parlé dans la région du Souss et reconnu sous le nom de Tachelhit [58], l'Amazigh atlas qui se concentre au milieu et l'atlas moyen qui fait référence au Tamazight [58] et l'Amazighe du Rif qui se diffuse au nord du pays nommé Tarifit [58]. Le Tamazight a connu une standardisation depuis son intégration dans les manuels scolaires après la création de l'Institut Royal de la Culture Amazighe Marocain (IRCAM) [58]. Ceci impose une vision interrogative des différentes variétés dialectales afin de construire un Amazighe standard. Dans ce sens, plusieurs articles traitent la politique linguistique et l'action sur le corpus de ce dialecte [59].

La standardisation de l'Amazighe revient, donc, à uniformiser les structures de cette langue, à réduire les divergences entre les trois familles en éliminant les occurrences non distinctives qui peuvent entraver l'intercompréhension. L'adaptation de Tifinaghe [59] en tant que graphie officielle de la langue Amazighe au Maroc, constitue le premier jalon de sa normalisation. Cette étape a permis d'adopter un corpus vocabulaire plus au moins normalisé. Ceci permet d'équilibrer les conventions à travers les trois familles existantes. Cette relative normalisation tend de mettre en disposition des lecteurs un corpus vocabulaire normalisé ce qui permet de faciliter la compréhension de ce dialecte.

Après la présentation détaillée du Tamazight et le Darija, nous aborderons le problème de la reconnaissance des mots isolés de ces deux dialectes. Ensuite, on termine ce chapitre par une analyse des résultats obtenus et une conclusion.

2. Structure du dialecte Darija

Le dialecte marocain Darija est un langage populaire dérivé de l'arabe classique, dont il extrait ses principales composantes. Ce dialecte se diffuse largement au Maroc, puisqu'il constitue l'élément principal de communication au Maroc. Le Darija utilise presque le corpus vocabulaire de l'arabe classique en plus de quelques composantes régionales qui se diffèrent d'une région à l'autre au niveau de la prononciation.

2.1. Composition phonétique de Darija

Le Darija comprend tous les phonèmes de l'arabe classique, il y ajoute quelques composantes régionales. Le corpus phonétique du Darija est donné dans la table 2.1 :

Darija	Correspondance arabe	Correspondance français
A	أ	A
B	ب	B
ḅ	Inexistant en arabe classique	Emphatique de b
T	ت	T
J	ج	J
ḥ	ه	H prononcé au fond de la gorge
ẖ	خ	Comme le J de l'espagnol
D	د	D
ḏ	-	Th de l'anglais
R	غ	Gh en français
ṛ	ر	Emphatique de gh
Z	ز	Z en français
ẓ	Emphatique de ز	Emphatique de z
S	س	S
S	ش	Ch
ṣ	ص	Emphatique de s
ḍ	ض	Emphatique de d
ṭ	ث	Emphatique de t
ɛ	ع	Sorte de vibration de la gorge
F	ف	F
Q	ق	K prononcé au fond de la gorge
K	ك	K
L	ل	L
ḷ	-	Emphatique de l
M	م	M
ṃ	-	Emphatique de m
N	ن	N
W	و	-
Y	ي	-
I	إ	I
A	آ	A long
I		I long

U	أ	-
U	-	U long

Table 2.1 : corpus phonétique du Darija

En Darija, toutes les lettres doivent être prononcées ce qui facilite la transcription phonétique. Néanmoins, les lettres Darija sont toujours prononcées sans voyelles, ceci implique que le signal du Darija contient moins d'énergie que celui de l'arabe classique (Figure 2.1).

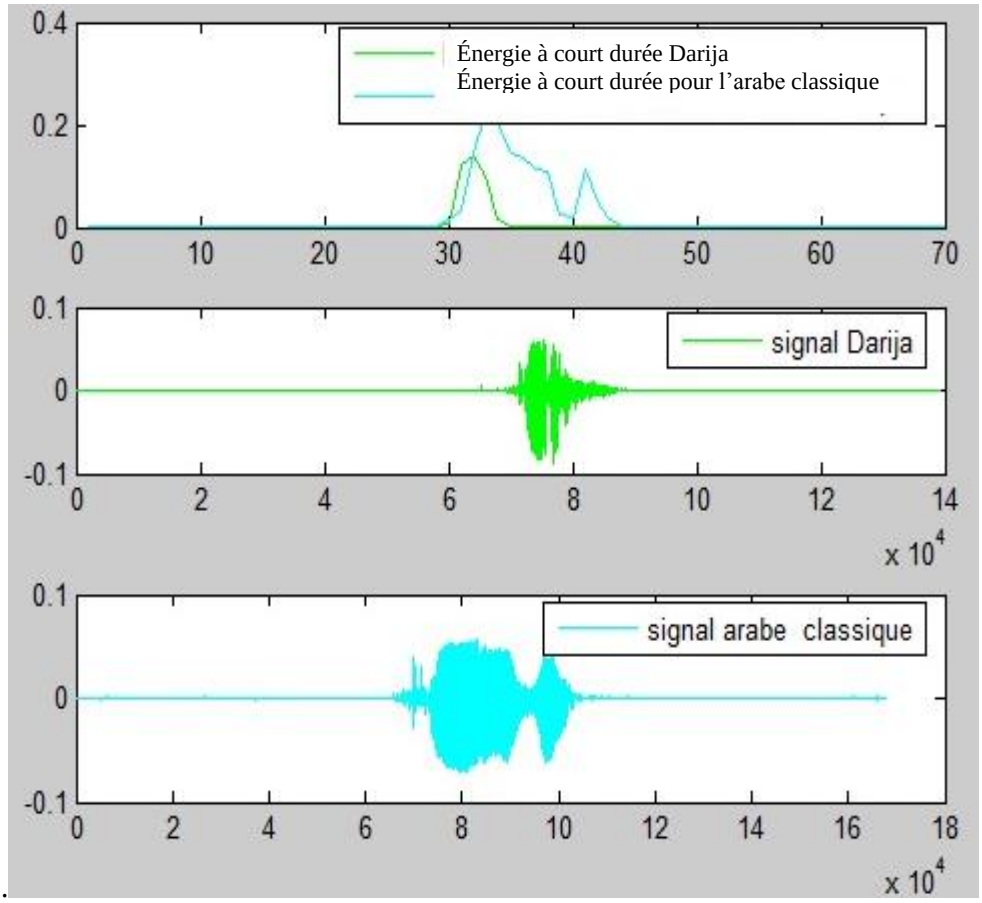


Figure 2.1 : Différence de prononciation du mot 'خَرَج', en Darija on le prononce خُرْجْ

3. Tamazight

3.1. Écriture de Tamazight

Pour transcrire les sons d'un langage, on a recours généralement soit à une transcription phonétique, soit à une transcription phonologique. Elle est dite phonétiques la transcription qui rend, dans ses détails, les sons et les séquences phoniques selon leurs prononciations effectives (exemple ⵙⵎⵙ= e~m~i qui signifie la bouche) [58, 59]. Tandis que la transcription phonologique prend en compte l'emplacement grammatical du mot (exemple : ⵙⵎⵙⵉⵏ qui signifie sa bouche).

L'alphabet Tifinaghe a été élaboré sur la base d'une analyse phonologique et d'un ensemble de critères qui sont :

- ✓ **L'univocité du signe :** elle renvoie au principe général selon lequel à un son correspond un graphème et un seul ; ce qui permet d'éviter l'écriture en digraphe (comme par exemple le *ch* ou le *ph* du français).
- ✓ **L'extension géographique :** elle permet de ne retenir que les oppositions distinctives communes aux trois variantes. Lorsqu'une opposition est très localisée, elle n'est pas retenue.
- ✓ **Le rendement fonctionnel :** ce principe renvoie à la productivité des oppositions phonématiques. En effet, une paire minimale isolée ne permet pas d'octroyer le statut d'unités distinctives aux sons en opposition (cas de I sans emphase /I avec emphase).
- ✓ **La neutralisation de la variation linguistique :** sont exclues du système phonologique présenté les variantes phonétiques non pertinentes. En revanche, plusieurs latitudes de réalisation phonétique selon les parlers et l'environnement phonétique sont permises au niveau de l'oral.

Le système phonologique de Tamazight est composé de 33 unités phonétiques décrites dans la table 2.2 :

	Tifinaghe	Correspondance arabe	Exemples	Prononciation
Graphèmes vocaliques	ⵄ	ا	ⵄⵛⵄⵏ	Achal (la terre)
	ⵉ	ي	ⵉⵎⵉ	Imi (la bouche)
	ⵓ	و	ⵓⴷⵎ	Odm (la face)
	ⵒ	ⴻ	ⵜⵒⵜⵜ	Titt (l'œil)
Semi-consonnes simples	ⵢ	يَ	ⵄⵢⵢⵉⵙ	Ayis (le cheval)
	ⵡ	وَ	ⵄⵡⵄⵏ	Awal (la parole)
Emphatiques	ⴹ	ض	ⵄⴹⵄⵇ	Adar (le pied)
	ⵟ	ط	ⵜⵉⵟⵟ	Tit (l'œil)
	ⵣ	ظ	ⵉⵣⵉ	Izi (la mouche)
	ⵚ	ص	ⵄⵚⵎⵉⴷ	Asmid (le froid)
	ⵕ	ر	ⵠⵕⵇⵄ	Rbbi (le dieu)
Labiovélaires	ⴽ⵽	‘ك	ⵄⵎⴷⴰⴽ⵽ⵏ	Amdakl (l'ami)
	ⵅ⵽	كَ	ⵄⵣⵅⵅ⵽ⵄⵢ	Azggagh(le rouge)
Consonnes simples	ⵠ	ب	ⵄⵠⵄⵉⴷ	Abrid (le chemin)
	ⵎ	م	ⵄⵎⵄⵍ	Aman (l'eau)
	ⴼ	ف	ⵄⴼⵓⵠ	Afoud (la jambe)
	ⵜ	ت	ⵜⵓⵠⵠⵍⵄ	Tossna (la culture)
	ⴷ	د	ⵄⴼⵓⴷ	Afoud
	ⵏ	ن	ⵉⵔⵏ	Irdn (les grains)
	ⵙ	س	ⵉⵍⵙ	Ils (la langue)
	ⵣ	ز	ⵄⵎⵄⵣⵉⵢ	Amazigh
	ⵍ	ل	ⵄⵎⵍⵄⵏ	Amlil (le blanc)
	ⵔ	ر	ⵓⵔⵢ	Orgh (l'acier)
	ⵛ	ش	ⵄⵛⵡⵡⵄⵏ	Achwal

Consonnes simples	ⵊ	ج	ⵄⵎⵊⵓⵔ	Amjoud
	ⴽ	ك	ⵄⴽⵇⴽⵓⵇ	Akrkour (les rauches)
	ⵅ	‘ك	ⵄⵅⵎⵄⵔ	Agmar (le cheval)
	ⵖ	غ	ⵄⵖⵇⵓⵎ	Aghroum (le pain)
	ⵄ	ع	ⵄⵄⵏⵄⵍⵡ	A3daw (l’ennemi)
	ⵇ	ق	ⵄⵇⵇⵄⵖ	A9rab (le cartable)
	ⵃ	ح	ⵄⵃⵙⵏⵓⵖ	Ahidous
	ⵅ	خ	ⵜⵙⵅⵖⵄ	Tikhba (les trous)
	ⵀ	ه	ⵄⵀⵏⵏⵓⵍ	Ahddoun (le tapi)

Table 3.2 : Système phonologique de l’Amazighe

3.2. Gémination en Tamazight

- ✓ La gémination concerne toutes les consonnes, elle est rendue, au niveau de l’écrit, par le dédoublement du graphème. Pour les labiovélaires géminées, seul le deuxième graphème porte l’indice de la labio-vélarisation (ⴽⴽ̈ et ⵅⵅ̈) [59].
- ✓ Un *schwa* prononcé ne sera noté que dans deux cas :
  - Dans des suites de plus de deux consonnes identiques (ⵜⵙⵉⵣⵉⵙ, ⵜⵣⵜⵜⵔ, ⵣⵣⵜⵜⵙ, ⵜⵣⵜⵜⵙ).
  - Dans les radicaux verbaux qui se terminent par deux consonnes identiques (ⵜⵉⵣⵉ « être blanc »).

La table 2.3 et la table 2.4 résument respectivement le système consonantique (Tifinaghe) et le système vocalique de l’amazigh standard :

Lieu d’articulation / mode d’articulation			Labiales	Dentales	Alvéolaires	Palatales	Vélaires	Labiovélaires	Uvulaires	Pharyngales	Laryngales
Occlusives	Non Emphatiques	Sourdes		ⵜ			ⴽ	ⴽ̈	ⵇ		
		Sonores	ⵖ	ⵍ		ⵅ	ⵅ	◌̈			
	Emphatiques	Sourdes		ⵟ							
		Sonores		ⵜ							
Constrictives	Non Emphatiques	Sourdes	ⵃ		ⵙ	ⵛ			ⵅ	ⵄ	ⵀ
		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					

Table 2.3 : Phonologique de l’Amazighe standard

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

Table 2.4 : le système vocalique de l'Amazighe standard

4. Reconnaissance automatique des dialectes marocains

Dans cette partie, nous utilisons le modèle de Markov caché (MMC) [38,52] pour élaborer un système de reconnaissance automatique de la parole relatif aux deux dialectes marocains Tamazight et Darija. La modélisation phonétique de chaque dialecte permet de déterminer le type du modèle MMC à utiliser. Ici nous avons pris comme unité de base le phonème. Chaque mot est divisé en phonèmes dont chacun sera présenté par un modèle de Markov caché à trois états (figure 2.2).

Le signal de la parole ne peut pas être exploité directement. En effet, il contient de nombreux autres éléments autre que le message linguistique : des informations liées au locuteur, aux conditions d'enregistrement, au bruit,…etc. Toutes ces informations ne sont pas nécessaires lors du décodage de la parole et rajoutant même du bruit. De plus, la variabilité et la redondance du signal de la parole les rend difficilement exploitable tel qu'il est. Il est donc nécessaire d'en extraire uniquement les paramètres qui sont dépendants du message linguistique.

Généralement, comme il est décrit dans la première partie, ces paramètres sont estimés via des fenêtres glissantes sur le signal. Cette analyse par fenêtre permet d'estimer le signal sur une portion du signal jugée stationnaire : généralement de 10 à 30 ms en limitant les effets de bord et la discontinuité du signal via une fenêtre de Hamming [52].

La majorité des paramètres représentent le spectre fréquentiel et son évolution sur une fenêtre de taille donnée. Dans ce travail, nous avons utilisé les coefficients de Mel MFCC (Mel Frequency Cepstrum Coefficient). Chaque fenêtre de 10ms donne un vecteur acoustique de 13 coefficients. Pour améliorer la qualité de reconnaissance nous avons ajouté la dérivée première et seconde des coefficients MFCC, ce qui donne en total 39 coefficients réels (39=13+Δ13+ΔΔ13).

Après la phase de décodage décrite ci-dessus, le signal est représenté par une matrice de taille n*m, avec *n* le nombre de lignes correspondant au nombre de fenêtres glissantes dans le

signal (il dépend de la durée du signal audio) et *m* le nombre de coefficients réels (on prend m=39).

La matrice des coefficients doit être segmentée selon le nombre de phonèmes dans le mot prononcé. Chaque portion correspondant à un phonème doit être elle-même segmentée en trois parties dont chacune est représentée par un état dans le modèle de Markov caché correspondant. Le vecteur des moyennes ainsi que la matrice de covariance sont calculés pour chaque coefficient en utilisant les relations suivantes :

$$m_{etat\ i}^{k} = \frac{\sum_{j=1}^{n} c_{jk}}{N} \qquad \textbf{2.1}$$

$$C_{(etat\ i)} = \frac{1}{N}\sum_{j=1}^{N}(x_j - m_i)(x_j - m_i)^T \qquad \textbf{2.2}$$

Avec :

$\boldsymbol{c_{jk}}$: Coefficient MFCC de la j[ème] ligne et k[ème] colonne.

$\boldsymbol{x_k}$: Valeur à l'instant k du signal échantillonné.

$\boldsymbol{m_i}$: Moyenne relative à l'état i.

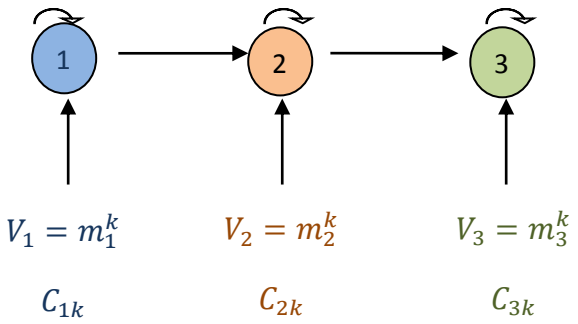


Figure 2.2 : MMC à trois états

5. Modèle mélange de gaussiennes

La distribution gaussienne s'applique sur les variables aléatoires qui ont des fluctuations autour d'une valeur moyenne. Ce modèle est très utilisé en reconnaissance automatique de la parole. Il repose sur le fait qu'on peut modéliser les données par une fonction de densité de probabilité qui combine plusieurs fonctions gaussiennes. Chacune de ces fonctions est appelée composante. Chaque composante correspond à une loi normale multidimensionnelle. Le modèle obtenu après apprentissage se constitue du vecteur des moyennes et la matrice de covariance pour chaque composante (figure 2.3).

Figue 2.3 : Modèle à trois ($\hat{\theta}_1$) et quatre ($\hat{\theta}_1$) composantes gaussiennes

Les vecteurs de paramètres obtenus du décodage des signaux peuvent être modélisés par cette distribution multidimensionnelle. Dans un état du modèle de Markov caché, les vecteurs acoustiques sont modélisés par des classes de gaussiennes. Chaque vecteur appartient à une classe donnée. L'intérêt de telle modélisation est de permettre une meilleure distribution des données. Cette distribution multi variée permet aussi d'améliorer le taux de reconnaissance en maximisant la probabilité obtenue lors de la phase de reconnaissance. Le calcul de la probabilité associée au mélange de gaussiennes se fait en utilisant la relation suivante :

$$b_j(y_t) = \sum_{k=1}^{G} \alpha_k . N_{(\mu_k, \sum_k)}(y_t) \quad \textbf{2.3}$$

Avec :

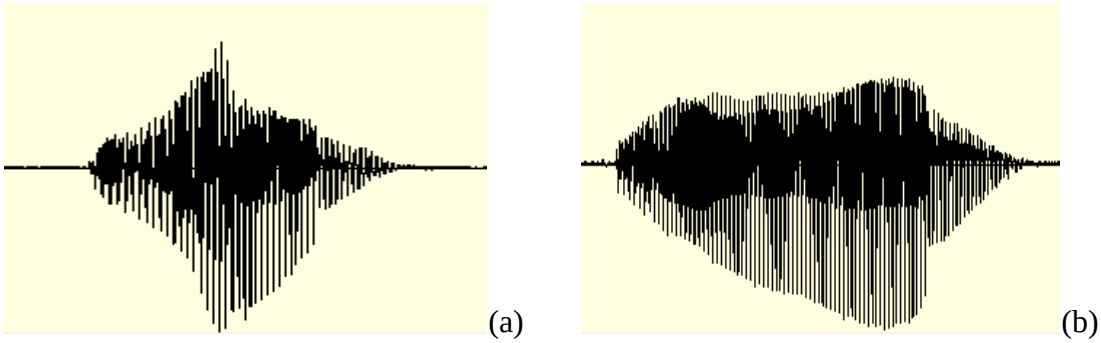
α_k : Poids lié à la $k^{ième}$ gaussienne.

$N_{(\mu_k, \sum_k)}$: Densité de probabilité normale de moyenne μ_k et de matrice de covariance $\sum_k$.

6. Influence de la production vocale sur le SRAP

6.1. Influence de l'état physique de locuteur

La parole est un signal non-stationnaire produit par un ensemble de mécanismes commençant par les poumons et passant par les cordes vocales et finissant par la cavité nasale ou labiale. Le changement de l'état physique ou psychique de l'un de ces mécanismes influe directement sur le signal vocal produit. La figure ci-dessous présente le mot 'un' prononcé en Tamazight (yan) par deux personnes, la figure 2.4(a) pour une personne normale et la figure 2.4(b) pour une personne malade.



(a) (b)

Figure 2.4 : 'yan' produit par une personne normale (a) et par une personne malade (b)

6.2. Influence de l'accent régional

Les dialectes marocains se diffèrent d'une région à l'autre et au sein de la même région. En effet, cette variabilité est due à l'influence des accents régionaux. Le Tamazight est aussi influencé par cette variabilité dû aux trois familles (Tachelhit, Tamazight, Tarifit). Par exemple, la prononciation du mot 'akal' (qui signifie la terre) a deux variantes : est '***acal'*** et '***achal***' selon les régions.

L'existence des accents régionaux exige la prise en compte de toutes les prononciations possibles d'un mot donné dans le système de reconnaissance. Néanmoins, l'IRCAM a pu mettre en disposition des chercheurs un corpus vocabulaire standard [59]. Ce corpus est composé d'une base de vocabulaire normalisé sans prendre en considération des composantes régionales les plus spécifiques.

6.3. Influence des zones intermédiaires

L'apparition de cet effet est confinée dans les régions de transition. Pour le Tamazight, dans les régions voisines de Souss, l'influence de cet accent apparaît dans la prononciation de ces régions. Par exemple, le mot 'cha' (qui signifie quelque chose) se prononce 'ka' dans les régions voisines de Souss.

6.4. Position des segments acoustiques

La position des segments acoustiques peut changer dans un même mot. Cette inversion est due à la difficulté de prononciation. On renverse les segments pour faciliter la prononciation. Le mot '*adil*' qui signifie '*cépage*' se prononce parfois '*alid*'.

7. Système de reconnaissance automatique des dialectes marocains

7.1. Le dialecte Darija

7.1.1. Base d'apprentissage

On propose un système de reconnaissance de mots isolés relatif au Darija, On a considéré les chiffres isolés de 0 à 9 suite à leur grande utilisation pour la saisie des mots de passe. Chaque chiffre est prononcé cinq fois par le même locuteur. La base des locuteurs est

prise diversifiée. Elle est composée de différents types de locuteurs (garçons, filles, adultes). La figure 2.5 présente la distribution des locuteurs selon leur sexe et leur âge.

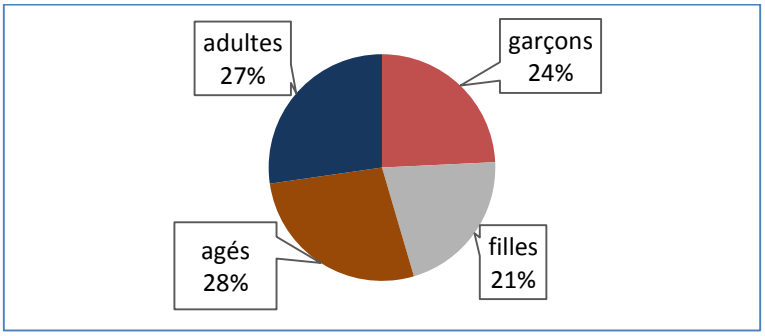


Figure 2.5 : base d'apprentissage

Les données audio sont capturées à l'aide d'un microphone. Ensuite, elles sont échantillonnées à une fréquence de 16Khz. Les fichiers obtenus sont soumis à un traitement pour extraire les coefficients de Mel (Mel Frequency Ceptrum Coefficients MFCC). La table 2.5 présente le contenu de la base d'apprentissage, des chiffres de 0 à 9 prononcés en Darija.

Chiffre	Prononciation française	Transcription arabe
0	SIFFER	صِفْرْ
1	WAHED	وَاحْدْ
2	JUJ	جُوجْ
3	TLATTA	ثْلاَثَة
4	REB3A	رْبْعَة
5	KHEMSSA	خْمْسَة
6	SETTA	سْتَّة
7	SEB3A	سْبْعَة
8	TMANIYA	ثْمْنْيَة
9	TES3OUD	تْسْعُودْ

Table 2.5 : Contenu de la base d'apprentissage

7.1.2. Segmentation des signaux audio

La segmentation du signal de la parole constitue une étape importante dans la reconnaissance automatique de la parole. Le système doit être capable de détecter le début ainsi que la fin d'un segment acoustique. Pour cela, plusieurs approches ont été utilisées pour la segmentation de la parole. Parmi les plus connues, la segmentation par modèle de Markov caché [19] et la segmentation basé sur l'algorithme à seuil [50]. Dans ce travail, la segmentation des signaux continus est basée sur le seuillage de l'énergie du signal. Pour améliorer la qualité de la segmentation on ajoute un autre critère basé sur le centre de gravité spectral.

On rappelle que la formule de calcul de l'énergie d'un signal est donnée par l'équation suivante :

$$E_S = \int_{-t}^{+t} |x(t)^2| dt \quad \textbf{2.4}$$

E_s est l'énergie totale sur un signal continu non échantillonné et *x(t)* représente la valeur du signal pour un instant donné t.

On utilise une sommation puisqu'on travaille avec un signal échantillonné, donc un signal à valeurs discrets. L'énergie d'un signal est liée directement à la puissance de production de la voix. Quand on parle à voix basse on utilise peu d'énergie et en traçant l'allure du signal, il aura une faible amplitude. Donc, on peut construire une estimation de l'énergie à partir de l'amplitude du signal. On met un carré pour récupérer les valeurs négatives du signal et éviter les équations qui peuvent suivre un signal à énergie nulle.

L'algorithme repose sur le seuillage de l'énergie à court terme et le centre de gravité spectral. En effet, l'énergie est calculée sur chaque fenêtre glissante du signal échantillonné (valeurs discrètes) en utilisant la formule suivante :

$$E_i = \frac{1}{N}\sum_{j=1}^{N} |X_{ij}|^2 \text{ (Avec j=1…..N suite d'échantillons de la fenêtre i)} \quad \mathbf{2.5}$$

L'allure de la figure 2.6 présente la variation de l'énergie à court terme pour la suite de mots 'wahed jouj' :

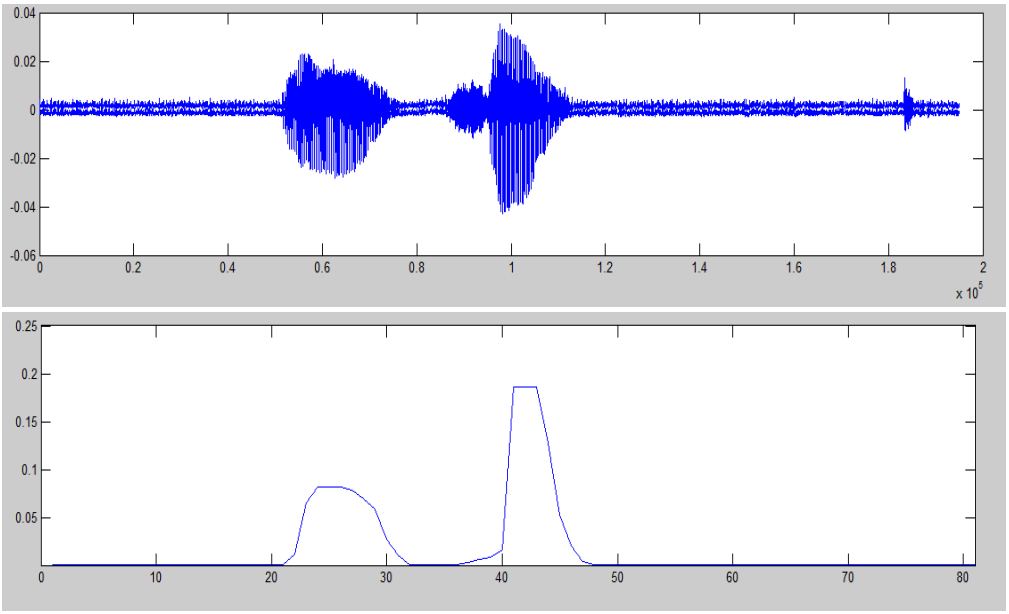


*Figure 2.6 : Signal de la parole pour la suite '**wahed jouj**' (en haut) et son énergie à court terme (en bas).*

L'énergie à court terme permet de détecter les zones de silence dans les signaux de la parole, comme il permet de différencier les classes d'un signal (fricatives, nasal, …).

Le deuxième paramètre utilisé pour la segmentation est le centre de gravité spectral qui correspond au centre de gravité du spectre du signal qui est mesurable en Hz. Il est calculé à

partir des moyennes des amplitudes pondérées par les fréquences des harmoniques du son selon la formule :

$$CGS = \frac{\sum_{k=1}^{N}(k+1)f_i(k)}{\sum_{k=1}^{N} f_i(k)} . f_i(k) \quad \textbf{2.6}$$

Avec :

$f_i(k)$: La $i^{ème}$ transformée de Fourier sur le $i^{ème}$ frame (figure 2.7).

N : nombre d'échantillons dans le frame i.

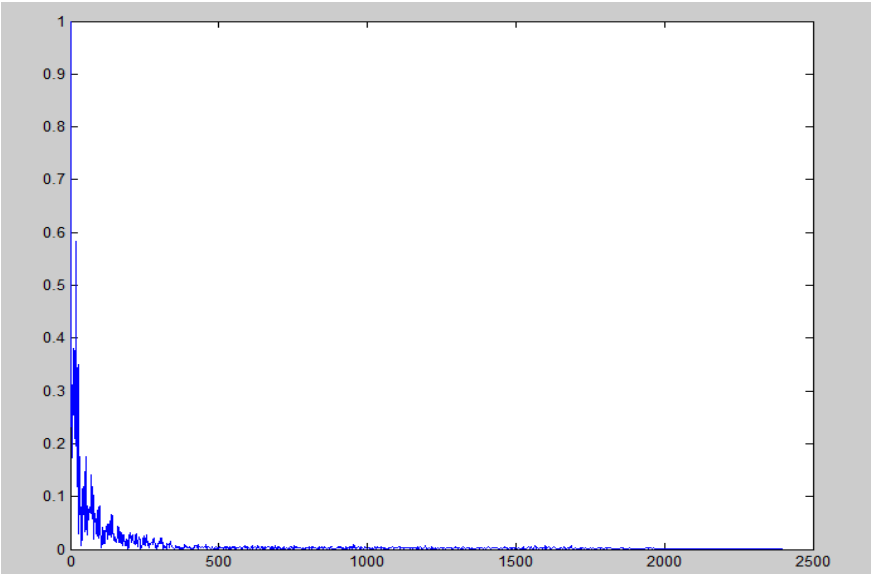


Figure 2.7 : Transformée de Fourier sur le frame 4 du signal

L'allure sur la figure 2.8 présente la variation du centre de gravité spectral pour la suite 'wahed jouj' :

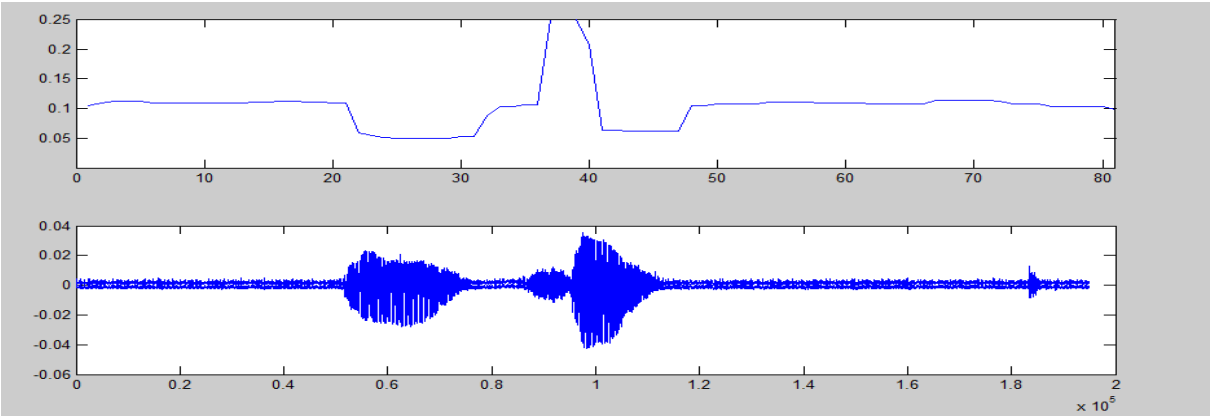


Figure 2.8 *: Signal de la parole pour la suite 'wahed jouj' (en bas) et son centroïd spectral (en haut).*

Le centre de gravité spectral est une caractéristique qui mesure la position spectrale, avec des valeurs élevées qui correspondent à des sons 'brillantes'. Cette caractéristique est notamment plus variante pour les zones qui correspondent à la parole.

Algorithme 2.1 : Détection de début et fin de la parole

Algorithme : détection du début et fin de la parole

Entrées : signal, Énergie E, centre de gravité spectral C

Sortie : Zones corresponds à la parole dans le signal

- ✓ ***Début : compter les valeurs de l'énergie et le centroïd spectral notées respectivement: HistE et HistC ;***
- ✓ ***Application du filtre médian pour le lissage des valeurs ;***
- ✓ ***Détection du premier et second maximum local pour l'énergie et le centroïd notés respectivement M1_Ec, M2_E et M1_C, M2_C ;***
  Calcul des seuils de décision définis par :

$$S_E = \frac{Cst * M1\_E + M2\_E}{Cst}$$

$$S_c = \frac{Cst * M1\_C + M2\_C}{Cst}$$

Avec Cst=5 ;

- ✓ ***Si $E \geq S_E$ et $C \geq S_c$, alors la partie est considérée parole sinon il se sera considérée silence ;***

La figure 2.9 présente la segmentation de la suite 'Sin_yan', les zones de silence ou bruit sont données en gris. La figure 2.10 présente la segmentation du mot isolé 'wahed'.

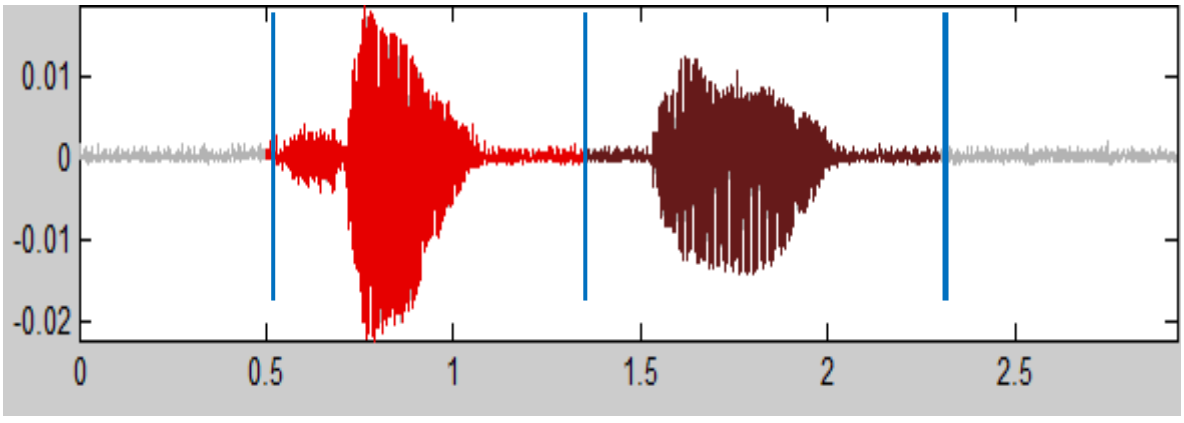


*Figure 2.9 : Résultats de la segmentation de la suite '**Sin_yan**'*

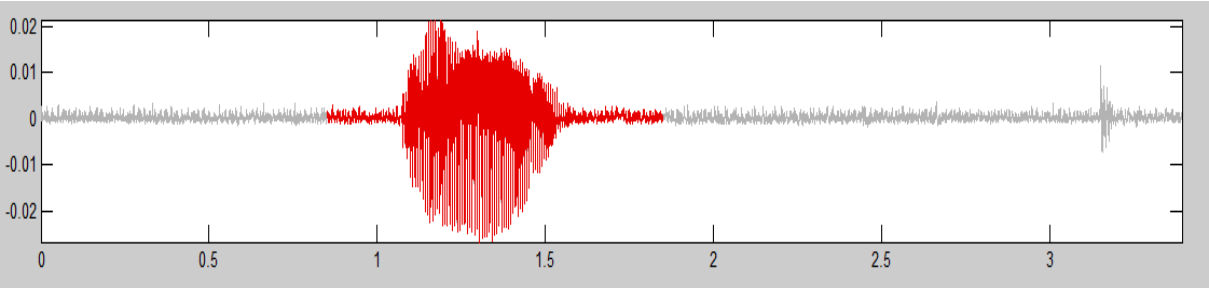


*Figure 2.10 : segmentation du mot isolé '**wahed**'*

La table 2.6 présente les résultats de la segmentation selon le nombre de mots dans le flux de la parole :

Nombre d'occurrences dans le flux	Taux de segmentation	Taux d'erreur
2	100%	0%
3	99%	1%
4	91%	9%
10	84%	16%

Table 2.6 : Résultats de la segmentation

Selon les résultats obtenus, le taux de segmentation est inversement proportionnel au nombre de mots dans le flux et proportionnel à la taille du silence entre les mots. Pour avoir une meilleure segmentation, il faut faire des pauses et augmenter la taille du silence entre chaque prononciation dans les signaux contenant plus que deux prononciations, chose qui n'est pas toujours vraie dans la réalité et, par conséquent, la segmentation devient plus difficile.

7.1.3. Apprentissage

L'apprentissage de la base de données est fait en utilisant l'algorithme de Baum-Welch [38]. Cet algorithme repose sur les étapes suivantes :

- ✓ Utilisation de l'algorithme de quantification vectorielle : cet algorithme permet d'initialiser les centres de gaussiennes, cet algorithme est donné en détail dans la partie annexe 10.
- ✓ Utilisation de l'algorithme de Forward-Backward [38] : cet algorithme est décrit dans le premier chapitre. Il est basé sur la topologie avant-arrière pour calculer la probabilité d'observation.
- ✓ Ajustement des paramètres du modèle en utilisant l'algorithme de Baum-Welch [38]. La mise à jour des paramètres du modèle dans chaque itération se fait via les relations suivantes :

$$C_j = \frac{\sum_{t=1}^{T} \gamma_t(j,k)}{\sum_{t=1}^{T} \gamma_{(j)}}, 1 \leq j \leq N, 1 \leq k \leq M \qquad 2.7$$

$$m_j = \frac{\sum_{t=1}^{T} o_t \gamma_t(j,k)}{\sum_{t=1}^{T} \gamma_t(j)}, 1 \le j \le N, 1 \le k \le M \qquad \textbf{2.8}$$

$$Co_j = \frac{\sum_{t=1}^{T} (o_t - m_j)(o_t - m_j)' \gamma_t(j,k)}{\sum_{t=1}^{T} \gamma_t(j)}, 1 \le j \le N, 1 \le k \le M \quad \textbf{2.9}$$

$$\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{p(o|\lambda)} \qquad \textbf{2.10}$$

Avec :

M : nombre de gaussiennes.

N : nombre de vecteurs acoustiques pour chaque état.

$$\gamma_t = \frac{\alpha_t(j)\beta_t(j)}{\sum_i \alpha_t(i)\beta_t(j)} = \sum_{j=1}^{N} \xi(i,j)$$

$$\gamma_t(j,k) = \gamma_t\left(\frac{\mathcal{N}(o_t, m_{jk}, Co_{jk}).c_{jk}}{\sum_{k=1}^{M} C_{jk}\,\mathcal{N}(o_t, m_{jk}, Co_{jk})}\right)$$

C_{jk} représente le poids de gaussienne k relatif à l'état j et les coefficients α et β sont calculés par l'algorithme de Forward-Backward [60]. Cet algorithme permet le calcul de la probabilité d'observation$p(o|\lambda)$, le principe considère que l'observation peut se faire en deux étapes :

Étape 1 :

- ✓ *La variable Forward :* représente l'émission de la séquence d'observations $\{o_1, o_2, \dots, o_t\}$ et la réalisation de l'état q_t au temps t [4,14] , soit :

$$\alpha_t(i) = p(o, q_t = s_i|\lambda) \quad \textbf{2.11}$$

La probabilité $\alpha_t(i)$ est calculée de manière récursive comme suit :

Algorithme 2.2 : Algorithme de Forward

Algorithme : Forward

Initialisation : $\boldsymbol{\alpha_t(i) = \Pi_i b_i(o_1)}$ ***1≤i≤N***

Récurrence : $\boldsymbol{\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] * b_j(o_{t+1})}$ ***t ϵ [1, T-1], 1≤j≤N***

Terminaison : $\boldsymbol{p(o|\lambda) = \sum_{i=1}^{N} \alpha_T(i)}$

Étape 2 :

- ✓ *La variable Backward :* représente l'émission de la suite d'observations$\{o_{t+1}, o_{t+2}, \dots, o_T\}$, en partant de l'état q_t au temps t [16, 19], soit :

$$\beta_t(i) = p(o|q_t = s_i, \lambda) \quad \textbf{2.12}$$

On déduit β_t et β_{t+1} par l'algorithme suivant :

Algorithme 2.3 : Algorithme de Backward

Algorithme : Backward

Initialisation : $\boldsymbol{\beta_T(i) = 1, 1 \leq i \leq N}$
Récurrence : $\boldsymbol{\beta_t(i) = \left[\sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)\right], T - 1 \geq t \geq 1, 1 \leq i \leq N ;}$
Terminaison : $\boldsymbol{p(o|\lambda) = \sum_{i=j}^{N} \Pi_j b_j(o_1)\, \beta_1(j)}$

L'algorithme Forward-Backward considère que l'observation peut se faire par l'émission de début de la séquence O(1:t) et d'aboutir à l'état q_t au temps t, puis émission de la fin de l'observation O(t+1 :T) en partant de l'état q_t au temps t. Le calcul de $\alpha_t(i)$ se fait avec t croissant tandis que celui de $\beta_t(i)$ se fait avec t décroissant, d'où l'expression Forward-Backward.

P(o|λ) peut être définie à chaque instant t ϵ [1,T] par :

$$p(o|\lambda) = \sum_{i=1}^{N} \alpha_t(i)\beta_t(i) \qquad \textbf{2.13}$$

Dans le cas où (t=0) (T=0) on obtient :

$$p(o|\lambda) = \sum_{i=1}^{N} \alpha_T(i) = \sum_{i=1}^{N} \Pi_i \beta_0(i) \quad \textbf{2.14}$$

Algorithme 2.4 : algorithme de Baum-Welch

Algorithme : Baum-Welch,

1. ***Initialisation : fixer les valeurs initiales :***
$$\boldsymbol{a_{ij}^0, b_j^0(k), \Pi_i^0, 1 \leq i, j \leq N, 1 \leq k \leq N}$$
1. ***Calculer à l'aide des fonctions de Forward-Backward :***
$$\boldsymbol{\xi_t(i,j), \gamma_t(i), 1 \leq i, j \leq N, 1 \leq t \leq T - 1}$$

Et $\boldsymbol{\bar{\lambda}}$ ***en utilisant les formules de ré-estimation.***
Recommencer de l'étape 2 jusqu'à ce que les critères de convergence soient remplis.

7.1.4. Reconnaissance

Le principe de la reconnaissance peut être expliqué comme le calcul de la probabilité P(W|S) : la probabilité qu'une suite de mots W correspond au signal S et de déterminer la suite de mots qui maximise cette probabilité [14, 19].

Selon la formule de Bayes la probabilité P (W|S) peut s'écrire :

$$P(W|S) = \frac{P(W).P(S|W)}{P(S)} \quad \textbf{2.15}$$

Avec :

- ✓ P(W) : probabilité a priori de la suite de mots W (Modèle de langage).
- ✓ P(S|W) : probabilité du signal S, étant donnée la suite de mots W (Modèle acoustique).
- ✓ P(S) : probabilité du signal acoustique S (indépendant de W).

Dans la phase de reconnaissance, on utilise souvent l'algorithme de Viterbi [29, 32]. Ce dernier se base sur un graphe de recherche composé de différents modèles de mots. La concaténation de la suite de phonèmes dans le chemin qui maximise la probabilité d'observation détermine le mot reconnu. Dans la figure 2.11 une illustration de la reconnaissance par l'algorithme de Viterbi.

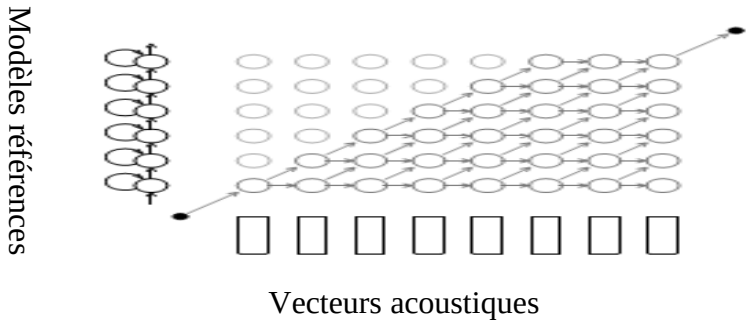


Figure 2.11 : Algorithme de Viterbi

On rappelle que l'approximation de Viterbi est donnée par :

$$p(y|w) = \sum_{s} p(s|w)p(y|s,w) \sim \max_{s} p(s|w)p(y|s,w)$$

Cet algorithme se base sur la recherche de chemin des états s qui maximise la probabilité d'observation. Il se base sur la recherche des meilleurs chemins dans un graphe de recherche en utilisant la formule suivante :

$$S(j,t) = \max_{i} S(i,t-1) + \ln(a_{ij}) + lnb_j(y_t)$$

Avec :

$\max_{i} S(i,t-1)$: La suite d'états qui maximise la probabilité d'observation jusqu'à l'instant t.

a_{ij} : Transition relative au chemin i→j.

$b_j(y_t)$: Probabilité d'observation de y_t par l'état j.

Algorithme 2.5 : Algorithme de Viterbi

Algorithme : Viterbi

Entrées : modèles MMC références de chaque unité phonétique, matrice MFCC pour le nouveau signal d'entée ;

Sortie : Suite d'états maximisant la probabilité d'observation, vraisemblance ;

Initialisation : t=1, 1≤i≤N, $\boldsymbol{\delta_1(i) = \Pi_i * b_i(o_1)}$

Récurrence : t ϵ [2, N], 1≤i≤N

$$\boldsymbol{\delta_{t+1}(i) = \underset{j=1..n}{max}\left(\delta_t(j) * a_{i,j}\right) * b_i(o_{t+1})}$$

$$\boldsymbol{\psi_t(i) = \underset{j=1..n}{arg\ max_j}(\delta_{t-1}(j) * a_{j,i})}$$

Terminaison : $\boldsymbol{s(N) = \underset{i}{argmax}\delta_T(i)}$

Retour en arrière : N-1≥t≥1 $\boldsymbol{s(t) = \psi_{t+1}(s(t+1))}$

Au lieu de l'algorithme de Viterbi, on peut utiliser dans la phase de reconnaissance, le célèbre algorithme A^* qui cherche le chemin le plus court dans un graphe entre un nœud initial et un nœud final [25]. Nous avons utilisé l'algorithme de Viterbi dans sa version classique suite à sa rapidité et à ses bons résultats au niveau de la reconnaissance.

7.1.5. Transcription phonétique du Darija Marocain

Dans un système de reconnaissance automatique de la parole, un mot est représenté par une suite de phonèmes ou syllabes. Chaque phonème est représenté par un MMC à trois états [101]. Dans la phase d'entrainement, chaque modèle de phonème est entrainé indépendamment des autres phonèmes constituants le mot en question. La table 2.7 présente la segmentation phonétique de chaque mot de la base d'apprentissage.

Chiffre	Transcription Phonétique
0	S I F R
1	W A H D
2	J U J
3	TT L A TT A
4	R E B 3 A
5	KH E M S A
6	S E TT A
7	S E B 3 A
8	TT M E N Y A
9	T E S 3 OU D

Table 2.7 : Chiffres de la base d'apprentissage et leurs transcriptions phonétiques

Les segments acoustiques qui représentent le même phonème seront représentés par les MMCs, ce qui facilite la reconnaissance de phonèmes issus des personnes différentes et suivre les différentes prononciations d'un mot.

7.2. Résultats expérimentaux pour Darija

Notre système de reconnaissance est testé sur une base de données diversifiée constituée de 300 prononciations prises par différentes personnes en y ajoutant des fichiers audio bruités. Le système est évalué en calculant le taux de reconnaissance qui définit le pourcentage des mots reconnus par le système, il est donné par :

$$T = \frac{\text{nombre de mots reconnus}}{\text{taille du corpus de test}} \quad \textbf{2.16}$$

On peut aussi calculer les quantités suivantes :

- ✓ T_R = taux de rejet, taux des mots qui ne sont pas reconnus par le système.

$$T_R = \frac{\text{nombre de mots non reconnus}}{\text{taille du corpus de test}} \quad \textbf{2.17}$$

- ✓ T_E =taux calculé sur l'erreur d'identification d'un mot.

$$T_E = \frac{\text{nombre de mots mal reconnus}}{\text{taille du corpus de test}} \quad \textbf{2.18}$$

Les résultats obtenus sont données dans la table 2.8 :

Base d'apprentissage	Base de test	Taux de reconnaissance
1h 20 min	200 prononciations isolées de 0 à 9	T=91% T_R=2,68% T_E=6,32%

Table 2.8 : résultats obtenus

La programmation dynamique ou DTW (Dynamics Time Warping), permet de comparer un signal de la parole avec des signaux de référence. Cette technique est déjà décrite dans le deuxième chapitre. Elle est basée sur le calcul de la distance euclidienne entre les vecteurs de paramètres acoustiques des deux signaux. La comparaison des résultats obtenus par le MMC et la programmation dynamique est donnée dans la table 2.9 :

	MMC	DTW
Temps d'exécution	Rapide quelques millisecondes	Plus lente surtout pour la comparaison des fichiers audio plus volumineux (la reconnaissance d'une suite de mots peut prendre jusqu'à 1 minute)
Taux de reconnaissance	91%	60%

Table 2.9 : Comparaison entre le temps d'exécution de MMC et DTW

L'allure dans la figure 2.12 présente la variation de taux de reconnaissance en fonction du nombre de gaussiennes. Le taux de reconnaissance est proportionnel au nombre de gaussiennes.

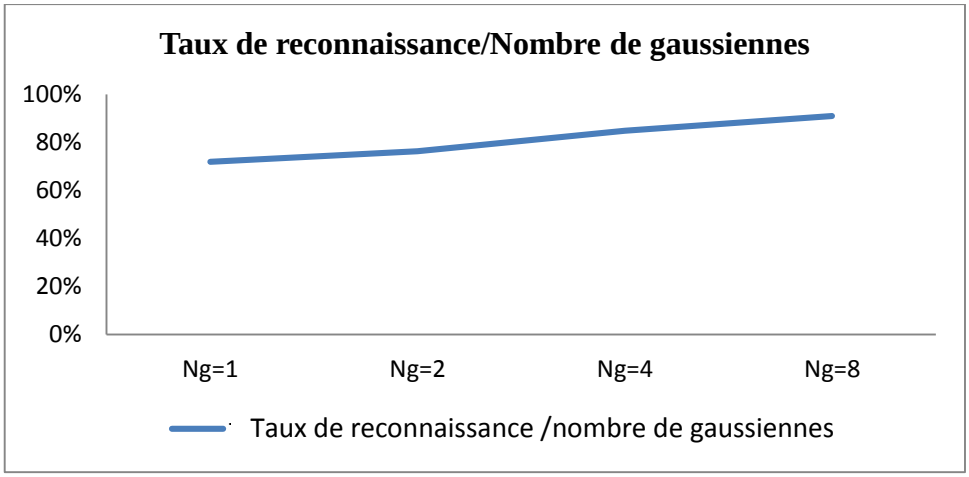


Figure 2.12 : Variation de taux de reconnaissance en fonction du nombre de gaussiennes

7.3. Système de reconnaissance automatique de Tamazight

7.3.1. Base d'apprentissage

La base d'apprentissage pour le dialecte Tamazight est composée des prononciations des chiffres isolés de 0 à 9. Ces chiffres sont prononcés par plusieurs personnes. La table 2.10 présente les caractéristiques en durée de la base d'apprentissage :

Durée de la base d'apprentissage	Nombres de personnes
1heures et 45min	14 masculins et 6 féminins

Table 2.10 : Caractéristiques de la base d'apprentissage

La table 2.11 présente la prononciation des chiffres de 0 à 9 en Tamazight ainsi que la transcription phonétique utilisée.

Chiffre	Transcription phonétique	Transcription Tifinaghe
0	I L E M	ⵉⵍⵎ
1	Y E N ou Y A N	ⵢⴻⵏ
2	S I N	ⵙⵉⵏ
3	C R A DD	ⵛⵔⴰⴷⴷ
4	K O Z	ⴽⵓⵣ
5	SS E M (S MM U S pour quelques régions)	ⵙⵙⴻⵎ
6	SS DD E SS	ⵙⵙⴷⴷⴻⵙⵙ
7	SS A	ⵙⵙⴰ
8	TT A M	ⵜⵜⴰⵎ
9	T Z A	ⵜⵣⴰ

Table 2.11 : Structure de la base d'apprentissage

7.3.2. Résultats

Le système de reconnaissance de Tamazight est testé en calculant le taux de reconnaissance sur un corpus de test constitué de 300 prononciations en introduisant les fichiers bruités. Les résultats obtenus sont donnés dans la table 2.12 :

Base de test	Résultat
300 prononciations différentes introduisant des fichiers audio plus bruités	T=90%

Table 2.12 : Résultats obtenus

La table 2.13 poursuit les résultats de comparaison entre MMC et la programmation dynamique :

	MMC	DTW
Taux de reconnaissance	90%	52%

Table 2.13 : Comparaison de taux de reconnaissance donné par MMC et DTW

L'efficacité de la programmation dynamique apparaît sur les fichiers audio non bruités. Cela est déduit d'après les résultats de comparaison dans la table 2.13, telle que la base de données de test utilisée est une base bruitée. Son inconvénient est que la durée d'exécution augmente proportionnellement avec la durée du fichier, ce qui influence sur le temps de reconnaissance. En comparaison avec la programmation dynamique, le modèle de Markov caché permet de modéliser un mot par une suite de phonèmes et une phrase par une suite de modèles de mots. Au niveau du temps d'exécution, l'algorithme de Viterbi implémenté pour la tâche de reconnaissance dans le MMC est plus rapide que la programmation dynamique (table 2.9). Ce dernier démontre ses points faibles en temps de reconnaissance pour les signaux de la parole continus.

7.3.3. Conclusion

Les résultats obtenus pour la reconnaissance de Darija et Tamazight en mots isolés sont satisfaisants au niveau du taux de reconnaissance, en comparaison avec la taille de la base d'apprentissage et la qualité des fichiers pris en compte dans la phase d'apprentissage. Dans le paragraphe suivant, nous présentons un système de reconnaissance automatique pour les chiffres isolés combinant Tamazight et Darija.

7.4. Combinaison de Tamazight et Darija dans un système de reconnaissance

Pour tester l'assemblage des dialectes marocains dans un seul système de reconnaissance, nous avons combiné les deux dialectes marocains Darija et Tamazight. À cause des caractéristiques phonatoires différentes de ces dialectes, nous avons séparé les deux bases d'apprentissage. Pour cela, la prononciation d'un chiffre doit contenir deux attributs : la correspondance en Darija et la correspondance en Tamazight. La table 2.14 présente les données d'apprentissage :

Chiffres en transcription numérique	Transcription phonétique
0	S I F R
0(1)	I LL E M
1	W A H D
1(1)	Y A N
2	J U J
2(1)	S I N
3	T L A TT A
3(1)	K R A D
4	R E B 3 A
4(1)	K U Z
5	KH E M S A
5(1)	SS MM U S
6	S TT A
6(1)	S D E SS
7	S E B 3 A
7(1)	S A
8	T E M N Y A
8(1)	T A M
9	T E S 3 OU D
9(1)	T Z A

Table 2.14 : Données de combinaison de Tamazight et Darija

Les résultats obtenus sont donnés dans la table 2.15 :

	Tamazight	Darija
Durée de la base d'apprentissage	1h45min	1h45min
Taux de reconnaissance	90%	92,33
Taux de rejet	3%	4,5%
Taux d'erreur	7%	3,17%

Table 2.15 : Résultats obtenus pour un système combinant Tamazight et Darija

7.5. Conclusion

Le taux de reconnaissance, généralement, dans un système RAP est proportionnel à la taille en durée et en diversité de la base d'apprentissage (durée de la base d'apprentissage dans la table 2.8 supérieur à celle dans la table 2.15 pour Darija). Plusieurs facteurs peuvent opter cette règle, à savoir la qualité des enregistrements. Les résultats obtenus démontrent la robustesse de la modélisation markovienne d'un côté, de l'autre côté, la reconnaissance automatique de la parole par la programmation dynamique est limitée pour les mots isolés à faible durée. Dans le paragraphe suivant, nous allons présenter une nouvelle approche pour la reconnaissance des chiffres enchainés en Tamazight.

8. Nouvelle Approche pour la reconnaissance des mots enchainés en Tamazight

Les chiffres en Tamazight se caractérisent par leur simplicité, ses chiffres enchainés sont reformulés à partir des chiffres isolés. En effet, Cette caractéristique simplifie la reconnaissance des chiffres continus à base des chiffres isolés. Dans ce sens, ce travail se focalise sur la création d'un système de reconnaissance des chiffres enchainés en Tamazight en se basant sur un corpus minimal d'apprentissage. Ce dernier est composé des chiffres isolés de 1 à 10 en ajoutant deux éléments importants qui jouent le rôle de coordinateurs pour formuler les chiffres continus. Ce travail se base sur un modèle de grammaire simple comportant les différents chiffres pouvant être constitués, les liens qui combinent entre ses différents chiffres sont les coordinateurs 'd' et 'id'. La figure 2.13 présente une simplification de cette caractéristique :

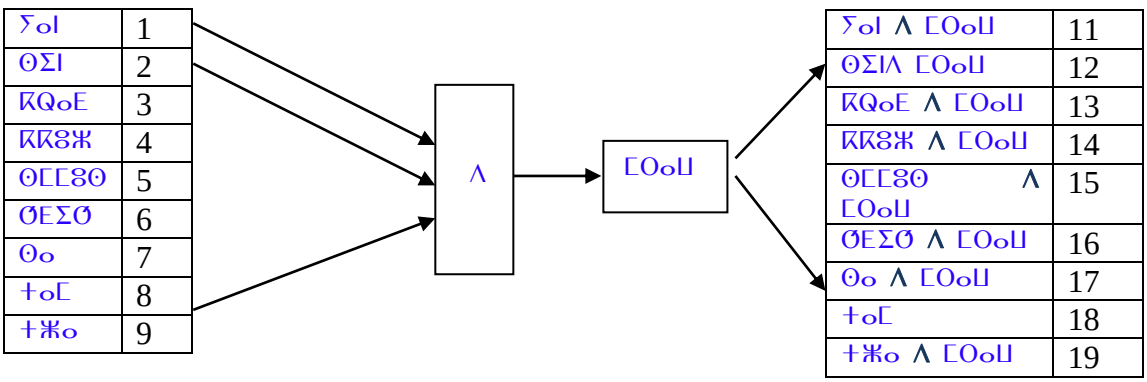


Figure 2.13 : exemples de formulation des chiffres enchainés en Tamazight

Dans la formulation des chiffres enchainés, on prend une base de données d'apprentissage contenant les chiffres isolés et les coordinateurs. Dans notre cas, nous avons pris les chiffres de 1 à 10 et 100 pour un système de reconnaissance limité à 199. De la même façon, on peut traiter le reste des chiffres enchainés.

8.1. Règles de construction des chiffres enchainés en Tamazight

Pour construire les chiffres enchainés en Tamazight, nous utilisons trois règles. La première concerne l'intervalle comportant les chiffres de 11 à 19, elle est représentée dans le paragraphe $8.3.1, la deuxième règle sur l'intervalle des chiffres de 20 à 99 et le troisième dans l'intervalle qui concerne les chiffres supérieur à 100. On définit les conventions suivantes :

- ✓ Les unités : les nombres qui représentent la partie unité des chiffres à titre d'exemple le chiffres 12 l'unité est 2.
- ✓ Les dizaines : le nombre qui représente le nombre de 10 dans un chiffre par exemple 20 la dizaine est 2.
- ✓ Les syntagmes : représentent les coordinateurs entre les unités et les dizaines, dans le dialecte Tamazight, ils sont représentés par la lettre '⵰' qui se prononce 'd'. Exemple : ⵙⵉⵏ ⵏ ⵜⵎⵎⵓⵔⵜ qui signifie un et dix (c'est à dire le chiffre 11).
- ✓ Les morphèmes : caractérisent la multiplication dans une expression des chiffres enchainés en Tamazight. Ils se mettent entre le nombre des dizaines et le chiffre dix, il s'écrit '⵵ⵏ' qui se prononce 'id'. Par exemple : ⵙⵙⵉ ⵵ⵏ ⵜⵎⵎⵓⵔⵜ qui signifie deux multiplié par dix c'est à dire le chiffre 20.

Ces différentes règles de construction des chiffres enchainés en Tamazight simplifient la tâche de reconnaissance, comme elles permettent d'élargir les intervalles de reconnaissance en se basant sur un corpus d'apprentissage minimal. Cela permet d'optimiser les ressources en quantité de stockage.

8.1.1. Construction des chiffres enchainés de 11 à 19

La construction des chiffres enchainés de 11 à 19 se base sur la prononciation des nombres de 1 à 10, en plus de syntagme 'd' qui permet de faire une addition pour construire le nombre final. En formule mathématique, le syntagme 'd' signifie l'addition, il permet d'ajouter les unités au chiffre dix pour construire le nombre désiré. Cette règle de construction peut être représentée par l'application bijective suivante :

$$f : [1,9] \rightarrow [11,19]$$

$$x \rightarrow x+10$$

La variable *x* a des valeurs dans l'intervalle [1,9], ce dernier comporte la prononciation des chiffres de 1 à 9. Le nombre résultant est calculé en ajoutant 10 au variable x. Il existe quelques cas particuliers en Tamazight qui dépendent généralement du contexte du chiffre. Parfois on a besoin de mettre le chiffre 10 ou le nombre d'unités au féminin. Par exemple :

ⵖⵙ ⴰ ⵎⵙⵓ : ici toutes les composantes du chiffre sont au masculin, il peut devenir ⵖⵙ ⴰ ⵎⵙⵓⵜ ou ⵖⵙⵜ ⴰ ⵎⵙⵓⵜ, ce cas particulier existe beaucoup dans la langue arabe et dépend du contexte du chiffre.

La table 2.16 présente quelques exemples qui résument cette règle de construction.

Premier chiffre	morphème	Deuxième chiffre	Chiffre résultant (Tifinaghe)	prononciation
ⵖⵙ	ⴰ	ⵎⵙⵓ	ⵖⵙ ⴰ ⵎⵙⵓ	*Yan d mraw (1+10) : f(x=1)=11*
ⵙⵉ			ⵙⵉ ⴰ ⵎⵙⵓ	*Sin d mraw(2+10) : f(x=2)=12*
ⴽⵕⴰⴹ			ⴽⵕⴰⴹ ⴰ ⵎⵙⵓ	*Krad d mraw(3+10) : f(x=3)=13*
ⵜⵣⴰ			ⵜⵣⴰ ⴰ ⵎⵙⵓ	*Tza d mraw(9+10) : f(x=9)=19*

Table 2.16 : Quelques exemples de règles de construction des chiffres de 11 à 19

8.1.2. Règles de construction pour l'intervalle de 20 à 99

Dans cet intervalle, les chiffres se composent soit de trois parties ou de cinq parties. Ils sont tous formulés à partir des chiffres isolés, en ajoutant les syntagmes 'd' et les morphèmes 'id'. Le morphème 'ⵙⴰ' se met entre le nombre des dizaines et le chiffre dix. La figure ci-dessous présente une simplification de cette règle.

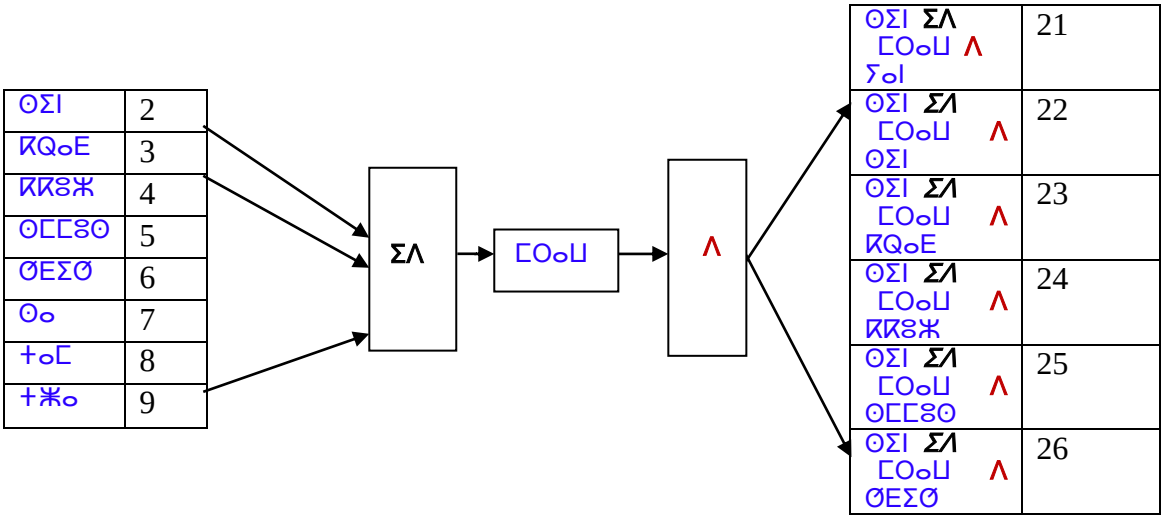


Figure 2.14 : Exemples de formulation des chiffres de 20 à 99

La règle de construction des chiffres de 21 à 99 consiste à commencer par le nombre de dizaines dans l'expression puis le morphème 'id' ensuite le chiffre dix et on met le syntagme 'd' et enfin le nombre d'unités. Cette règle peut être donnée par l'application suivante :

$$f : [2,9] \times [1,9] \rightarrow [21,99]$$

$$(x, y) \rightarrow x*10+y$$

La table 2.17 donne quelques exemples de formulation de ces chiffres :

Premier chiffre	Syntagme	deuxième chiffre	Morphème	3ème chiffre	Chiffre résultant (Tifinaghe)	Chiffre résultant
ⵙⵉⵏ	ⵙⴰ	ⵎⵕⴰⵡ	ⴷ	ⵢⴰⵏ	ⵙⵉⵏ ⵙⴰ ⵎⵕⴰⵡ ⴷ ⵢⴰⵏ	*Sin id mraw d yan(21)) (f(2,1)=2*10+1=21)*
ⴽⵔⴰⴹ					ⴽⵔⴰⴹ ⵙⴰ ⵎⵕⴰⵡ ⴷ ⵢⴰⵏ	*Krad id mraw d yan (31) (f(3,1)=3*10+1=31)*
ⵜⵣⵉⵜ					ⵜⵣⵉⵜ ⵙⴰ ⵎⵕⴰⵡ ⴷ ⵢⴰⵏ	*Tza id mraw d yan (91) (f(9,1)=9*10+1=91)*

Table 2.17 : exemples de formulation des chiffres de 21 à 99

8.1.3. Chiffre supérieur à 100

Pour les chiffres supérieur à 100, la règle consiste à commencer par le nombre 100 puis le morphème ‘d’ et on suit les règles décrites dans les paragraphes 8.1.1 et 8.1.2 pour le reste de l’expression. La figure 2.15 donne une illustration de cette règle.

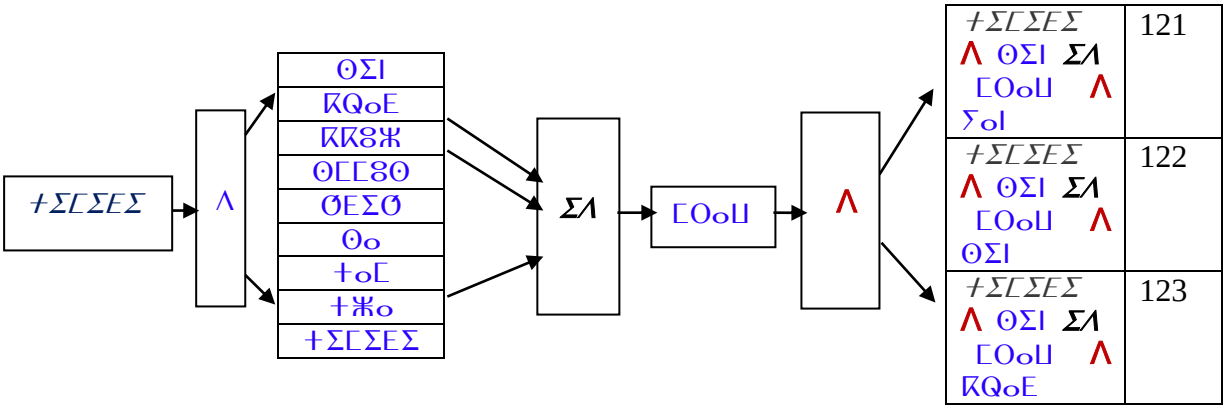


Figure 2.15 : Exemples de chiffres enchainés supérieurs à 100

Cette règle peut être représentée par l’application bijective suivante :

$$f : [0,9] \times [0,9] \times [0,9] \rightarrow [100,199]$$

$$(x, y, z) \rightarrow x*10+y+z+100$$

La table 2.18 présente quelques exemples de formulation de ces chiffres :

Premier nombre	Syntagme	2éme nombre	Morphème	3éme Nombre	Chiffre résultant (Tifinaghe)	Chiffres résultant
ⵜⵙⵙⵉⵙ	ⵏ	ⵎⵙⵏ	ⵙⵏ	ⵜⵎⴰⵡⴰⵢ	ⵜⵙⵙⵉⵙ ⵏ ⵎⵙⵏ ⵜⵎⴰⵡⴰⵢ	*Timidi dsin idmraw (120)) F(2,0,0)=100+2*20+0+0=120*
		ⵙⴰⵏ	ⵏ	ⵜⵎⴰⵡⴰⵢ	ⵜⵙⵙⵉⵙ ⵏ ⴽⵕⴰⵜ ⵏ ⵜⵎⴰⵡⴰⵢ	*Timidi dyan dmraw(111) F(1,1,0)=100+1*10+1=111*
		ⵜⵎⴰⵜ	ⵏ	ⵜⵎⴰⵡⴰⵢ	ⵜⵙⵙⵉⵙ ⵏ ⵜⵎⴰⵜ ⵏ ⵜⵎⴰⵡⴰⵢ	*Timidi dtza dmraw (119) F(1,9,0)=100+1*10+1+9+0=119*

Table 2.18 : Quelques exemples des chiffres enchainés supérieurs à 100

8.2. Expérimentation

8.2.1. Base d'apprentissage

La base d'apprentissage se compose de la prononciation des chiffres isolés de 1 à 10, le chiffre 100 puis les syntagmes et les morphèmes. Le système proposé se base sur un corpus d'apprentissage minimal, il permet de formuler les chiffres enchainés à partir des nombres isolés en se basant sur un modèle de grammaire simple. Ce dernier met en place les différentes phrases ou chiffres possibles à être reconnus par le système. L'utilisation de la transcription française de l'alphabet Tifinaghe permet aussi de faciliter le décodage de ces lettres. La liste des phonèmes utilisés ainsi que leurs transcriptions sont données dans la table 2.19. Dans la phase d'apprentissage le modèle de Markov caché à trois états est créé pour chaque phonème. En reconnaissance, on se base sur un graphe de recherche qui permet de parcourir les modèles de phonèmes selon le chemin qui maximise la probabilité d'observation. L'algorithme de Viterbi [25, 60, 61] est l'un des algorithmes qui permet de réduire les chemins à prendre en compte dans la recherche, il ne prend que le chemin qui maximise la probabilité d'observation. Cet algorithme est présenté en détail dans l'annexe 6.

Phonèmes en transcription française	Correspondance Tifinaghe
Y	ⵢ
A	ⴰ
N	ⵏ
S	ⵙ
I	ⵉ
K	ⴽ
R	ⵔ
U	ⵓ
Z	ⵣ
M	ⵎ
SS	ⵚ
DR	ⵕ
RR	ⵕ
ZZ	ⵥ
MM	ⵎⵎ
DD	ⴹ
D	ⴷ
KK	ⴽ
E	ⵄ
T	ⵜ
W	ⵡ

Table 2.19 : phonèmes Tifinaghe et leurs transcriptions françaises ainsi que le nombre d'occurrences dans la base d'apprentissage

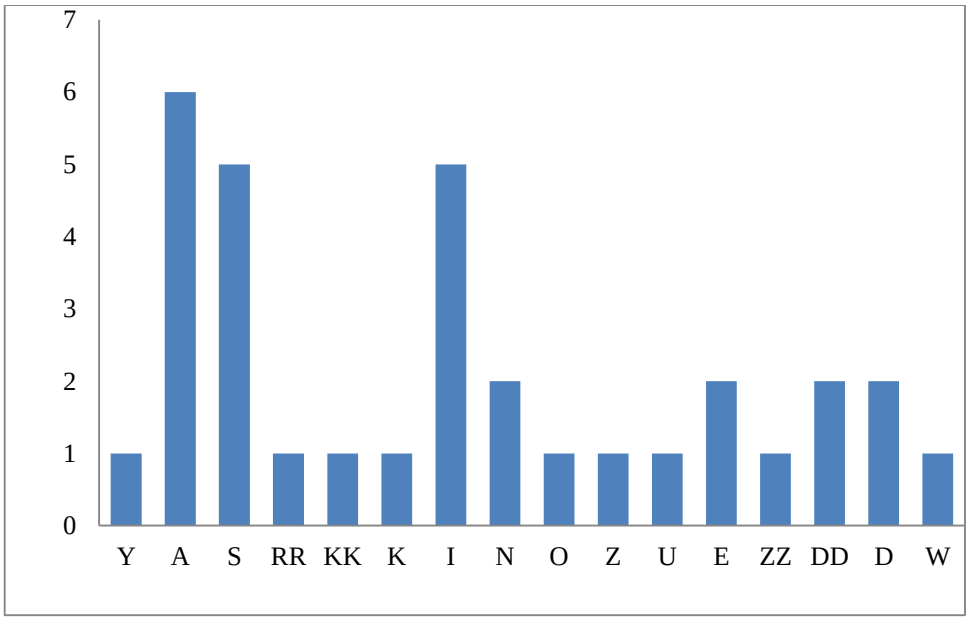


Figure 2.16 : statistiques des phonèmes Tifinaghe en transcription française dans la base d'apprentissage

Les caractéristiques temporelles de la base d'apprentissage sont données dans la table 2.20 :

Nombre de prononciations	Durée totale
Plus de 5192 prononciations pour chaque chiffre isolé	6,28 heures

Table 2.20 : caractéristiques temporelles de la base d'apprentissage

8.2.2. Résultats

Le système de reconnaissance réalisé est testé avec une base de données composée de 200 prononciations. Les résultats obtenus sont donnés dans la table 2.21 :

Base de test	Taux d'évaluation
200 prononciations	T=94,32% T_R=2,68% T_E=3%

Table 2.21 : résultats obtenus

La courbe dans la figure 2.17 présente la variation de taux de reconnaissance en fonction de la taille de la base de données d'apprentissage :

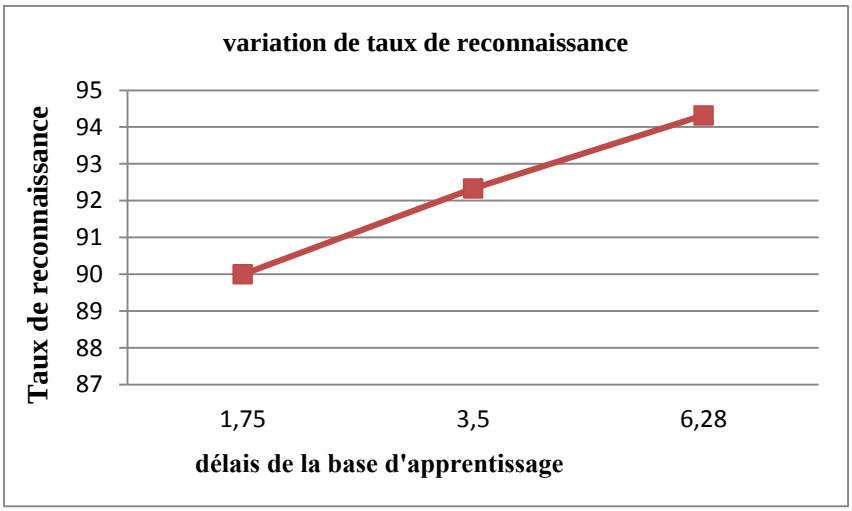


Figure 2.17 : Variation de taux de reconnaissance en fonction du délai de la base d'apprentissage

9. Conclusion

Dans ce chapitre, nous avons mis en place un système de reconnaissance des deux dialectes marocains Tamazight et Darija. Ce système est évalué en se basant sur le taux de reconnaissance. Dans ce sens, les résultats obtenus sont satisfaisants vue le nombre de locuteurs pris en compte dans la base d'apprentissage et voire aussi la qualité des enregistrements (bruit, microphone,…).

Les systèmes RAP peuvent être intégrés dans les systèmes avancés de dialogue homme-machines et les systèmes de sécurité. Vu l'importance de ces applications, nous nous sommes basés sur le système de reconnaissance réalisé pour Tamazight et Darija pour élaborer un système de sécurité applicable au Maroc. Ce système est fondé sur la validation de mot de passe et l'identification du locuteur. Il sera développé dans le chapitre suivant.

Chapitre 3

Reconnaissance et vérification du locuteur

1. Introduction

La sécurité est l'un des domaines qui évoluent proportionnellement au développement de l'informatique. Les empreintes digitales constituent la clé d'identification personnelle qui se base sur le traitement d'images. Le système doit comparer directement une empreinte parmi des milliers d'autres dans une base de données. A cause de la simplicité des images traitées, les résultats obtenus sont souvent corrects dans le sens d'identification. Plus que les images des empreintes sont incomplètes, plus le système nous montre ces points faibles au niveau de reconnaissance. Pour améliorer la reconnaissance, plusieurs travaux de recherche ont été lancés [62, 63, 64]. Parmi les systèmes d'identification existants, l'identification vocale, l'identification par utilisation de la cornée de l'œil et l'identification par les images personnelles. Ces techniques ont montré leur puissance en reconnaissance et dans l'identification au domaine criminel, elles ont été implantées dans les systèmes judicaires de plusieurs pays et ont été prises comme outil d'identification officiel.

L'un des plus importants outils d'identification est la vérification personnelle par la voix, c'est ce qu'on appelle l'identification du locuteur ou la vérification de locuteur. Ce système est basé sur les empreintes vocales. Celles-ci permettent de discriminer les personnes en se basant sur les caractéristiques de la parole produite. La différence intra-locuteurs des caractéristiques du conduit vocal permet de donner une caractéristique personnelle à la voix produite.

Dans ce chapitre, nous allons présenter les différentes techniques liées à la reconnaissance automatique du locuteur. On peut regrouper les techniques les plus utilisées selon les deux approches suivantes :

- ✓ Approche vectorielle : elle est basée sur la discrimination entre les vecteurs de paramètres acoustiques pour distinguer entre les locuteurs. Cette approche comporte la programmation dynamique et la quantification vectorielle.
- ✓ Approche statistique : elle est basée sur les lois de probabilité pour modéliser les locuteurs. Cette approche comporte le mélange de gaussiennes et le modèle de Markov caché.

Dans la suite de ce chapitre nous donnerons les différentes définitions liées à la reconnaissance, identification et vérification du locuteur, ainsi que les différentes approches

théoriques utilisées dans ce sens.

2. La reconnaissance automatique du locuteur

2.1. Généralités

La caractérisation automatique du locuteur est un vaste domaine dans lequel le système de reconnaissance a pour tâche d'extraire du signal de la parole les informations de nature à renseigner sur les spécificités d'un individu : identité, caractéristiques physiques, émotivité, état pathologique, particularités régionales, etc. Elle s'applique à différents thèmes de recherche traitant des informations extralinguistiques véhiculées par la voix tels que la classification d'individus, ou l'étude psychique ou physiologique d'une personne.

La Reconnaissance Automatique du Locuteur (RAL) [63, 64] est un sous problème de la caractérisation automatique du locuteur. Son objectif est de reconnaitre l'identité d'une personne à l'aide de sa voix. La variabilité de la parole entre locuteurs (variabilité interlocuteur) est l'essence même de la RAL. Sans cette variabilité, il serait impossible d'identifier une voix.

La RAL, contrairement à la reconnaissance automatique de la parole (RAP) s'intéresse tout particulièrement aux informations extralinguistiques véhiculées par un signal vocal (signal de parole). Pourtant, la RAL a très souvent bénéficié des avancées de la RAP. Ainsi, de nombreuses techniques ont été appliquées en RAP avant d'être adaptées au domaine de la RAL.

Les applications de la reconnaissance automatique du locuteur sont liées principalement aux problèmes d'authentification ou de confidentialité à savoir les informations liées aux comptes bancaires.

2.2. Niveau de dépendance de texte

Une première classification des systèmes de RAL repose sur le niveau de dépendance de texte. En premier lieu, on distingue généralement les systèmes dépendants du texte et des systèmes indépendants du texte. En mode dépendant du texte, la reconnaissance d'une personne est réalisée sur la base d'un message dont le contenu linguistique (mot de passe, phrase,…) est connu du système. En mode indépendant du texte, le système de reconnaissance n'a aucune connaissance sur le message linguistique prononcé par la personne.

Concernant le mode dépendant du texte, une terminologie plus fine peut être donnée à un système suivant l'application visée. Celle-ci est inspirée de la littérature :

- ✓ Systèmes à messages fixés : la personne doit prononcer un message qu'elle aura fixé au préalable (mots de passe personnalisés) ou qui sera imposé par le système.
- ✓ Systèmes à messages promptes : un message, différent à chaque nouvelle session de reconnaissance, est imposé par le système sous forme visuelle [65] ou auditive. Ces systèmes ont pour première motivation de se protéger des attaques de

personnes malveillantes (imposteurs) qui disposeraient d'un enregistrement de la voix d'une personne.

- ✓ Systèmes à unités segmentales fixées : la personne doit prononcer un message comportant soit une séquence de mots (séquence de chiffres), soit des traits phonétiques (séquence de phonèmes) connu du système.

La connaissance *a priori* partielle ou totale du message prononcé par la personne rend généralement les systèmes dépendants du texte plus performants que les systèmes indépendants du texte. En mode dépendant du texte, les systèmes s'affranchissent du problème de la variabilité linguistique.

2.3. Différentes tâches de RAL

La vérification automatique du locuteur et l'identification automatique du locuteur sont les tâches principales du domaine de la RAL. Plus récemment, les besoins applicatifs ont fait naitre de nouvelles tâches comme l'indexation par locuteurs de flux audio [66] ou le suivi de locuteur [67] ou de nouvelles variantes telles que la détection d'un locuteur dans une conversation [68].

2.3.1. Identification Automatique du Locuteur

L'identification Automatique du Locuteur (IAL) est le processus qui consiste à déterminer, parmi une population de locuteurs connus, la personne ayant prononcé un message donné.

D'un point de vue schématique (figure 4.1), une séquence de parole est donnée en entrée du système d'IAL. Pour un locuteur connu par le système, la séquence de parole est comparée à une référence caractéristique du locuteur. L'identité du locuteur dont la référence est la plus proche de la séquence de parole est donnée en sortie du système d'IAL [69,70].

Deux modes sont proposés en identification automatique du locuteur : l'identification en ensemble fermé pour lequel on suppose que la séquence de parole est effectivement prononcée par un locuteur connu du système et l'identification en ensemble ouvert pour lequel le locuteur peut ne pas être connu. En mode 'ensemble ouvert', le système d'IAL doit décider de la fiabilité de son jugement en acceptant ou rejetant l'identité qu'il a trouvée.

De par son principe, déterminer une identité parmi les identités potentielles, les performances des systèmes d'identification automatique du locuteur se dégradent généralement au fur et à mesure que la population de locuteurs augmente.

En IAL, les applications sont peu nombreuses. On peut retenir, par exemple, l'utilisation d'un système d'IAL en vue de faciliter l'adaptation au locuteur des systèmes de reconnaissance automatique de la parole. Par ailleurs, il peut être intéressant pour les applications commerciales d'associer un même mot de passe pour une petite population de locuteurs (membre d'une famille d'une société). Dans une telle situation, un système d'identification automatique du locuteur en ensemble ouvert et dépendant du texte peut être utilisé pour contrôler l'accès à des données sensibles dans un réseau ou dans un bâtiment [67].

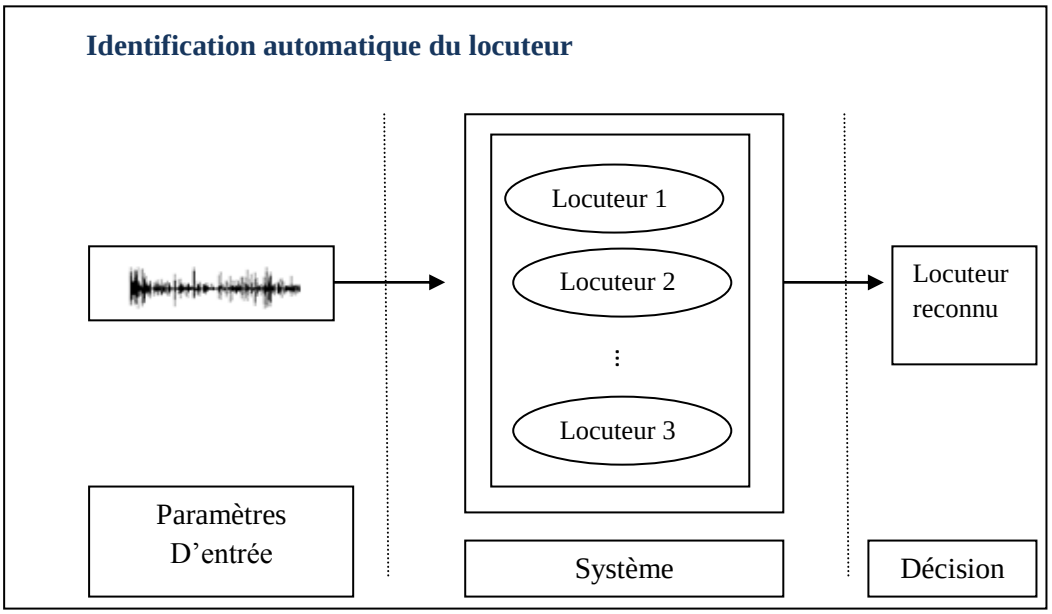


Figure 3.1 : principe de base d'identification automatique du locuteur

2.3.2. Vérification Automatique du Locuteur

La vérification Automatique du locuteur (VAL) est le processus décisionnel permettant de déterminer, au moyen d'un message vocal, la véracité de l'identité revendiquée par un individu (figure 3.2). L'identité ainsi que le message vocal constituent les deux entrées du système de VAL. L'identité, nécessairement connue du système, désigne automatiquement la référence caractéristique d'un locuteur. Une mesure de similarité est calculée entre cette référence et le message vocal puis comparée à un seuil de décision. Dans le cas où la mesure de similarité est supérieure au seuil, l'individu est accepté. Dans le cas contraire, l'individu est considéré un imposteur et rejeté [71, 72].

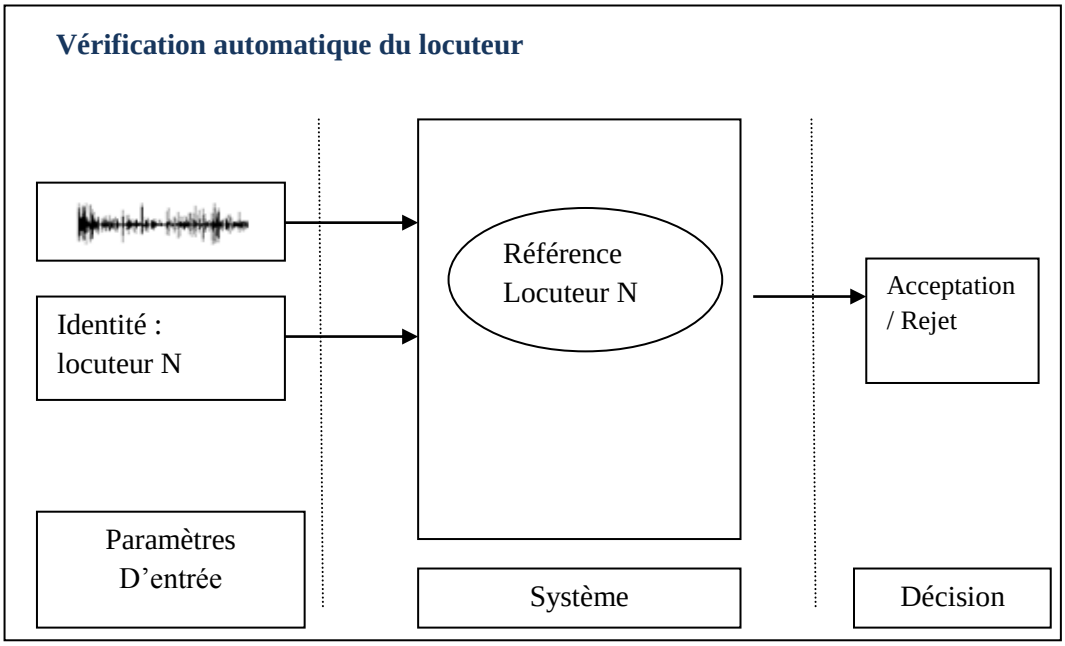


Figure 3.2 : principe de base de vérification automatique du locuteur

Les applications de VAL sont multiples et principalement commerciales [73] :

- ✓ Serrures vocales pour le contrôle d'accès à des locaux ;
- ✓ Authentification pour l'accès à distance à des données sensibles ou à des services spécifiques à travers le réseau téléphonique (consultation ou transaction de boites vocales, télé-achat, etc.) ;
- ✓ Protection de matériel contre le vol (téléphones portables, voitures, etc.) ;
- ✓ Incarcération à domicile nécessitant une authentification régulière du prévenu.

2.3.3. Détection de locuteurs

La détection de locuteurs dans un flux audio est une variante de la vérification automatique de locuteur [67, 74,75]. Sa particularité est de considérer un flux audio composé de séquences de parole produites par plusieurs locuteurs (conversations multi locuteurs, programme télévisé, etc.). Dans ce contexte, la tâche de détection consiste à déterminer si un locuteur donné intervient ou non dans le document audio. Dans le cas d'un flux de la parole mono-locuteur la tâche de détection se résume à la tâche de vérification.

La tâche de détection est évidemment motivée par les instances militaires ou judiciaires. Néanmoins, elle demeure très intéressante dans le domaine de l'indexation de documents audio pour laquelle la détection d'un locuteur connu peut permettre de cibler plus facilement un document audio particulier (séquence d'un journal télévisé ou d'une émission radio).

2.3.4. Indexation par locuteur et ses variantes

La tâche d'indexation automatique de locuteur consiste à cibler les interventions des locuteurs dans un flux audio (figure 3.3). En d'autres termes, indexer un document audio en locuteurs revient à indiquer à quel moment un individu prend la parole et qui est cet individu. La seule entrée d'un système d'indexation est le document audio à indexer. Aucune information n'est donnée au système concernant le nombre de locuteurs présents dans le document ou leur identité. Contrairement aux systèmes d'IAL ou de VAL, les systèmes d'indexation ne détiennent pas de référence pour les locuteurs présents dans un document audio. Leur principe repose généralement sur une phase de segmentation 'aveugle' en locuteurs suivie d'une phase de regroupement. Un système d'IAL permet d'identifier les différents locuteurs existants dans le document audio. la sortie d'un système d'indexation ressemble à la séquence suivante : le locuteur A est intervenu aux instants t_1, t_2, t_4, le locuteur B aux instants t_3, t_5.

La tâche de suivi de locuteurs peut être considérée comme une version simplifiée de l'indexation par locuteur d'un flux audio (figure 3.4). Le principe reste le même : déterminer les interventions d'un ou plusieurs locuteurs, appelés locuteurs cibles, dans un flux audio. La simplification réside dans le fait que le système de suivi de locuteur connait nécessairement les locuteurs présents dans le document à indexer. Il possède une référence caractéristique pour chacun des locuteurs. Malgré cette simplification, le suivi de locuteurs reste une tâche très complexe. Trois grandes approches sont recensées dans la littérature :

1. Une segmentation 'aveugle' en locuteurs, identique à celle employée pour l'indexation par locuteur d'un flux audio, elle est appliquée sur le signal de test. Les segments résultats de la segmentation sont soumis à un système de VAL classique afin de déterminer les segments appartenant effectivement au locuteur cible [76].

2. Le signal de test est découpé en une suite de blocs de trames, de taille fixe, sur ces blocs on applique un système de VAL. Un processus de décision, à base de seuils, permet en phase finale d'accepter ou de rejeter les blocs appartenant au locuteur cible [76].

3. La troisième approche est similaire à la précédente sauf pour le processus de décision. Dans ce cas, la décision repose sur un HMM ergodique composé d'états correspondant au locuteur cible. Dans ce sens, deux modèles de ce type sont mis en place : un modèle générique de parole et un modèle générique de non parole (silence, bruit…) [75].

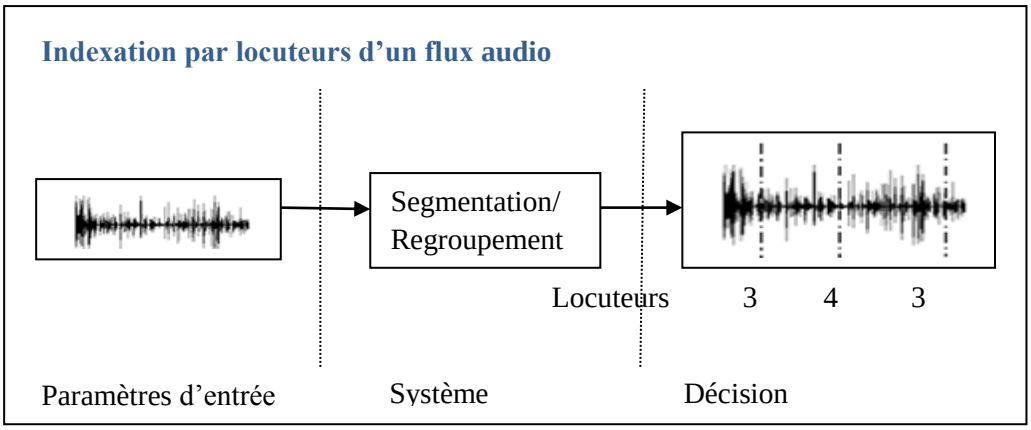


Figure 3.3 : La tâche d'indexation par locuteurs d'un flux audio

Les systèmes d'Indexation Automatique par Locuteur (IAL) d'un flux audio sont principalement utilisés pour le traitement des bases de données audio (recherche de séquences d'émissions télévisées ou radiophoniques par le suivi du présentateur, estimation du temps de parole de chaque intervenant lors de débat, etc.…). D'autres applications sont envisageables comme la recherche des messages par locuteur sur un répondeur téléphonique ou sur une boite vocale.

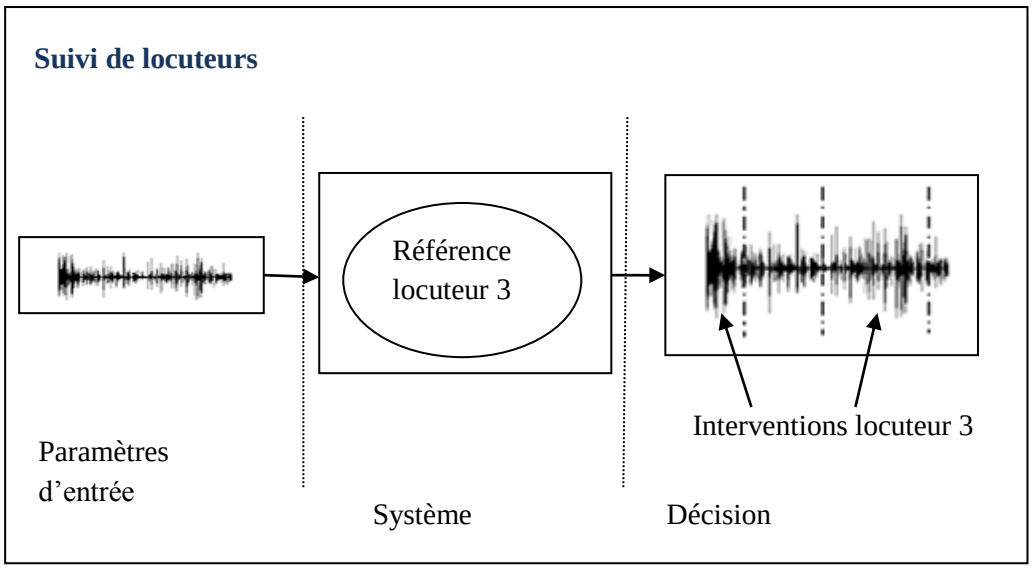


Figure 3.4 : Tâche de suivi de locuteur

2.3.5. Applications criminalistiques

Parmi les utilisations les plus importantes de la RAL, on trouve les domaines judicaires ou criminalistiques [63]. Il s'agit de rechercher un individu parmi une population de suspects potentiels (tâche d'IAL) ou encore de comparer un enregistrement vocal issu d'une écoute téléphonique à la voix d'un suspect potentiel (tâche de VAL).

Dans ce contexte, il est important de souligner que la voix est très souvent assimilée à une empreinte vocale au même titre que les empreintes digitales ou génétiques et peut constituer une preuve dans une procédure pénale, ceci dépend du règlement judiciaire de chaque pays. Ce terme d'empreinte vocale est une aberration sachant que la voix ne possède pas de caractéristiques qui peuvent la rendre unique [63].

2.4. Mise en place d'un système de RAL

Le système de reconnaissance automatique du locuteur pour une application donnée se décompose en deux phases distinctes. La première phase est nécessaire à la construction de références ou modèles pour chaque locuteur connu du système i.e. de chaque client ou utilisateur de l'application. Elle consiste à collecter, auprès de ces clients, des signaux de parole dits d'apprentissage, lors de sessions d'enrôlement. La deuxième phase est la phase de reconnaissance à proprement parler qui consiste, pour un client, à se présenter devant le système de RAL. Cette phase est dite la phase de test.

2.5. Problèmes rencontrés en RAL

Le signal de la parole est un signal très complexe, il constitue un mélange qui combine l'information linguistique, l'information caractéristique de locuteur et l'information relative au matériel utilisé pour la transmission ou l'enregistrement du signal, etc. En outre, le signal de la parole est très redondant. Cette caractéristique est d'ailleurs reconnue pour faciliter la communication entre deux personnes dans un environnement très bruyant. Par ces différents aspects, le signal de parole présente une très grande variabilité.

La puissance d'un système de reconnaissance automatique du locuteur repose essentiellement sur la variabilité interlocuteur i.e., la disposition du signal de la parole à faire varier entre individus. Néanmoins, le signal de parole présente d'autre variabilité qui rend difficile la tâche de reconnaissance, telles que la variabilité intra locuteur ou la variabilité due au matériel ou l'ambiance du travail. Par ailleurs, les systèmes de reconnaissance automatique du locuteur doivent faire face à d'autres difficultés liées davantage au domaine applicatif, comme l'utilisation des systèmes dans des conditions difficiles, les tentatives d'imposture, etc.

2.5.1. Variabilité due au locuteur

Le signal de la parole est variable pour deux individus, il varie également pour un même locuteur. Cette variabilité intra-locuteur est induite par l'évolution naturelle ou volontaire de la voix d'une personne. Cette évolution peut être :

- ✓ L'état pathologique d'une personne provoque des altérations momentanées dans sa voix. Dans ce sens, la voix d'une personne peut évoluer entre le début et la fin de la journée (fatigue, changement climatique, etc.). D'autre part, un individu ne peut pas répéter deux phrases consécutives de la même façon. Une légère variation est toujours observée. Finalement une personne a la possibilité de modifier volontairement sa voix.

- ✓ Un individu, en interaction avec un système, se modifie au fur et à mesure de son utilisation. Il devient plus en plus confiant et sa voix évolue dans ce sens.
- ✓ La voix d'une personne varie en fonction de son âge.

La variabilité intra-locuteur pose le problème de la représentativité des signaux de parole collectés lors des sessions d'enrôlement au sein d'un système de reconnaissance automatique du locuteur. Des travaux ont montré que les performances d'un système sont fortement corrélées au temps qui sépare les sessions d'enrôlement et les tests [77]. Plus ce temps augmente, plus les performances se dégradent. Néanmoins, même les variations à court terme (émotion, état pathologique) peuvent être très préjudiciables aux systèmes de reconnaissance automatique du locuteur.

2.5.2. Variabilité due au matériel

Le signal de parole est porteur d'informations qui caractérisent le matériel utilisé lors de sa capture (ex : microphone, enregistrement téléphonique), de sa transmission (ex : lignes téléphoniques, air ambiant) et de son enregistrement (ex : microphones, convertisseurs). Ces informations apparaissent le plus souvent sous la forme de déformations/dégradations du signal de la parole. Ces déformations sont différentes selon le type de matériel utilisé.

De nombreux travaux expérimentaux ont montré que la variation du matériel entre les phases d'apprentissage et de test, sont à l'origine de graves dégradations des performances [78].

2.5.3. Robustesse en environnement difficile

Le bruit est l'un des contraintes qui menace la qualité d'un système de reconnaissance automatique de locuteur. En effet, le bruit existe et ne peut pas être séparé d'un signal de parole. Les systèmes de reconnaissance automatique de locuteur doivent renforcer leur robustesse au bruit ambiant. En effet, d'une manière similaire à la variabilité intra-locuteur ou à la variabilité due aux changements de matériel, la variabilité du niveau de bruit entre apprentissage et test peut susciter une baisse de performances des systèmes de reconnaissance automatique du locuteur.

2.5.4. Tentatives d'imposture

Selon l'application visée, un système de reconnaissance automatique de locuteur peut faire l'objet d'attaque d'individus imitant l'identité d'une autre personne. Ces attaques peuvent, par exemple, avoir pour dessein des transactions frauduleuses sur le compte bancaire d'un client ou l'accès à des données confidentielles. Un système de reconnaissance automatique du locuteur doit par conséquent être robuste face à de telles attaques [79].

Dans un contexte judicaire, le système de reconnaissance automatique du locuteur peut être soumis à des locuteurs non-coopératifs i.e. des locuteurs qui ne désirent pas être reconnu par le système. Dans ce cas de figure, les locuteurs tentent fréquemment de transformer leur voix.

2.5.5. Contraintes imposées par le domaine applicatif

Les domaines d'application et plus particulièrement commerciales imposent des contraintes fortes quant à l'utilisation d'un système de reconnaissance automatique du locuteur. Néanmoins, les sessions d'enrôlement doivent être peu contraignantes pour les clients d'une application. Dans ce sens, elles sont généralement peu nombreuses et peu espacées dans le temps. Aussi, la quantité de signaux de parole collectés lors de ces sessions s'avère insuffisante pour une bonne estimation des modèles clients.

D'autre part, lors des sessions d'enrôlement, les conditions d'utilisation du système (réseau téléphonique, microphone direct), sont peu variées. Aussi peu de variabilité due au matériel est introduite dans les signaux d'apprentissage.

La figure 3.5 présente une structure générale d'un système de reconnaissance automatique du locuteur. En effet, un système de RAL comprend deux phases importantes. La première phase consiste à entrainer une base de données audio sur un ensemble de clients ou utilisateurs de l'application. Cette phase permet d'obtenir un modèle caractérisant un locuteur avec des paramètres caractérisant sa voix. La deuxième phase concerne la phase de décision ou la phase de test. Elle consiste à tester le système de reconnaissance automatique de locuteur sur une base de données qui peut contenir la même structure de données selon le type de système visé.

3. Techniques associées à la reconnaissance automatique du locuteur

Un système de reconnaissance automatique du locuteur quel que soit son type caractérise l'enchainement de trois processus principaux qui sont : la paramétrisation, la reconnaissance et la décision. Contrairement au processus de paramétrisation (généralement commune à d'autre système comme la reconnaissance automatique de la parole), les principes de la paramétrisation et la reconnaissance sont liés à la tâche visée. Le processus de reconnaissance est différent selon qu'il repose sur la modélisation des caractéristiques connues par le système (modèle client pour les tâches de l'identification automatique du locuteur et la vérification automatique du locuteur) ou non (indexation du locuteur dans le flux audio). Dans le paragraphe suivant, nous nous intéresserons aux différentes étapes de création d'un système de reconnaissance automatique du locuteur. Nous commençons par la phase de paramétrisation qui est similaire en quelque sorte au cas de paramétrisation dans le cas de reconnaissance automatique de la parole. Puis, nous donnerons quelques techniques de modélisation et d'apprentissage pour la RAL. Ces techniques se diffèrent selon le type du système visé. La figure 3.5 présente la structure générale d'un système de reconnaissance du locuteur.

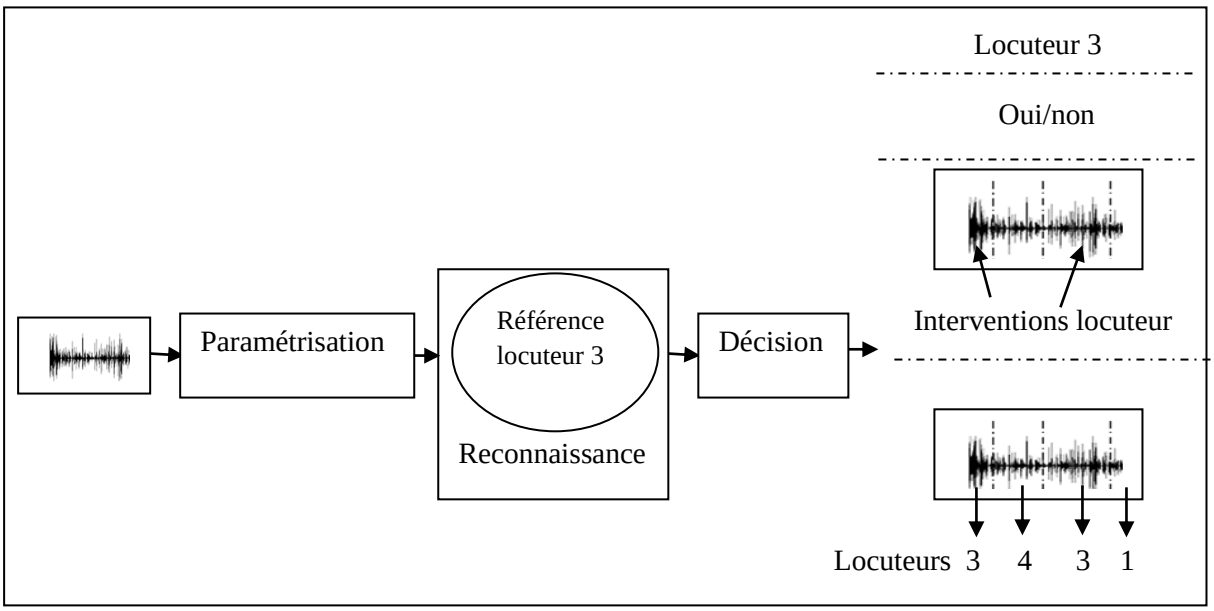


Figure 3.5 : structure d'un système de reconnaissance automatique du locuteur

3.1. Paramétrisation acoustique pour la reconnaissance automatique du locuteur

La paramétrisation, comme c'est le cas pour la reconnaissance automatique de la parole, consiste à extraire du signal de la parole les informations pertinentes en vue de la reconnaissance. Le signal de parole, de par sa complexité (multitude d'informations et redondance), ne peut être exploité directement [9,19]. Une représentation simplifiée du signal de parole est par conséquent nécessaire. Cette représentation repose généralement sur des vecteurs de paramètres acoustiques, calculés périodiquement sur le signal de parole.

La première étape, comme elle est expliquée dans le chapitre 2 de ce mémoire, consiste à décomposer le signal de parole, à des cadence régulières (généralement toutes les 10 millisecondes), en trames de signal (d'une longueur variante généralement de 20 à 30 ms). Un traitement particulier est ensuite appliqué à ces trames afin de produire les vecteurs de paramètres acoustiques.

Les travaux de recherche sur le traitement du signal proposent un grand nombre de traitements selon la nature des informations à extraire du signal de parole. IL existe généralement trois grandes classes de paramètres : les paramètres de l'analyse spectrale, les paramètres prosodiques et les paramètres dynamiques. Néanmoins, d'autres classifications sont envisageables. Par exemple, de séparer les traitements suivant qu'ils s'intéressent ou non aux informations de nature statique ou dynamique véhiculées par le signal de parole.

3.1.1. Paramètres d'analyse spectrale

L'analyse spectrale est l'analyse la plus utilisée en reconnaissance automatique du locuteur. Les paramètres qui en découlent sont généralement représentatifs des caractéristiques physiques de l'appareil phonatoire de chaque individu. En effet, le conduit vocal constitue l'un des organes de l'appareil phonatoire qui permet de déterminer la forme particulière du signal produit.

Dans ce sens, plusieurs paramètres ont été étudiés dans la littérature. On trouve une description détaillée dans [80,81]. Nous donnons ici les paramètres les plus utilisés en reconnaissance automatique du locuteur :

- ✓ Les coefficients issus d'une analyse par prédiction linéaire [82] : LPCC (Linear Predictive Cepstral Coefficients) ou LPC (Linear Predictive Coefficients).
- ✓ Les coefficients cepstraux issus d'une analyse en banc de filtres : LFSC (Linear Frequency Spectral Coefficients) ou MFSC (Mel Frequency Spectral Coefficients)[83].
- ✓ Les coefficients cepstraux issus d'une analyse en banc de filtres : LFCC (Linear Frequncy Cepstral Coefficients) ou MFCC (Mel Frequency Cepstral Coefficients)[84].

3.1.2. Paramètres prosodiques

Les paramètres prosodiques présentent le style d'élocution d'un locuteur : vitesse d'élocution (débit), durée et fréquences des pauses, ainsi que les caractéristiques de la source glottale (fréquence fondamentale, énergie, taux de voisement).

Néanmoins, ces paramètres, notamment la fréquence fondamentale et ses variations [85], ne sont pas suffisants pour être utilisés dans un système de reconnaissance automatique du locuteur. Ils sont généralement associés aux paramètres de l'analyse spectrale pour améliorer les performances du système de reconnaissance automatique du locuteur.

3.1.3. Paramètres dynamiques

L'information dynamique circulée par le signal de la parole est une source potentielle d'information pour la caractérisation du locuteur, qui reste encore mal exploitée par les systèmes de reconnaissance automatique du locuteur.

Les paramètres dynamiques les plus répandus demeurent les coefficients dérivées des vecteurs acoustiques, appelés aussi coefficients Delta (la dérivée première) et Delta-Delta (la dérivée seconde) [82]. Ils existent d'autres paramétrisations dans la littérature pour exploiter l'information dynamique dans un signal de la parole telles que l'utilisation des composantes principales temps-fréquence (TFPC : Time Frequency Principal Components), la concaténation de trames successives du signal [86].

3.2. Modélisation des mesures dans la reconnaissance automatique du locuteur

Le processus de RAL s'appuie généralement sur une modélisation des caractéristiques de chaque locuteur connu du système (modèle de locuteur ou modèle client). Cette modélisation est réalisée à partir des données d'apprentissage. Une mesure de vraisemblance entre le modèle client et un signal de la parole permet de décider sur l'acceptation ou le rejet du locuteur.

On distingue quatre grandes approches pour la construction du modèle client : les approches vectorielles, statistiques basées sur le modèle de Markov caché, prédictives et

connexionnistes. Nous présentons par la suite chacune de ces approches et les techniques qui leur sont associées.

3.2.1. Approche vectorielle

L'approche vectorielle se base sur un ensemble de vecteurs paramètres relatifs à chaque client ou utilisateur du système [87]. Ils sont obtenus à partir de la phase de paramétrisation des signaux d'apprentissage. Ces ensembles de vecteurs acoustiques sont présentés dans l'espace acoustique. La reconnaissance à base de cette technique, consiste à calculer la distance euclidienne entre les vecteurs acoustiques issus des signaux de test et les vecteurs des signaux de références. Cette approche comporte deux techniques : la programmation dynamique et la quantification vectorielle.

3.2.1.1. Programmation dynamique

La programmation dynamique (Dynamics Time Warpping) [15, 17], comme est le cas pour la reconnaissance automatique de la parole, consiste à aligner temporellement deux séquences de vecteurs acoustiques : vecteurs de paramètres de test avec une séquence de vecteurs d'apprentissage. Dans ce cas de figure, le modèle client est tout simplement un ensemble de vecteurs acoustiques qui correspondent au paramètres issus de son enregistrement. La distance calculée est moyennée sur l'ensemble de séquences. La programmation dynamique est utilisée exclusivement pour le système de reconnaissance automatique du locuteur dépendant de texte. Cet algorithme est sensible à la qualité d'alignement et notamment au choix du point de départ (Figure 3.6).

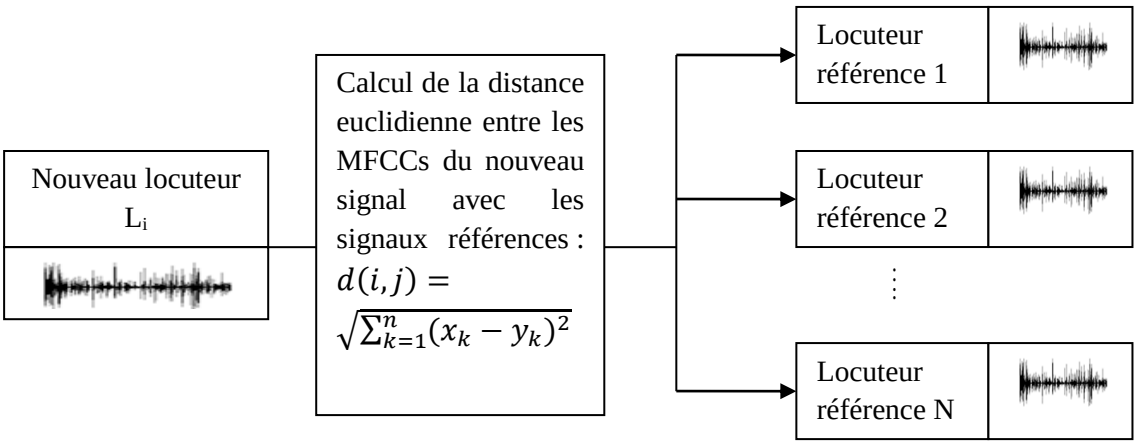


Figure 3.6 : Principe de la programmation dynamique pour la reconnaissance du locuteur

La décision sur l'identité du nouveau locuteur L_i est faite en prenant le signal référence qui minimise la distance d, soit :

$$L_i = argmind(s_i, r_j) \qquad 3.1$$

Avec s_i : signal produit par le locuteur inconnu L_i.

r : signaux références (1<j<N)

3.2.1.2. Quantification vectorielle

La quantification vectorielle (Vector Quantisation : VQ) [49,87] repose sur le principe de partitionnement de l'espace des données acoustiques en sous espaces. Chaque sous espace est représenté par un vecteur centroïde. Ce dernier représente l'ensemble de vecteurs acoustiques qui composent le sous espace. Le modèle du locuteur est un ensemble de centroïdes, appelé dictionnaire de quantification.

La reconnaissance dans le cas de quantification vectorielle consiste à calculer la distance entre un vecteur de test et tous les centroïdes du dictionnaire de quantification. La distance finale est obtenue en moyennant les distances minimales attribuées à chacun des vecteurs de test.

La quantification vectorielle s'applique dans le cas de reconnaissance du locuteur dépendant et indépendant du texte. La performance du système de reconnaissance dépend de la taille du dictionnaire : plus la taille augmente, meilleurs sont les performances. La figure 3.7 montre un exemple de quantification vectorielle, les trois points en rouge présentent le dictionnaire. Le calcul de la distance euclidienne entre un point et chaque centroïde du dictionnaire permet de décider sur la classe d'appartenance.

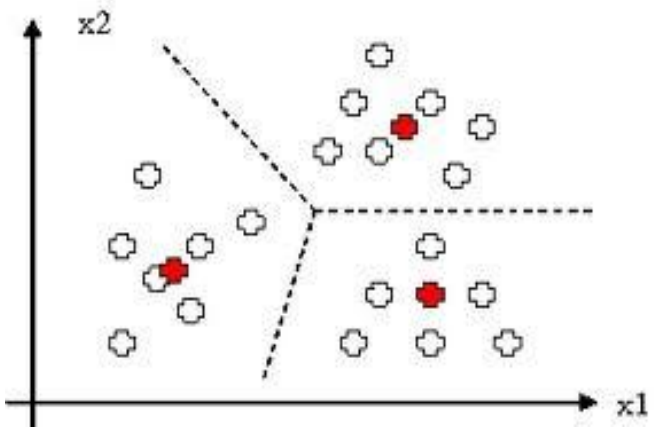


Figure 3.7 : exemple de quantification vectorielle à trois centroides

La quantification vectorielle se base sur le calcul d'une distance entre un vecteur X et son représentant noté $\hat{X}$ selon la relation :

$$d(X,\hat{X}) = \frac{1}{K}\sum_{k=1}^{K}(x_k - \hat{x}_k)^2 \quad \textbf{3.2}$$

Avec :

K : dimension de l'échantillon X.

L'apprentissage se fait par l'algorithme LBG (Voir l'annexe) et vise la division de l'espace des vecteurs acoustiques en des partitions non chevauchées. La génération du dictionnaire est donnée par la recherche d'un partitionnement qui minimise la distorsion moyenne des données d'apprentissage (les paramètres acoustiques). L'ensemble des *clusters* obtenus est appelé le dictionnaire des vecteurs qui représente un seul locuteur.

3.2.2. Approche statistique

L'approche statistique en reconnaissance automatique du locuteur permet de présenter une séquence de vecteurs issus de la phase de paramétrisation par des statistiques à long terme (représentation multigaussiennes). Ainsi, les paramètres du spectre représentent le modèle du locuteur. Dans la phase de reconnaissance, le spectre moyen issu des paramètres de test est comparé avec le spectre moyen issu de l'apprentissage. Par la suite, les statistiques du seconde ordre ont été introduites, celles-ci permettent d'introduire la variation des paramètres acoustiques (vecteur des moyennes et matrice de covariance).

Considérons une population de locuteurs i=1…N avec M_i la référence associée à chaque locuteur *i* [88]. L'identité retournée M, présente dans le signal X, est alors celle qui maximise la probabilité :

$$M = \underset{i}{argmax} P(M_i|X) \quad \textbf{3.3}$$

Sans informations *a priori* sur l'apparition des locuteurs $P(M_i)$, et en appliquant la règle de Bayes la relation 3.3 devient :

$$M = \underset{i}{argmax} P(M_i|X) = \underset{i}{argmax} \frac{P(X|M_i).P(M_i)}{P(X)} = \underset{i}{argmax} P(X|M_i)$$

Où $P(X \mid M_i)$ est la fonction de vraisemblance du locuteur *i* qui approxime la densité de probabilité des observations du locuteur *i*.

3.2.2.1. Mélange de gaussiennes

Le modèle de mélange de gaussiennes, comme est le cas pour la reconnaissance vocale, consiste à représenter un ensemble de vecteurs acoustiques d'apprentissage par un mélange de gaussiennes i.e. une somme pondérée de *M* distributions gaussiennes multidimensionnelles, chacune est caractérisée par un vecteur moyen et une matrice de covariance. Lors de l'apprentissage, les paramètres des modèles clients (vecteur moyen x_i, matrice de covariance Σ_i, pondération p_i de chaque distribution gaussienne) sont généralement estimés à l'aide de l'algorithme EM (Expectation-Maximisation) [89,90] couplé à l'approche par Estimation du Maximum de Vraisemblance (EMV) [90].

Lors de la reconnaissance, la mesure de similarité entre un modèle client et une séquence de vecteurs de test repose à nouveau sur l'approche EMV.

La vraisemblance pour qu'un vecteur de test, y_t, soit produit par le mélange de gaussienne χ s'exprime par:

$$L(y_t|\chi) = \sum_{i=1}^{M} p_i . L_i(y_t) \quad \textbf{3.4}$$

$$L_i(y_t) = \frac{1}{(2\pi)^{D/2}|\sum_i|^{1/2}} \, exp\left\{-\frac{1}{2}(y_t - \bar{x}_i)^T(\sum_i)^{-1}(y_t - \bar{x}_i)\right\}$$

Où p_i, $\bar{x}_i$ et $\sum_i$ représentent, respectivement, le poids, le vecteur des moyennes (de dimension D) et la matrice de covariance (de dimension D*D) de la $i^{ème}$ distribution gaussienne.

Par les performances qu'ils obtiennent, les mélanges de gaussiennes sont considérés comme la modélisation 'état de l'art' des systèmes de reconnaissance automatique du locuteur en mode indépendant de texte. L'inconvénient majeur de cette technique est la quantité de signaux d'apprentissage requise pour une bonne estimation des paramètres des modèles.

Pour créer un modèle statistique du locuteur, il est nécessaire de déterminer les paramètres de ce modèle $(\omega_i, \mu_i, \Sigma_i)$ [88]. Cette étape est généralement réalisée à partir d'un jeu de données dit d'apprentissage. Un algorithme est utilisé pour estimer ces paramètres en maximisant un critère choisi, vis à vis des données d'apprentissage. Le critère le plus utilisé pour l'apprentissage des modèles GMM, est le critère de maximum de vraisemblance *ML* (*maximum likelihood*). L'estimation des paramètres du GMM consiste à trouver ceux qui maximisent la fonction de vraisemblance des données d'apprentissage.

$$\tilde{\lambda}_X = \underset{\lambda}{argmax}(p(X|\lambda)$$

Où X est l'ensemble de trames d'apprentissage : $X = \vec{x}_1, \vec{x}_2, \ldots, \vec{x}_T$ et $p(X|\lambda)$. La vraisemblance de X sachant le modèle GMM est :

$$p(X|\lambda) = \prod_t p(\vec{x}_t|\lambda) \quad \textbf{3.5}$$

Il est très complexe de résoudre l'équation 3.5 à cause du manque d'information des données d'apprentissage. En effet, il est difficile de savoir quelle gaussienne dans le mélange a généré une trame d'apprentissage donnée. Pour résoudre ce problème, appelé problème des données manquantes, l'algorithme Expectation Maximisation (EM) [35] est communément utilisé. L'étape Expectation détermine les probabilités *a posteriori* que les gaussiennes aient généré les trames d'apprentissage. Ensuite, l'étape Maximisation modifie les paramètres du modèle pour maximiser le critère choisi. Cet algorithme est itératif, il garantit l'augmentation de la vraisemblance des données sachant le modèle $p(X \mid \lambda)$.

En pratique, il existe plusieurs critères pour l'apprentissage des modèles GMM de locuteurs : le critère de maximum de vraisemblance ML (*maximum likelihood*) présenté, le critère MMI (*maximum mutual information*) et le critère MAP (*Maximum a posteriori*).

Le choix du critère d'apprentissage dépend de la quantité de données d'apprentissage disponible. Lorsque cette quantité est limitée, l'estimation au sens du maximum de vraisemblance pose le problème du sur-apprentissage. Le modèle résultant est alors trop spécifique aux données et perd sa capacité de généralisation. L'estimation au sens du critère MAP permet d'introduire les densités de probabilité des paramètres *a priori* du modèle GMM λ :

$$\tilde{\lambda}_X = argmax_{\lambda} p(X|\lambda).p(\lambda)$$

Le critère MMI (Maximum Mutual Information) [91] vise à intégrer un critère discriminant lors de l'apprentissage. Les paramètres du GMM sont alors modifiés selon une fonction objective, dont le but est de diminuer l'influence des densités de probabilité qui modélisent des informations communes, entre le modèle d'apprentissage et des modèles de contre-exemple.

La phase d'initialisation est très importante lors de l'apprentissage d'un modèle GMM. Les techniques les plus courantes sélectionnent aléatoirement des données dans l'ensemble d'apprentissage pour initialiser les moyennes, la matrice de variance est la matrice unité et les poids suivent la loi d'équiprobabilité. Les moyennes initialisées du GMM peuvent être réactualisées par l'utilisation de l'algorithme de classification de type *k-means* [49].

3.2.2.2. Modèle de Markov Caché

Empruntés à la RAP, les modèles de Markov cachés (Hidden Markov Model : HMM) [92, 93, 94] permettent de caractériser les variations temporelles du signal de parole. Ils reposent sur une machine à états, i.e. une succession d'états associés à des probabilités de transition d'un état à l'autre. Une ou plusieurs distributions de probabilité associées à chaque état caractérisent les probabilités d'émissions des vecteurs acoustiques par un état.

Lors de la reconnaissance, la vraisemblance pour une séquence de vecteurs de test est calculée.

3.2.2.3. L'approche connexionniste

L'approche connexionniste, telle que nous l'entendons ici, repose sur la discrimination entre locuteurs. Elle consiste à fournir à un réseau de neurones un ensemble de signaux de parole issus d'une population de locuteurs clients afin que ce réseau apprenne comment discriminer un locuteur des autres. L'approche connexionniste se résume, par conséquent, à une tâche de classification. Un modèle client se présente sous forme d'un ou plusieurs réseaux de neurones pour lesquels la séquence de vecteurs d'apprentissage du client concerné ainsi que celles des autres clients du système sont fournies en entrée. Plusieurs types de réseaux de neurones sont proposés dans la littérature : le réseau de neurones multicouches MLP

[95,96,97,98] et apprentissage de la quantification vectorielle LVQ [93]. Nous nous intéressons ici au réseau de neurones multicouches. Lors de la reconnaissance, la vraisemblance pour une séquence de vecteurs de test soit produite par un réseau de neurones est calculée [99,100,101,102,103] . La figure 4.8 présente un exemple de réseau de neurones multicouches :

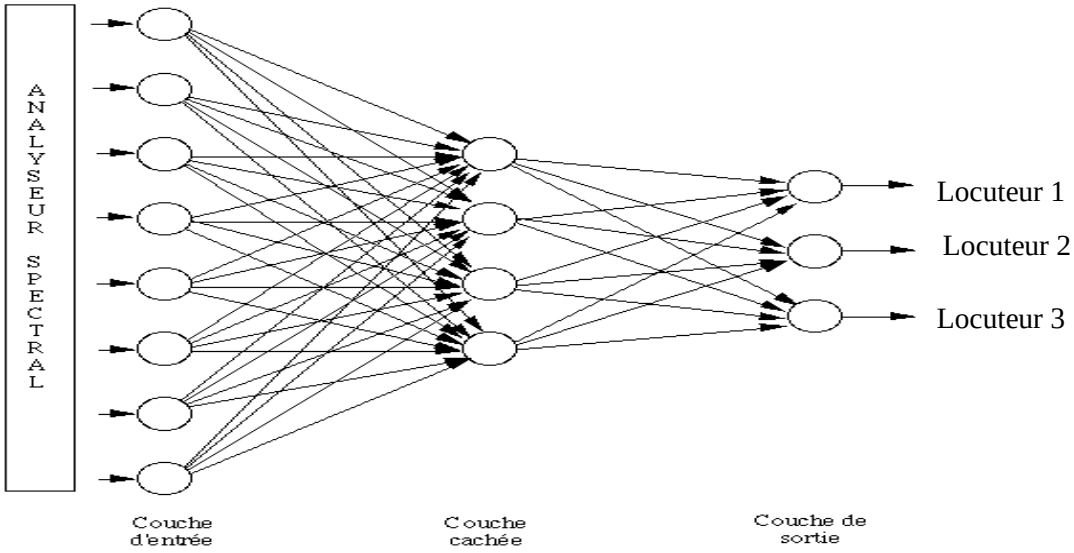


Figure 3.8 : exemple de réseau de neurones multicouches

Pour une entrée *x*, la sortie est calculée en utilisant la fonction d'activation f telle que :

$$f(x) = \sum_{i=0}^{n} (w_i . a_i) - t$$

Avec :

a : le vecteur d'entrée

w : les poids associés à chaque neurone

t : le seuil de décision associé au réseau

Cette fonction permet de donner une interprétation binaire sur la couche de sortie en se basant sur un seuillage des valeurs des neurones de cette couche.

Généralement pour la fonction d'activation, on utilise la fonction sigmoïde définie par :

$$sigmoid(x, c) = \frac{1}{(1 + e^{-cx})}$$

$$\frac{d(sigmoid(x;c))}{dx} = c.sigmoid(x;c).\left(1 - sigmoid(x;c)\right)$$

La structure du réseau utilisée est un réseau de neurones à *feed-forward* [104], Cela implique que les neurones sont organisés en ensembles. En effet, un neurone de la couche j, possède des entrées de couche j-1 et en sortie à la couche j+1 uniquement. Cette structure facilite l'évaluation et la formation d'un réseau. Par exemple, lors de l'évaluation d'un réseau sur un vecteur I d'entrée, la sortie du neurone de la première couche est calculée, suivie par la deuxième couche, et ainsi de suite.

L'entrainement de ce type de réseau de neurones se base sur la rétro-propagation de l'erreur. La correction de l'erreur à travers le réseau commence par la couche de sortie. La correction des poids se fait suivant la relation :

$$\omega_{i,j} = \beta\omega_{i,j} + \alpha.a_j.\Delta_i$$

Avec :

$\Delta_i = Err_i.\frac{df}{dx(in_i)}$ Pour le neurone i à la couche de sortie.

$\Delta_i = \frac{df}{dx(in_i)}.\sum_{j=0}^{n}\Delta_j$ Pour les autres couches.

Avec :

Err_i : l'erreur sur le neurone i.

F : fonction d'activation.

in_i : l'entrée au neurone i.

Les paramètres α et β sont utilisés pour éviter des minimums locaux dans le processus d'optimisation de l'entrainement. Ils permettent de pondérer la combinaison de l'ancien poids avec l'ajout de la nouvelle modification. Les valeurs habituelles pour ces paramètres sont déterminées expérimentalement.

Comme un classificateur, le réseau multicouche est utilisé pour pointer les vecteurs acoustiques vers les identifiants du locuteur. Les neurones d'entrée correspondent à chaque caractéristique du vecteur acoustique. La sortie du réseau est une interprétation binaire de la couche de sortie. Ainsi, la couche d'entrée est de taille M, ou M est la taille des vecteurs acoustiques et la couche de sortie comporte $log_2 n$ ou n est le nombre maximum des locuteurs à identifier.

Chapitre 4

Traitement de la parole et sécurité

1. INTRODUCTION

Dans le domaine de la sécurité militaire ou civile, les techniques d'authentification biométriques habituellement employées sont : la reconnaissance d'empreintes digitales ou de l'iris. Ces dernières autorisent un niveau de sécurité maximal et des taux d'erreur particulièrement réduits. Néanmoins, ces méthodes qui sont alors dites «intrusives» et restent lourdes à mettre en œuvre. L'intrusivité d'une méthode biométrique est définie comme le niveau d'acceptation de la méthode par l'utilisateur [62,105]. Les empreintes digitales sont ainsi peu acceptées du public pour des raisons d'éthique car elles ont une connotation criminalistique. Ainsi, la biométrie idéale doit présenter une fiabilité élevée, une grande facilité d'utilisation et un faible coût. Dans ce sens, l'utilisation de la voix pour l'authentification offre l'avantage d'être bien acceptée par l'utilisateur et d'être simple à mettre en œuvre. Basée sur un échantillon de voix du locuteur, elle n'implique que la prise de son à travers un microphone. De plus, c'est souvent le seul média disponible. Les systèmes de reconnaissance automatique du locuteur (RAL) s'appuient sur les caractéristiques de la parole permettant de reconnaître les individus. La reconnaissance du locuteur en tant que technique d'authentification présente les avantages suivants :

- ✓ L'acquisition du signal audio est très simple à mettre en œuvre,
- ✓ L'enregistrement du signal audio n'est pas considéré comme intrusif (mais peut cependant présenter des difficultés au niveau législatif), le signal audio est naturellement véhiculé dans la majorité des réseaux de communication,
- ✓ Les techniques de stockage et de compression du signal audio sont très efficaces,
- ✓ Dans de nombreuses applications (serveurs vocaux), l'utilisateur emploie déjà la parole pour communiquer avec la machine. Le coût supplémentaire de la RAL, en coût de mise en œuvre comme en contraintes ergonomiques, est faible.

Dans ce chapitre, nous allons exploiter notre système réalisé de reconnaissance automatique des dialectes marocains pour élaborer un nouveau système de sécurité. Ce dernier utilise la voix comme paramètre d'entrée et il est divisé en deux parties :

- ✓ la première partie consiste en vérification de mot de passe, elle permet de valider un mot de passe prononcé par un locuteur en se basant sur les informations déjà enregistrées dans une base de données.
- ✓ La deuxième partie consiste en vérification du locuteur en se basant sur le mot de

passe prononcé.

Les deux couches du système de sécurité sont complétives, la vérification de mots de passe permet de valider la première étape de l'authentification. Cette étape se base sur la puissance du système de reconnaissance de la parole. La deuxième étape de vérification de locuteur permet d'accomplir la tâche d'authentification en vérifiant l'identité de la personne qui a prononcé le signal de la parole.

Dans la suite de ce chapitre nous donnerons une description des deux parties qui composent le système, puis nous présentons l'architecture du système de sécurité proposé et nous finirons par les résultats expérimentaux.

2. Traitement de la parole et sécurité

Les systèmes de sécurité actuels se basent généralement sur les empreintes digitales. La plupart des systèmes de sécurité intégrés dans divers appareils reposent sur les mots de passe comme outil de verrouillage. En effet, cette technique ne considère que les informations textuelles sans prendre en compte l'identification personnelle du propriétaire de l'appareil. L'apparition de verrouillage avec des empreintes digitales a résolu à une certaines limites le problème de sécurité. Malgré la fiabilité de cette technique, elle n'a pas pu être intégrée en somme dans tous les appareils. Dans ce sens, l'introduction de la reconnaissance vocale dans le domaine de sécurité commence à donner ces résultats depuis le début de $20^{ème}$siécle. Dans cette partie de ce mémoire, nous présentons un système de sécurité basé sur la reconnaissance vocale. Ce système est divisé en deux couches : la première se base sur un système de reconnaissance automatique de la parole pour vérifier le mot de passe saisi par un utilisateur de système. La deuxième couche consiste à identifier le locuteur qui a saisi le mot de passe et vérifier son identité. La figure 4.1 montre la structure générale du système :

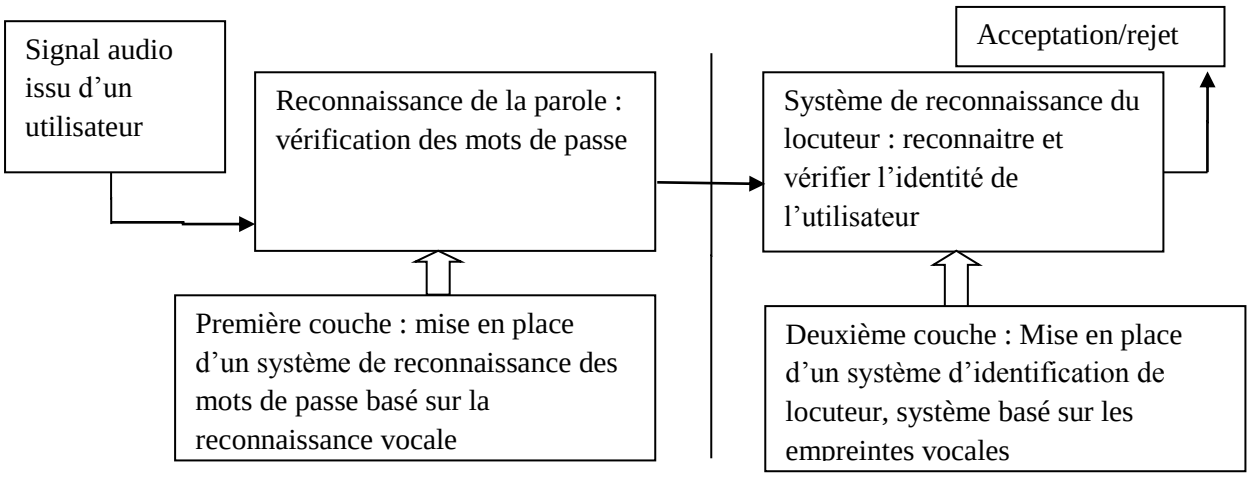


Figure 4.1 : architecture du système de sécurité basé sur la reconnaissance vocale et la vérification du locuteur

2.1. Vérification automatique du mot de passe

Dans cette partie, nous exploitons le système de reconnaissance automatique, présenté au chapitre 2, des dialectes marocains Tamazight et Darija pour élaborer un système de vérification du mot de passe. Ce système est basé sur les bases d'apprentissage utilisées dans le chapitre 2.

La vérification de mot de passe se base sur un système de reconnaissance automatique de la parole. Un client est invité à prononcer son mot de passe, un système de reconnaissance transmet en paramètre le mot reconnu à un programme de vérification. Ce dernier permet de comparer l'information issue du système de reconnaissance avec le mot de passe existant dans la base de données textuelle et décide sur l'acceptation ou le rejet du signal. La figure 4.2 présente une illustration de cette partie du système :

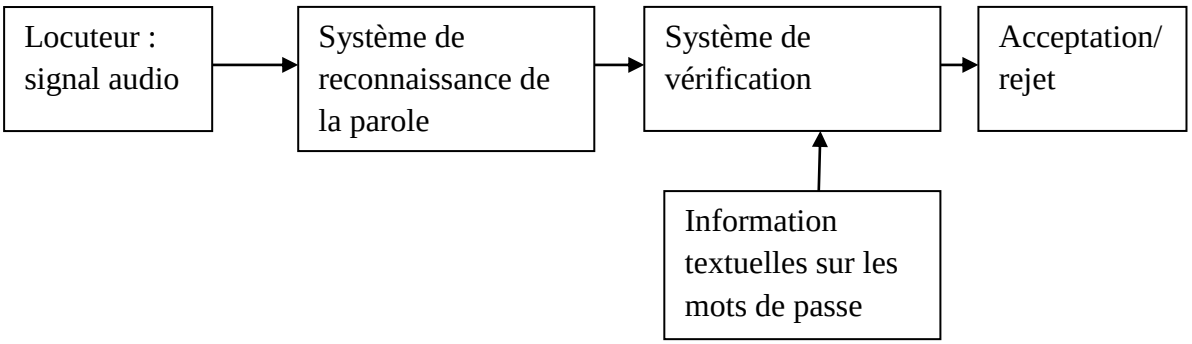


Figure 4.2 : Vérification automatique du mot de passe

Le système de vérification de mot de passe est basé sur les résultats de reconnaissance obtenus pour la reconnaissance des chiffres enchainés en Tamazight et les chiffres isolés en Darija. Il repose sur un modèle de grammaire contenant tous les combinaisons des chiffres possibles.

2.2. Reconnaissance et vérification du locuteur

La deuxième étape du système de sécurité consiste en vérification du locuteur. Cette tâche se base sur les empreintes vocales pour vérifier l'identité de la personne qui fournit le mot de passe. Cette étape, s'agit plutôt de la vérification car les données personnelles du locuteur doivent être lus à partir d'un support externe (exemple : carte guichet). Dans ce sens, l'envoi du mot de passe prononcé à la vérification doit être accompagné de l'identité de la personne qu'il l'a saisi. La prise des empreintes vocales nécessaires pour la tâche d'authentification est faite au moment de la prononciation du mot de passe, ces empreintes suivent le profil du locuteur comme des variables sessions.

La vérification du locuteur constitue une tâche importante dans les systèmes de sécurité. La vérification de mot de passe n'accomplit pas la tâche d'authentification surtout dans les systèmes de reconnaissance de la parole en cas du multi-locuteur. La figure 4.3 présente une illustration de ce système :

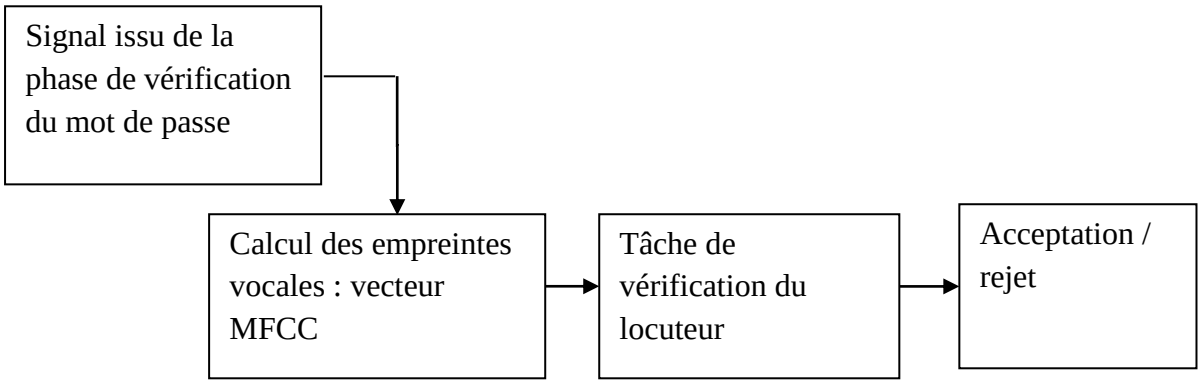


Figure 4.3 : tâche de vérification du locuteur

2.3. Choix des paramètres acoustiques

Nous avons précisé que les coefficients cepstraux peuvent être déterminés en utilisant une méthode non paramétrique, l'analyse cepstrale (MFCC ou LFCC), ou une méthode paramétrique, l'analyse LPC du signal (LPCC). Le calcul des coefficients cepstraux est détaillé dans la partie annexe.

L'analyse cepstrale est l'approche la plus utilisée en RAL, notamment parce qu'elle présente une plus grande robustesse d'estimation sur des signaux bruités [105]. En RAL et VAL, entre 13 et 20 coefficients cepstraux sont utilisés pour modéliser un locuteur. Ils sont généralement extraits toutes les 10 ms (hypothèse de pseudo-stationnarité) et calculés sur une fenêtre d'analyse de Hamming de 20 à 30ms.

Une analyse en banc de filtres à échelle linéaire ou échelle de Mel est utilisée dans le calcul des coefficients cepstraux (MFCC). Les dérivées premières, ou coefficients Δ (vitesse) et parfois secondes ou coefficients ΔΔ (accélération), des coefficients cepstraux sont ajoutés au vecteur de paramètres pour modéliser leur trajectoires dans le temps.

L'énergie du signal joue aussi un rôle important en RAL tant au niveau de la sélection des données utiles que comme paramètre. En effet, ce paramètre est souvent utilisé pour la détection d'activité vocale, et sa trajectoire (Δ-log-énergie) est souvent ajoutée au vecteur de paramètres.

Dans cette partie qui concerne la vérification du locuteur nous avons utilisé 12 coefficients MFCC en plus de l'énergie. Le nombre de coefficients est 13 en plus des dérivées premières et secondes ce qui donne en total 39 coefficients.

2.4. Modélisation des locuteurs par les réseaux de neurones

Les locuteurs dans la base d'apprentissage sont modélisés par les réseaux de neurones multicouches [106]. On a utilisé une couche d'entrée, une couche cachée et une couche de sortie. Les neurones d'entrée représentent les données des vecteurs de paramètres MFCC, la couche de sortie présente l'interprétation binaire de la couche d'entrée. La figure 4.4 présente la structure du réseau utilisé :

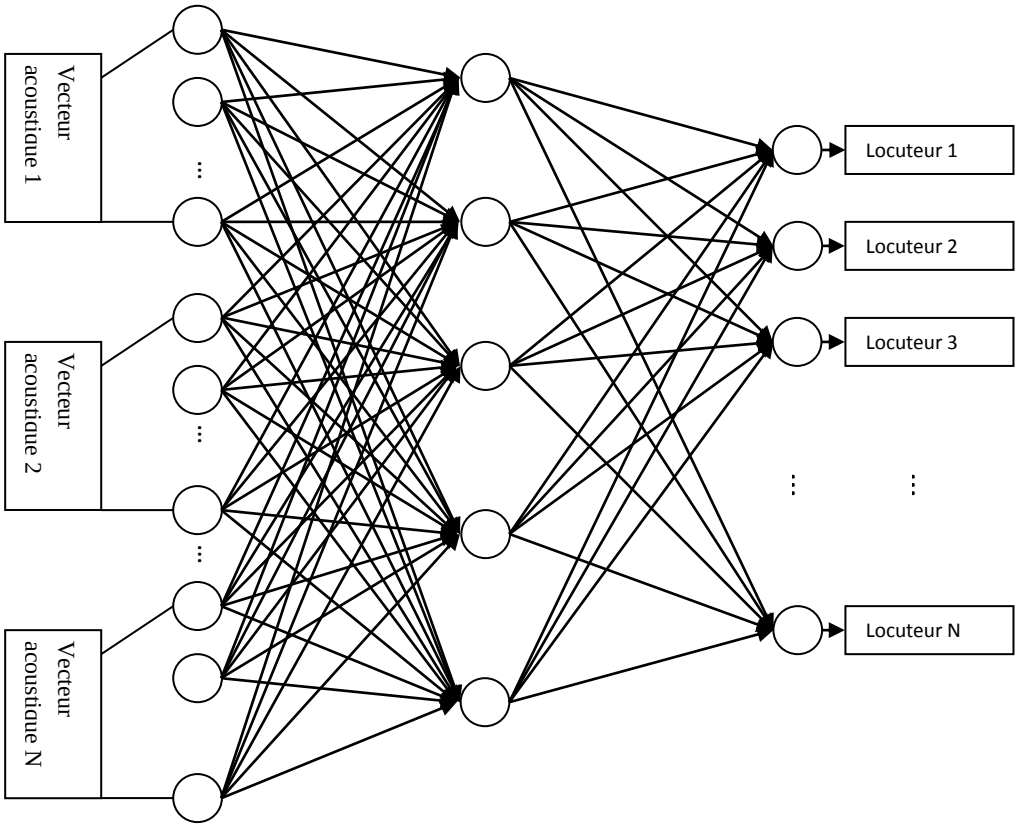


Figure 4.4 : Réseau de neurones Multicouches pour la reconnaissance du locuteur

L'apprentissage du réseau se fait par l'algorithme de rétro-propagation du gradient [106] donné dans le chapitre 3. La table 4.1 présente une interprétation binaire des sept premiers locuteurs sur la couche de sortie :

Locuteurs	**Identifiant**	**Interprétation binaire**
Locuteur 1	1	1000000000000000000
Locuteur 2	2	0100000000000000000
Locuteur 3	3	0010000000000000000
Locuteur 4	4	0001000000000000000
Locuteur 5	5	0000100000000000000
Locuteur 6	6	0000010000000000000
Locuteur 7	7	0000001000000000000

Table 4.1 : interprétation binaire des sept premiers locuteurs

2.5. Architecture du système proposé

Le système proposé se compose de deux processus consécutifs. Le premier processus est basé sur un système de reconnaissance automatique de la parole, il permet la vérification du mot de passe en se basant sur un signal de la parole issu de l'utilisateur. Le deuxième processus consiste en vérification de l'identité de cette personne. La figure 4.5 présente un organigramme global du système de sécurité.

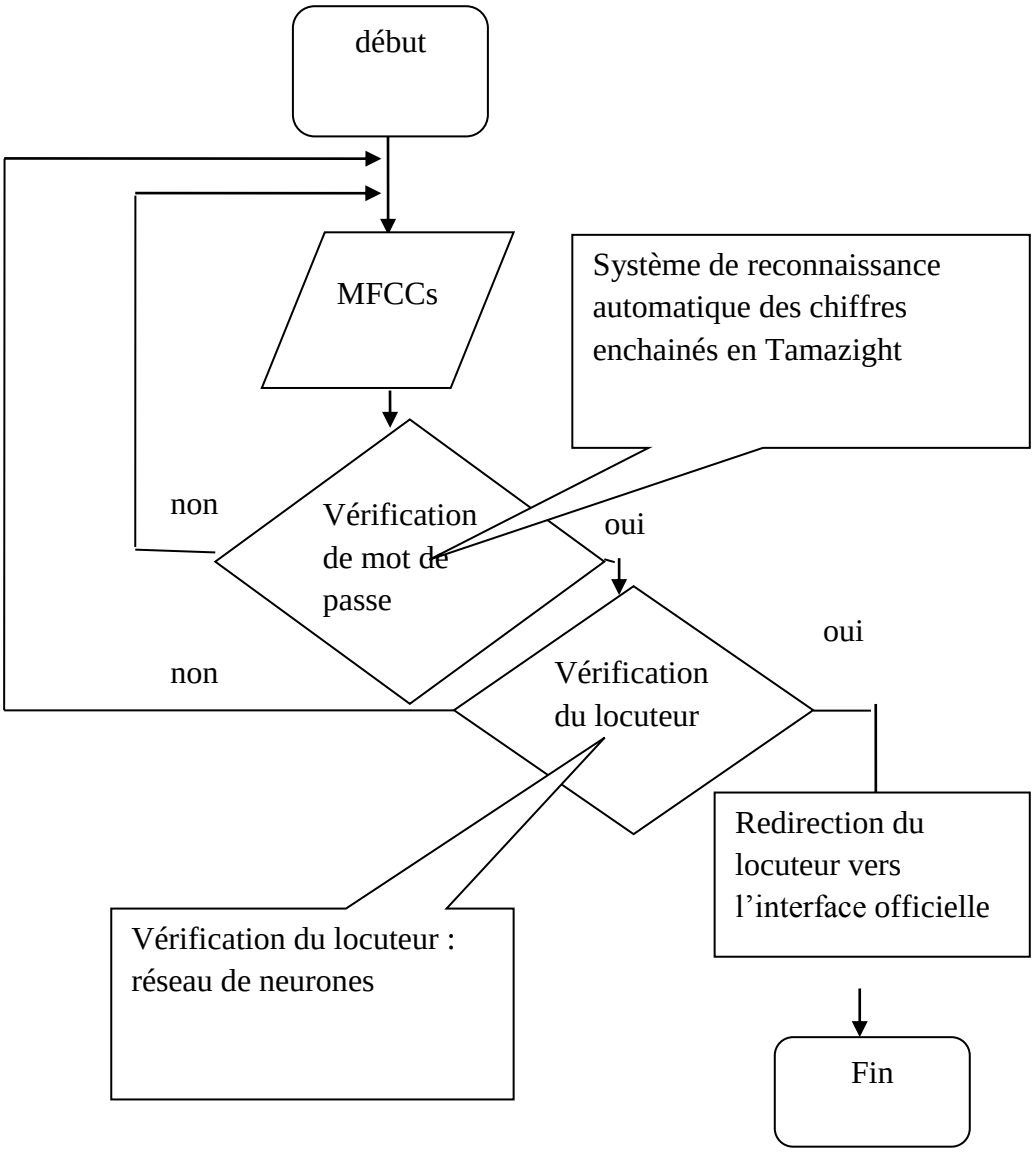


Figure 4.5 : organigramme du système de sécurité

La figure 4.6 présente une architecture détaillée du système proposé :

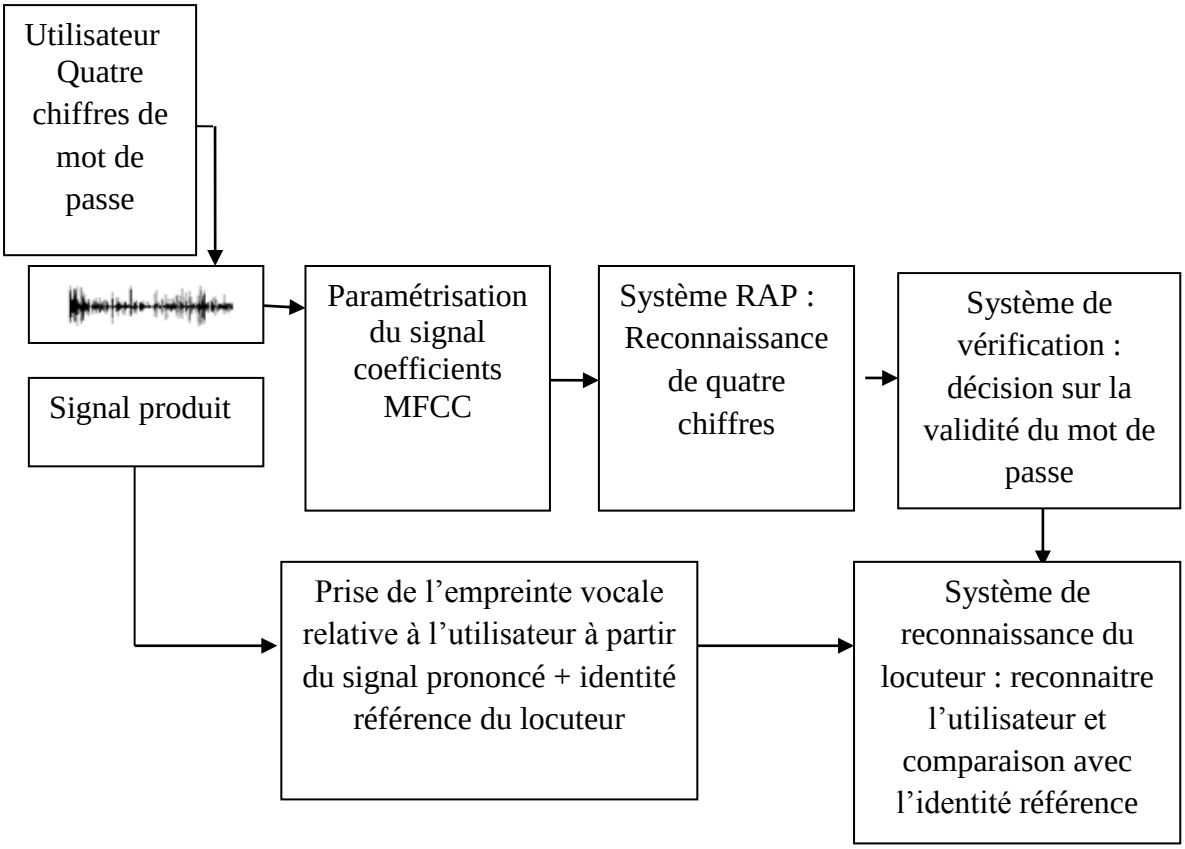


Figure 4.6 : Architecture détaillée du système

3. Résultats expérimentaux

3.1. Reconnaissance du locuteur

3.1.1. Base d'apprentissage pour la reconnaissance du locuteur

La base d'apprentissages comporte 20 locuteurs, chacun de ces derniers est invité à prononcer des phrases différentes en Tamazight et Darija. En effet, le système de reconnaissance réalisé est un système indépendant du texte. Un locuteur client est invité à prononcer un chiffre en Tamazight de 1 à 199, le système de reconnaissance permet d'identifier cette personne parmi 20 personnes utilisées dans la phase d'apprentissage. La table 4.2 présente les caractéristiques de la base d'apprentissage :

Locuteurs	Type	Taille de l'enregistrement
Loc_1	Masculin	8,5min
Loc_2	Masculin	6,4min
Loc_3	Masculin	7,6min
Loc_4	Féminin	3,5min
Loc_5	Masculin	8,2min
Loc_6	Masculin	5,4min
Loc_7	Masculin	7,1min
Loc_8	Masculin	6,5min
Loc_9	Féminin	8,4min
Loc_10	Féminin	6,4min
Loc_11	Féminin	6,7min
Loc_12	Masculin	8,3min
Loc_13	Masculin	9,4min
Loc_14	Masculin	5,6min
Loc_15	Masculin	5,8min
Loc_16	Féminin	6,7min
Loc_17	Féminin	4,8min
Loc_18	Féminin	9,3min
Loc_19	Féminin	5,8min
Loc_20	Féminin	6,8min

Table 4.2 : Distribution des locuteurs dans la base d'apprentissage

Contrairement à la reconnaissance automatique de la parole, un système de reconnaissance automatique du locuteur peut donner des bons résultats avec une base de données d'apprentissage minimale.

3.1.2. Résultats

Le système de reconnaissance automatique du locuteur est vérifié sur une base de données de test. Il est évalué en calculant le taux de reconnaissance selon la relation suivante :

$$T_r = \frac{\text{nombre de locuteurs reconnus}}{\text{taille de la base de test}}$$

On calcule aussi le taux de faux rejet et le taux de fausse acceptation selon les relations suivantes :

$$T_{fr} = \frac{\text{nombre de locuteurs rejetés}}{\text{taille de la base de test}}$$

$$T_{fa} = \frac{\text{nombre de locuteurs en fausse acceptation}}{\text{taille de la base de test}}$$

Les résultats obtenus sont illustrés dans la table 4.3 :

	Taux d'évaluation
T_r	86%
T_{fr}	8,5%
T_{fa}	5,5%

Table 4.3 : Résultats expérimentaux

3.2. Vérification du mot de passe

3.2.1. Base d'apprentissage pour la vérification du mot de passe

Pour l'apprentissage du système, nous avons utilisé les bases d'apprentissage données dans le chapitre 2 : base de données pour les chiffres enchainés en Tamazight et la base de données pour les chiffres isolés en Darija. La table 4.4 présente les caractéristiques de ces deux bases d'apprentissage :

	Caractéristiques temporelles de la base d'apprentissage
Darija	1h 20 min
Tamazight	6,28 heures

Table 4.4 : Base d'apprentissage pour la reconnaissance du mot de passe

3.2.2. Résultats

La base de données de test est composée de la prononciation des mots de passe. Chaque mot de passe est composé de quatre chiffres (figure 4.5).

Base de test	**Taux de Reconnaissance**	**Taux d'erreur**
200 mots de passe prononcés en Darija et Tamazight	T=92.2%	7,8%

Table 4.5 : Résultats de vérification du mot de passe

3.3. Evaluation des performances du système

Le système de sécurité est évalué en calculant un taux d'identification complète du locuteur en question. Ce rapport de vérification est définit par la relation suivante :

$$T_{ident} = \frac{\text{nombre de locuteurs identifiés}}{\text{Taille de la base de test}}$$

On calcule aussi les taux d'erreurs suivants :

- ✓ T_{pf} : le taux de rejet à partir de la vérification du mot de passe.
- ✓ T_{lf} : le taux de rejet à partir de l'étape de vérification du locuteur.

$$T_{pf} = \frac{\text{nombre de locuteurs rejetés dans l'étape de vérifcation du mot de passe}}{\text{taille de la base de test}}$$

$$T_{lf} = \frac{\text{nombre de locuteurs rejetés dans l'étape de véririfcation du locuteur}}{\text{taille de la base de test}}$$

La figure 4.7 montre les taux d'évaluation du système :

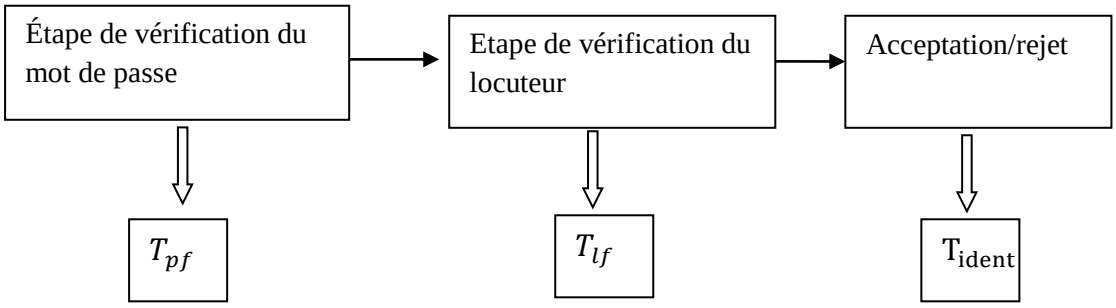


Figure 4.7 : taux d'évaluation du système

3.3.1. Base de test

La base des signaux de test est constituée des fichiers audio récupérés, généralement, auprès des personnes qui ont participé à la construction des deux bases d'apprentissage (la base de vérification de mot de passe et la base de vérification du locuteur). Nous constatons que les personnes qui n'ont pas participé à la base de reconnaissance du locuteur (respectivement la base de vérification du mot de passe) sont rejetées dans la deuxième phase (respectivement la première phase). Le diagramme 4.8 présente la structure de la base d'apprentissage :

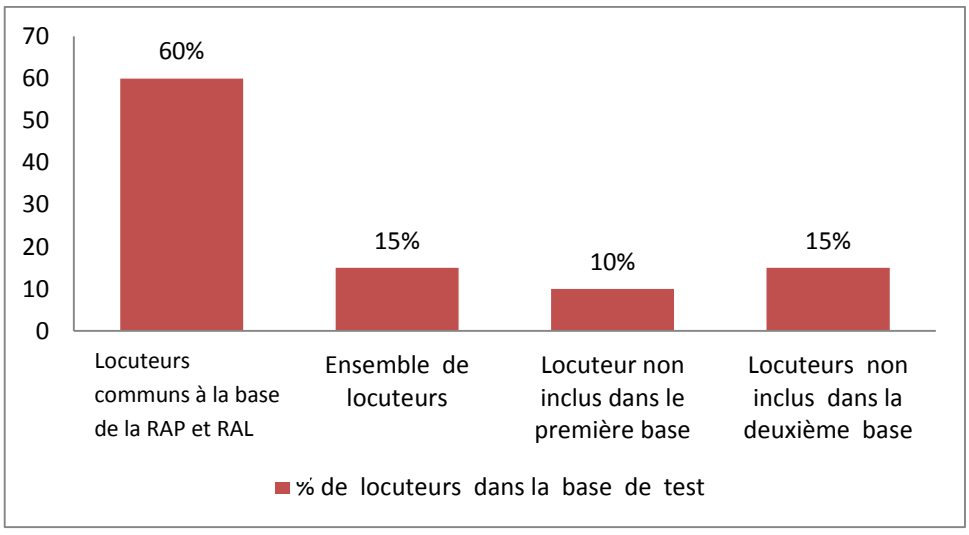


Figure 4.8 : pourcentages de locuteurs dans la base de test

3.3.2. Résultats

Les résultats obtenus en calculant les taux d'évaluation sont donnés dans la table 4.6 :

	Taux de reconnaissance
T_{ident}	89%
T_{pf}	7,8%
T_{lf}	3,2

Table 4.6 : Résultats obtenus

4. Conclusion

Le domaine de sécurité est l'un des domaines les plus sensibles aux données personnelles. Dans les systèmes de sécurité, les informations porteuses d'identités personnelles sont toujours cachées. Dans ce chapitre, nous avons exploité la reconnaissance vocale et l'identification du locuteur pour réaliser un système de sécurité. Ce système a tendance à être intégré dans les appareils mobiles, les guichets bancaires, l'identification via des appels téléphoniques, etc.

Les résultats obtenus en fin de ce chapitre confirment les conclusions tirées des travaux présentés tout au long de ce document. Ces résultats ont, une nouvelle fois, démontrés :

- ✓ La pertinence de l'information liée au locuteur dans le cadre de la reconnaissance automatique du locuteur ;
- ✓ Dans un système de sécurité, la vérification du locuteur permet d'identifier une personne en se basant sur les informations en ligne saisies par la personne en question. Ce processus, repose sur les dialectes Tamazight et Darija, ce qui permet de faciliter l'accès au système par tout le monde.
- ✓ La combinaison d'un système de reconnaissance de mots de passe avec un système de vérification du locuteur permet de rendre la tâche d'authentification, toute entière, basée sur un signal produit par la personne consignée.
- ✓ Le volume d'information qu'importe le conduit vocal peut être comparé à l'empreinte digitale. Néanmoins, les empreintes vocales se voient comme un signal variable selon l'état physique de la personne.

Les résultats obtenus sont satisfaisants, vis-à-vis de la taille des bases d'apprentissage. Par ailleurs, le signal de la parole n'est pas stable, il peut varier selon les conditions externes ou internes. En effet, les systèmes de sécurité basés sur la voix doivent être entrainés avec des signaux pris dans différentes situations des locuteurs.

Chapitre 5

Conclusion et perspectives

L'application du traitement automatique de la parole dans le domaine de sécurité permet de faciliter cette tâche et rendre plus efficace l'identification personnelle d'un individu. Néanmoins, le signal vocal présente une grande variabilité interlocuteur et intra-locuteurs ce qui rend la tâche de sécurité assez difficile, celle-ci nécessite l'identification du contenu du signal de la parole ainsi que la personne qui l'a prononcé. C'est sur cette base qu'ont été développées toutes les méthodologies de ce travail.

Dans le chapitre 2 de cette thèse, nous avons décrit la conception de quelques nouveaux systèmes de reconnaissance automatique de la parole relatifs aux dialectes marocains Tamazight et Darija. Nous avons utilisé des transcriptions phonétiques issues de la segmentation en phonèmes de ces deux dialectes, chaque unité de base est modélisée par un modèle de Markov caché (MMC) à trois états. En premier lieu, nous avons présenté deux systèmes de reconnaissance relatifs aux mots isolés en Darija et Tamazight et ensuite un troisième système hybride rassemblant les deux dialectes. Ces systèmes sont évalués sur une base de données de test créée au sein de notre laboratoire. Le calcul des taux de reconnaissance a permis d'évaluer l'efficacité des paramètres qu'on a utilisés pour caractériser les signaux de la parole des deux dialectes. Le deuxième volet de ce chapitre est consacré à la reconnaissance des chiffres enchainés en Tamazight. Dans ce sens, nous avons proposé une approche permettant de reformuler les chiffres enchainés à partir des nombres isolés. Au vu des résultats obtenus pour le cas des chiffres enchainés, nous constatons qu'il y a une amélioration du taux de reconnaissance par rapport au cas des chiffres isolés au niveau de l'ensemble de test. De même nous constatons qu'il une réduction significative du taux d'erreur.

Dans la quatrième partie de cette thèse, nous nous sommes amenés à créer un nouveau système de sécurité rassemblant la reconnaissance vocale et l'identification automatique du locuteur. Ce système de sécurité se compose de deux phases consécutives :

- ✓ La première phase permettant la validation d'un mot de passe. Elle est basée sur les systèmes de reconnaissance automatique des dialectes marocains créés dans le troisième chapitre.
- ✓ La deuxième phase permet la vérification du locuteur. Celle-ci se base sur un modèle référence de chaque utilisateur du système pour vérifier l'identité de la personne qui a prononcé le mot de passe.

Sur le plan pratique, notre approche permet l'amélioration de la sécurité dans les

systèmes utilisant le traitement automatique de la voix comme méthode d'identification personnelle. Elle admet aussi de mettre en place un système complet qui se base sur un signal vocal produit auprès d'une personne utilisateur d'une application donnée.

Les résultats obtenus démontrent une autre fois l'importance de la modélisation statistique et neuronale dans le traitement automatique des signaux. Malgré la variation et la non-stationnarité du signal de la parole, les résultats sont satisfaisants. Le système de sécurité élaboré en combinant la reconnaissance vocale et l'identification automatique du locuteur présente l'intérêt d'être utilisé dans plusieurs appareils mobiles et dans les guichets automatiques. Ce système vise à améliorer la sécurité basée sur les empreintes vocales, pourtant l'instabilité interlocuteur des signaux vocaux rend insuffisante la tâche d'identification du locuteur dans le domaine de sécurité. Dans ce sens, de nouvelles techniques peuvent être mises en place, à savoir la combinaison de la reconnaissance vocale et la reconnaissance des visages ou encore l'introduction de nouveaux paramètres qui prennent en compte l'information linguistique du signal permettant de rendre la tâche de sécurité assez fiable.

Perspectives

Bien que l'approche statistique se soit avérée pertinent dans la modélisation des signaux vocaux, elle a montré ses limites dans le domaine de reconnaissance. Dans ce sens, nous pensons à introduire de nouveaux paramètres qui prennent en compte l'information linguistique du signal, d'un autre côté ces paramètres doivent définir le plus possible les caractéristiques liées au locuteur à savoir les particularités articulatoires et la forme du conduit vocal.

Par ailleurs, une étude approfondie sur l'utilisation de la reconnaissance vocale et l'identification automatique du locuteur, peut s'avérer dangereuse dans le domaine bancaire. Ce système doit être testé sur des systèmes de dialogue homme-machine dont les transactions ne présentent aucun risque sur les bases de données sensibles.

Les perspectives des travaux réalisés peuvent poster sur différents niveaux, ils concernent tout d'abord la création d'un système de reconnaissance automatique de la parole multi-locuteur avec une base de données d'apprentissage standardisée, d'un autre côté le dialecte Tamazight doit être pris avec une segmentation phonétique normalisée. Pour cela, une étude linguistique approfondie de Tamazight doit être faite, ensuite un système de reconnaissance automatique de la parole continue doit être mis en place avec un modèle de langage plus approprié. Finalement, l'utilisation du système hybride MMC/RNA, comme une approche d'apprentissage du modèle acoustique peut améliorer davantage le taux de reconnaissance ainsi que la fiabilité du système réalisé.

Dans un système de sécurité, la vérification du locuteur apparait comme une tâche insuffisante à l'authentification, la voix d'une personne peut être imitée par un autre individu. Ainsi, nous percevons ajouter une nouvelle couche de vérification, celle-ci concerne la reconnaissance des visages à partir des images prises en ligne.

Bibliographie

[1] A. Sadiqui & N. chenfour, "Reconnaissance de la parole arabe basé sur CMU Sphinx", Séria Informatica. Vol VIII fasc.1, 2010.

[2] D. Genoud, "Reconnaissance et transformation du locuteur", thèse présenté à l'école polytechnique de Lausanne, 2007.

[3] A. PRITI, "Surveillance des réseaux professionnels de communication par la reconnaissance du locuteur", Thèse à l'école Doctorale 166 I2S Mathématiques et Informatique, Laboratoire d'Informatique d'Avignon 2008.

[4] T. Pellegrini et Raphael Durée, "Suivi de la voix parlée garce au modèle caché", Rapport de stage DEA ATIAM, France 2003.

[5] Y. Grenier, "Identification du locuteur et adaptation au locuteur d'un système de reconnaissance phonétique". Thèse de doctorat, Ecole Nationale Supérieur des Télécommunications (ENST), Paris(France), 1977.

[6] S. Jamoussi, "Méthodes statistiques pour la compréhension automatique de parole", Ecole doctorale IAEM Lorraine, 2004.

[7] D. MERHEJ, " Intégration de connaissances a priori dans la reconstruction des signaux parcimonieux : Cas particulier de la spectroscopie RMN multidimensionnelle", École doctorale Sciences et Technologies, L'UNIVERSITE LIBANAISE – LIBAN 2012.

[8] G. SEMET & G. TREFFOT, "La reconnaissance de la parole avec les MFCC", TIPE juin 2002.

[9] J. Baker, "The DRAGON system-An overview", IEEE Trans. Acoust., Speech signal processing, vol. 23, no.1; pp. 24-29, 1975.

[10] J.P Haton, JM Pierrel, G. Pérennou, J. Caelen et J.L. Gauvain, "reconnaissance automatique de la parole", AFCET, édition DUNOD Informatique, 1991.

[11] A. BALA, "voice command recognition system based on MFCC and DTW" International Journal of Engineering Science and Technology Vol. 2 (12), 2010.

[12] Y. Pan, Peipei Shen and Liping Shen, "Speech Emotion Recognition Using Support Vector Machine", International Journal of Smart Home Vol. 6, No. 2, April 2012.

[13] S. Gaetan & TREFFO, "Grégory,' Reconnaissance de la parole avec les coefficients MFCC", TIPE jiun 2002.

[14] JM Pierrel, G. Prénnou, J. Caelen et J.L. Gauvain : "Reconnaissance automatique de la parole", AFCET, édition Dunod Informatique, 1991.

[15] R. Bellman: "Dynamic programming", Princeton University Press, 1957.

[16] D. Jouvet, M. Dautremont et Q. Gossart , “Comparaison des multi modèles et des densités multi gaussiennes pour la reconnaissance de la parole par des modèles de Markov”, Actes des 20èmes JEP, pp. 159-164, 1994.

[17] T.K. Vintsjuk, "Recognition of words of oral speech by dynamic programming”, Kibernetika, vol.81, no. 8, 1968.

[18] C. Fang, "From Dynamic Time Warping (DTW) to Hidden Markov Model" (HMM) University of Cincinnati 2009.

[19] R. André-Obrecht, "A new statistical approach for the automatic segmentation of continuous speech signal", IEEE Trans. On acoustics, speech, signal processing, vol.36, nu 1, Janvier 1988.

[20] A. Shamma, ”Neural Networks for speech processing and recognition”, in first International Conference on Neural Networks (M. Caudill and C. Butler, eds.), (San Diego), IEEE, 1987.

[21] R. L. Watrous, "Speech recognition using Connectionist Networks", PhD thesis, University o Pennsylvania, Philadelphia, 1988.

[22] H. Bourlard and C. Wellekens, "Multilayer perceptron and automatic speech recognition", in Proceedings of IEEE First International Conference on Neural Networks, San Diego (M. Caudill and C. Butler, eds.), pp. 407-416, IEEE, 1987.

[23] H. Bourlard and N. Morgan, "A continuous speech recognition system embedding MLP into HMM", in Advances in Neural Information Processing System 2(D. Touretzky, ed.), pp. 186-193, Morgan, Kaufmann, 1990.

[24] H. Bourlard, N. Morgan, C. Wellekens, "Statistical inference in multilayer perceptrons and hidden Markov models with applications in continuous speech recognition", Neuro computing Algorithms, Architectures and Application pp. 217-226. F. Fogelman Soulie and J. Herault (eds), Nato ASI Series, 2001.

[25] M. Petit, Henning Christiansen, "Un calcul de Viterbi pour un Modèle de Markov Caché Contraintes", Department of Communication, Business and Information Technologies’, Roskilde University, P.O Box 260, DK-4000 Roskilde, Denmark, 2009.

[26] J. Tebelskis, A. Waibel, B. Petek, and O. Schmidbauer, ”Continuous speech recognition using liked predictive neural networks”, in Proceedings of the International Conference on Acoustic, Speech and Signal Processing, (Toronto, Canada), pp. 61-64, May 1991.

[27] F. Freitag, E. Monte, and J. M. Salavedra :”Predictive neural networks applied to phoneme recognition”, in Proc. Eurospeech’97, (Rhodes, Greece), pp. 2831-2834, Sept. 1997.

[28] D. Olivier, "Modèle dépendant du contexte et méthodes de fusion de données appliquées à la reconnaissance de la parole par modèle hybride HMM/MLP", Thèse du 3 cycle, 22 décembre 1998.

[29] F. Jelinek, “Continuous speech recognition by statistical methods”, proc. of IEEE, vol. 64, no. 4, pp. 532-556, 1976.

[30] K.-H. Lee, “Large vocabulary speaker-independent continuous speech recognition”, the SPHINX system, ph. D. Thesis, Carnegie Mellon University, 1988.

[31] Y. L. Chow, M.O. Dunham, O. Kimmball, M.A Krasner, G. Kubala, J. Makhoul, P.J. Price, S.Roucos & R. Schwartz, “BYBLOS, the BBN continuous speech recognition system”, Proc. ICASSP, pp. 89-92, Dallas, 1987.

[32] J.G. Wilpon, C.H. Lee & L.R. Rabiner, “connected digit recognition based on improved acoustic resolution”, Computer Speech and Language, vol.7, pp.15-26, 1993.

[33] B. LECOUTEUX, ‘’ Reconnaissance automatique de la parole guidée par des transcriptions a priori’’. Académie d'Aix Marseille - Université d'Avignon et des Pays Vaucluse - Laboratoire d'Informatique d'Avignon, Thèse Doctorat, 2002.

[34] L. R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition". IEEE transactions Speech Audio Processing, volume 77(2), pages 257-285, 1989.

[35] T. Matsui, S. Furui, "Speaker adaptation of tied-mixture –based phoneme model for text prompted speaker recognition". International Conference on Acoustics, speech, and signal Processing (Icassp), Istanbul (Turquie), 2000.

[36] D. Jouvet, “Reconnaissance de mots connectés indépendamment du locuteur par méthodes statistiques”, Thèse du 3 cycle, juin 1988.

[37] A. Cornijeol and L. Miclet , " Apprentissage Artificielle : méthode et concept" 1988.

[38] C. H. Lee, C.H.Lin & B.H.Juang, “Astudy on speaker adaptation of the parameters of continuous density hidden Markov models”, IEEE Trans, On Signal Processing, vol.39, no.4, pp. 806-814, 1991.

[39] S.J. Young et P. C. Woodland, ”The use of state tying in continuous speech recognition”, Proc, ESCA Eurospeech’93,3, pp.2203-2206, Berlin, Germany, september 1993.

[40] J.L. Gauvain & C.H.Lee, “Speaker adaptation based on MAP estimation of HMM parameters”, proc. IEEE ICASSP, pp II 558-561, Minneapolis, 1993.

[41] L. R. Rabiner, S. E. Leveinson, ”A speaker independant syntax directed connected word recognition system based on hidden Markov models and level building”, IEEE

Trans. ASSP, vol. 33, no.3, pp.561-573.

[42] L.A Liporace, “Maximum Likelihood estimation for multi-variant observation of Markov sources”, proc. IEEE trans IT, Vol.28, nu 5, pp. 729-734, 1982.

[43] L. Baum, “An inequality and association maximization technique in statistical estimation for probabilistic function of Markov processes”, Inequality, vol.3, 1972.

[44] J. Rice, Mathematical Statistics and data analysis”, page 511-540, 2006.

[45] R. Stern & M. Lasry, Dynamic speaker adaptation for feature based isolated word recognition”, proc. IEEE trans. Acoust. Speech Signal Process., vol. ASSP-35, no 6, June 1987.

[46] J. L. Gauvain & C. H. Lee, “Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains”, IEEE Trans, On Speech and Audio Processing, vol. 2, no. 2, pp. 291-298, 1994.

[47] L. R. Ahl, P. F. Brown, P.V. de Souza & R.L. Mercer,”Maximum mutual information estimation in hidden Markov model parameters for speech recognition”, Proc. ICASSP, pp. 49-52, Tokyo, 1986.

[48] P.F. Brown, ”The acoustic modeling problem in automatic speech recognition”, PhD Dissertation, Carnegie Mellon University, May 1987.

[49] H. Bourlard, C. J. Wellekens, H. Ney, ”Connected digit recognition using vector quantization”, Proc. IEEE Int. Conf. ASSP 1984, San Diego. CA. (March 1984).

[50] J. S. Bridle & M. D. Brown, R. M. Chamberlin, ”An algorithm for connected word recognition”, Proc. IEEE Int, Conf. ASSP pp.899-902, Paris, 1982.

[51] L. R. Rabiner, J. G. Wilpon & F. K. Soong, ”High performance connected digit recognition using hidden Markov models”, IEEE Trans. Acoust., Speech, Signal Processing, vol 37, no.8, pp.1214-1225,1989.

[52] T. Pellegrini et R. Duée, “Suivi de la voix parlée grâce aux modèles de Markov Caché“, lieu : IRCAM , Stravinsky 75004 PARIS juin 2003.

[53] M. Hwang and X. Huang, “Sub phonetic modeling with Markov states-senone”, Proc, IEEE ICASSP-92, San Francisco, CA, 1, pp.33-36, mars 1992.

[54] S. J. Young, ”The general use of tying in phoneme based HMM speech recognizers”, Proc, IEEE ICASSP-92, San Francisco, CA , pp.569-572, mars 1992.

[55] O. Le Blouche, "Décodage acoustico-phonétique et applications à l'indexation audio automatique", Thèse à l'Université Toulouse III – Paul Sabatier 2009

[56] Jen-Tzung Chien, "Online Hierarchical Transformation of Hidden Markov Models for Speech Recognition", IEEE transactions on speech and audio processing, vol. 7, no. 6, November 1999.

[57] S. Austin, R. Schwartz and P. Placeway, "The forward-backward search strategy for real time speech recognition", Proc. IEEE ICASSP-91, pp.697-700, Toronto, mai 1991.

[58] A. Boumalek, "Variation syntaxique en Amazighe", publication IRCAM, 2004.

[59] M. Amour, A. Bouhjar & F. Boukhris IRCAM, publication, "initiation à la langue Amazigh", 2004.

[60] Calliope, "La parole est son traitement automatique", édition Masson, 1989.

[61] S. Frui, "An analysis of long-term variation of feature parameters of speech and its application to talker recognition", Electron Communication, volume 57-A, pages 34-42, 1977.

[62] A. J. Robinson and F. Fallside, "A recurrent error propagation network speech recognition system", Computer Speech and Language, vol.5, pp. 257-286, 1991.

[63] L. J. Boe, "L'identification juridique de la voix : le cas français, historique, problématiques et proposition", Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C), pages 222-239, Avignon (France), Avril 1998.

[64] S. Boudjellal, "Détection et identification des personne par méthode biométrique", Université Tizi-ouzou (Algérie) 2005.

[65] BS. Atal, "Automatic recognition of speakers from their voices", IEEE transaction, volume 64(4), pages 460-475, 1976.

[66] P. Delacourt, "La segmentation et le regroupement par locuteurs pour l'indexation de document audio", Thèse de doctorat, Institut Eurecom, Nice (France), 2000.

[67] Rosenberg et al, I. Magrin-chagnolleau, "Speaker detection in broadcast speech databases". International Conference on Spoken Language Processing (ICASSP), pages 81-84, Atlanta (USA), 1996.

[68] A. Martin, M. Przybocki, "The NIST 1999 speaker recognition evaluation an overview". Digital Signal Processing (DSP), a review journal – Special issue on NIST 1999 speaker recognition workshop, 10(1-3), 2000.

[69] S. E. Frederickson, "Radial Basis functions for speaker identification. Workshops on Automatic Speaker Recognition, Identification, Verification", pages 107-110, Martigny (Suisse), Avril 1994.

[70] G.R. Doddington, "Speaker recognition evaluation methodology-An overview and

perspective", Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C), pages 60-66, Avignon (France), Avril 1998.

[71] C. Srividya1, "Speaker identification using cepstrum in Kannada language", Physics Section Forensic Science Laboratory Bangalore, Forensic Science Journal, 2011.

[72] A. Ouzounov, "Cepstral Features and Text-Dependent Speaker Identification A Comparative Study", CYBERNETICS AND INFORMATION TECHNOLOGIES Volume 10, No 1 2010.

[73] J. Oglesby J. S. Mason, "Optimization of neural models for speaker identification". International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 261-264, 1990.

[74] L. Boves, "Commercial application of speaker verification: overview and critical success factors'. Workshop on Speaker Recognition and its Commercial and Forensic Application (RLA2C) ", pages 150-159, Avignon (France), Avril 1998.

[75] K. Sonmez, Heck L. P., "Speaker tracking and detection with multiple speakers". European Conference on Speech Communication and Technology (Eurospeech), Budapest (Hongrie), September 1999.

[76] P.Delacourt, T. Merlin, "Différentes stratégies pour le suivi du locuteur". Reconnaissance des formes et intelligence Artificielle (RFIA), pages 123-129, Paris (France) , 2000a.

[77] S. Furui, "Cepstral analysis technique for automatic speaker verification", IEEE Transaction Acoustics, Speech, and Signal Processing (ASSP), volume 29(2), pages 254-272, Avril 1981.

[78] S. Van Vuuren, "Comparison of text-independent speaker recognition methods on telephone speech with acoustic mismatch", International Conference on Spoken Language Processing (ICSLP), pages 1788-1791, Philadelphia (USA), 1996.

[79] M. M.Hamayounpou, "Vérification vocal d'identité : dépendant et indépendant du texte", Thèse de doctorat, Université de Paris-sud centre d'Orsay, Paris (France) 1995.

[80] D. A. Reynolds, "Experimental evaluation of features for robust speaker identification", IEEE transactions Speech Audio Processing, volume 2, pages 639-643, 1994.

[81] D. Charlet, "Authentification vocale par téléphone en mode dépendant du texte". Thèse de doctorat, Ecole Nationale Supérieur des Télécommunication (ENST), Paris(France), 1997.

[82] Y. Mami, "Reconnaissance du locuteur par localisation dans un espace de locuteurs de référence", Ecole Supérieur de Télécommunication, Paris 2003.

[83] J. Andén, S. Mallat, "Scattering representation of modulated sounds", Proc. of the 15th Int. Conference on Digital Audio Effects (DAFx-12), York, UK , September 17-21,

2012.

[84] S. Sigurdsson, Kaare Brandt Petersen and Tue Lehn-Schiøler, "Mel Frequency Cepstral Coefficients : An Evaluation of Robustness of MP3 Encoded Music", Informatics and Mathematical Modelling Technical University of Denmark Richard Petersens Plads - Building 321 DK-2800 Kgs. Lyngby – Denmark, 2002.

[85] Atal B.S, “Automatic recognition of speakers from their voices”, IEEE transaction, vol(64), pages 460-475, 1979.

[86] M. Islam, "A Novel Approach for Text-Independent Speaker Identification Using Artificial Neural Network". International Journal of Innovative Research in Computer and Communication Engineering Vol. 1, Issue 4, June 2013.

[87] R.M. Gray, “Vector quantization”, IEEE ASSP Mag., vol.1(2) : 4-29,1984.

[88] Alexandre PRETI, "Surveillance de réseaux professionnels de communication par la reconnaissance du locuteur", these à l’Université d’Avignon et des Pays de Vaucluse 2008.

[89] Dempster, "ML from incomplete data via the EM algorithm", Journal of acoustical society of America (JASA), volume 39, pages 1-38, 1977.

[90] C. J. Leggetter and P. C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models”, Computer Speech and Language, 171–185, 1995.

[91] C.H. Lee, C. H. Lin, et B.H Juang, “Astudy on speaker adaptation of continuous density HMM parameters”, proc.IEEE Int. Conf. on Acoustic Speech and signal processing, pp 145-148, Albuquerque, New Mexico, ICASSP 90.8, Avril 1990.

[92] G. Celux, J. Clairambault, “Estimation de chaines de Markov cachées : méthodes et problèmes”, Journées thématiques CNRS sur les approches markoviennes en signal et images, septembre 1992.

[93] S. Renals, N. Morgan, and H. Bourlard, "Probability estimation by feed-forward networks in continuous speech recognition", Tech. Rep. TR-91-030, International Computer Science Institute, Berkeley, 1991.

[94] D. A. Reynolds and C. Rose, "Robust text-independent speaker identification using GMM", IEEE transaction on speech recognition vol. 3 January 1995.

[95] **PPS. Subhashini**, M.S. Sairam , D. Srinivasarao, "Speaker Identification with Back Propagation Neural Network Using Tunneling Alogorithm", International Journal of Engineering Trends and Technology (IJETT) – Volume 9 Number 3 - Mar 2014.

[96] Md. Ali Hossain, Md. Mijanur Rahman, Uzzal Kumar Prodhan, Md. Farukuzzaman Khan, "Implementation Of Back-Propagation Neural Network For Isolated Bangla

Speech Recognition", International Journal of Information Sciences and Techniques (IJIST) Vol.3, No.4, July 2013.

[97] N. Pushpa, R. Revathi, C. Ramya, S. Shahul Hameed, "Speech Processing Of Tamil Language With Back Propagatsion Neural Network And Semi-Supervised Training", International Journal of Innovative Research in Computer and Communication Engineering, Vol.2, Special Issue 1, March 2014.

[98] Y. Benani, "Connexionist approach for automatic speaker recognition", International conference on Acoustic Signal and Signal processing (ICASSP), pages 265-268, 1990.

[99] H. Hattori, "Text-independent speaker recognition using neural networks". International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 153-156, San Francisco(USA), ,1992.

[100] J. Oglesby J. S. Mason, "Optimization of neural models for speaker identification". International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 261-264, 1990.

[101] S. Renals, N. Morgan, M. Cohen, and H. Franco, "A real time recurrent error propagation network word recognition system", Computer Speech and Language, vol. 5, pp. 617-620, New York, USA, , March 1992.

[102] I. Trancoso, "Speaker Recognition Experiments using Connectionist Transformation Network Features", INTERSPEECH 2010.

[103] T. Robinson, "A real time recurrent error propagation network word recognition", Computer Speech and Language, vol. 5, pp. 617-620, New York, USA, March 1992.

[104] K. J. Lang and A. H. Waibel, "A time-delay neural network architecture for isolated word recognition", Neural Networks, vol.3, pp. 23-43, 1990.

[105] J. Tierney, "A study of LPC analysis of speech in additive noise". IEEE transactions on Acoustics Speech and Signal Processing 28(4), 389– 397, 1990.

[106] J. Hennebert, M. Hasler et H. Dedieu, "neural networks in speech recognition", Department of Electrical Engineering Swiss Federal Institute of Technology 1015 Lausanne, Switzerland.

Annexe 1 : Preuve des formules des paramètres LPC

Nous avons le modèle de la régression linéaire :

$$x_n = \sum_{i=1}^{T} a_i . x_{n-i} + \epsilon_n \qquad \textbf{5.1}$$

On pose :

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_T \end{pmatrix}, X^1 = \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{T-1} \end{pmatrix}, \ldots, X^p = \begin{pmatrix} x_{1-p} \\ x_{2-p} \\ \vdots \\ x_{T-p} \end{pmatrix}, \text{ et } \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_1 \\ \vdots \\ \epsilon_T \end{pmatrix}$$

Avec : $x_k = 0$ si $k<0$, et $\sum_{i=1}^{T}(\epsilon_i)^2 = \sigma^2$

Le modèle 5.1 devient alors :

$$X = \sum_{i=1}^{p} a_i . X^i + \epsilon \qquad \textbf{5.2}$$

Les $(a_i)_{i=1,\ldots,p}$ qui vérifie l'équation 5.2 sont les solutions du problème de minimisation suivant :

$$\min_{a_1,\ldots,a_p} \sum_{n=1}^{T}(\epsilon_n)^2 = \min_{a_1,\ldots,a_p} ||\epsilon||^2 \qquad \textbf{5.3}$$

Ce qui revient à dire :

$$\min_{a_1,\ldots,a_p} dist(X, \sum_{i=1}^{p} a_i . X^i) \qquad \textbf{5.4}$$

Avec :

- ✓ $||x||$ la norme euclidienne de x dans $\mathbb{R}^T$.
- ✓ $dist(.,.)$ est la distance euclidienne dans $\mathbb{R}^T$.

La solution de ce problème noté $\hat{a}_1, \ldots, \hat{a}_p$ vérifie :

$\sum_{i=1}^{p} \hat{a}_i . X^i$ est la projection orthogonale de X sur l'espace engendré par les vecteurs $X^1, \ldots, X^p$, donc nous avons :

- ✓ $\epsilon \perp X^i$ pour $i=1,\ldots,p$.

En suite on note par :

- ✓ $< X, Y >= \sum_{n=1}^{T} x_n . y_n$ le produit scalaire des deux vecteurs X et Y dans $\mathbb{R}^T$.
- ✓
- ✓ $R_{ij} = R_{ji} =< X^i, X^j >$.

$\epsilon \perp X^i$ implique que :

$$< \epsilon, X^i >= 0 =< X - \sum_{k=1}^{p} \hat{a}_k . X^p, X^i >=< X, X^i > - \sum_{k=1}^{p} \hat{a}_k . < X^k, X^i >= 0$$

Si on note par $R_{i0} =< X, X^i >$, et on pose $\hat{a}_0 = -1$ l'équation 5.3 devient :

$$0 = -\hat{a}_0 . R_{i0} - \sum_{k=1}^{p} \hat{a}_k . R_{ik}$$

Donc pour *i=1,…,p* nous avons :

$$\sum_{k=0}^{p} \hat{a}_k . R_{ik} = 0$$

De plus nous avons :

$$< \epsilon, \epsilon \geq < \epsilon, X - \sum_{k=1}^{p} \hat{a}_k . X^k \geq < \epsilon, X \geq < X - \sum_{k=1}^{p} \hat{a}_k . X^k, X >$$

$$=< X, X > - \sum_{k=1}^{p} \hat{a}_k . < X^k, X >$$

$$\sigma^2 =< \epsilon, \epsilon >= -\hat{a}_0 . R_{00} - \sum_{k=1}^{p} \hat{a}_k . R_{k0} = -\sum_{k=0}^{p} \hat{a}_k . R_{k0} \quad \textbf{5.5}$$

Annexe 2 : Algorithme Forward-Backward

L'algorithme de Forward-Backward permet de simplifier la tâche de calcul de probabilité d'observation. cet algorithme est basé sur la division de l'observation en deux partie successives. La première partie du début jusqu'à l'instant t et la deuxième de l'instant t+1 jusqu'à la fin de l'observation. L'algorithme de Forward se base sur un paramètre α qui est la probabilité d'observation à partir de l'instant 1 jusqu'à l'instant t, et l'algorithme de Backward se base sur un paramètre β qui est la probabilité d'observation de l'instant T jusqu'à l'instant t+1 dans un processus arrière. Ci-dessous les deux algorithmes :

Algorithme 5.1 : algorithme Forward-Backward

Algorithme de Forward :

Pour i=1, n faire

$$\alpha_1(i) = \Pi_i b_i(O_i)$$

Fin pour

Tant que t<T faire

j←1

Tant que j≤n faire

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

j←j+1

fin tant que

t←t+1

fin tant que

$$P(O|\wedge) \leftarrow \sum_{i=1}^{n} \alpha_T(i)$$

Algorithme de Backward :

Pour i=1,n faire

$$\beta_T(i) \leftarrow 1$$

Fin pour

t←T

tant que t>1 faire

j←1

Tant que j≤n faire

$$\beta_t(i) \leftarrow \sum_{i=}^{n} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)$$

j←j-1

Fin tant que

t←t-1

fin tat que

$$P(O|\wedge) = \sum_{i=1}^{n} \beta_i(i)$$

Finalement la probabilité d'une séquence d'observations est obtenue en prenant les valeurs de α et β à un instant t quelconque : $P(O|\wedge) = \sum_{i=1}^{n} \alpha_t(i)\beta_t(i)$. Cependant on utilise le plus souvent les valeurs obtenues pour deux cas particuliers (t=0) ou (t=T), ce qui donne :

$$P(O|\wedge) = \sum_{i=1}^{n} \alpha_T(i) = \sum_{i=1}^{n} \alpha_t(i)\beta_t(i)$$

Annexe 3 : Quelques unités de base

Le phonème :

L'avantage principal des phonèmes réside dans le nombre (de 30 à 40 selon les langages) et dans la facilité de coder les mots du lexique en une chaine phonétique. Certains phonèmes sont fortement influencés par les contextes phonétiques avoisinants. De ce fait il n'est pas facile de les identifier. Un phonème est caractérisé par plusieurs segments acoustiques.

Le phone :

Afin de remédier au problème d'influence contextuelle, on définit des unités de taille plus petites que les phonèmes représentant une homogénéité acoustique. Ces phones homogènes caractérisent les gestes articulatoires successifs. Ces unités sont en général de taille inférieure aux phonèmes (un phonème est constitué de plusieurs phones).

Le diphone :

Le diphone est une unité sub-lexicale contenant les transitions entre deux phonèmes successifs. Cette unité correspond à un segment de parole allant du milieu d'un phonème au milieu du phonème suivant. Donc cette unité inclut certaines règles de coarticulation. L'inconvénient principal de ces unités reste leur nombre très élevé à représenter.

La représentation du mot "deux" en diphones est :

Deux= d.d eu.eu

L'allophone :

L'allophone est une réalisation acoustique particulière d'un phonème qui dépend du contexte phonétique gauche et/ou droit. Le triphone est un modèle d'allophone qui dépend simultanément du deux contextes phonétique gauche et droit. L'avantage de ces modèles est qu'ils modélisent mieux les phénomènes de la coarticulation. La représentation des mots par allophones crée un très grand nombre de combinaison, si nous avons *n* phonèmes différents, alors le nombre d'allophones est :

- ✓ n^2 dans le cas de la prise en compte d'un seul contexte phonétique.
- ✓ n^3 le cas de la prise en compte des deux contextes.

Le fénone :

Le fénone est un modèle acoustique très simple, il est caractérisé par un modèle constitué d'un seul état et une seule densité de probabilité d'observation acoustique. Le niveau phonétique devient superflue, on passe directement des fénones au niveau mot.

La syllabe :

C’est une unité constituée de consonnes et de voyelles (CV, VC, CVC). Parmi les avantages de cette unité on cite :

- ✓ Le noyau syllabique est assez facile à localiser.
- ✓ L’information de coarticulation est incluse dans la représentation syllabique.
- ✓ 5=sin.k ou sin.ke

Annexe 4 : Hypothèses sur les MMCs

L'utilisation des MMC est toujours liée à des hypothèses qu'il faut fixer, parmi ces hypothèses nous citons :

Si on prend l'équation de l'approche statistique :

$$M^* = arg\max_{M} \frac{\Pr(Y|M) * \Pr(M)}{\Pr(Y)}$$

H1. Dans cette équation la Pr(M) est calculée indépendamment des observations Y, Cette probabilité est estimée sur un modèle de langage.

H2. Pr(Y) est supposée constante et indépendante du modèle M (les paramètres du modèle global sont fixés sur l'ensemble d'apprentissage, ils sont indépendants de M).

H3. Les modèles de Markov cachés sont supposés du premier ordre :

$$\Pr(X_{t+1} = q_j | X_t = q_i, \dots., X_1 = q_1) = \Pr(X_{t+1} = q_j | X_t = q_i) = a_{ij}$$

H4. Les observations sont supposées indépendantes les unes des autres conditionnellement à la suite d'états, et chaque observation ne dépend que de l'état courant (c-à-d) :

$$b_i(y_t) = \Pr(Y_t = y_t | Y_{t-1} = y_{t-1}, \dots, Y_1 = y_1, X_t = q_i, X_{t-1} = q_{i-1}, \dots, X_1 = q_1)$$

$$= \Pr(Y_t = y_t | X_t = q_i)$$

Annexe 5 : Algorithme k-moyenne

L'algorithme de quantification vectorielle ou k-moyennes est un processus itératif qui permet de rassembler un nuage de points en k classes. Cet algorithme est utilisé pour l'initialisation des paramètres de classes gaussiennes en cas de reconnaissance automatique de la parole :

Algorithme 5.2 : k-moyenne

Initialisation

On se donne un dictionnaire D_0 de taille K.

Construction de la partition

On possède le dictionnaire $D_t=\{D_{it}\}_{i=1,k}$ après t étapes. On cherche la partition qui minimise l'erreur de quantification associée à D_t composée des classes c_{it} : $\min_{i=1,k} d(y_n, \mu_{it})$.

L'erreur de quantification vaut : $D_t = \frac{1}{N}\sum_{n=1}^{N}[\min_{k=1,k} d(y_n, \mu_{it})]$.

Test d'arrêt

Si (par exemple) : $\frac{D_{t-1}-D_t}{D_t} < \varepsilon$ alors on s'arrête sinon aller en d.

Recalcule des centres de gravité

A chaque classe c_i de la partition, on associe le 'centre de gravité' D_t. Le dictionnaire est désormais composé des nouveaux D_{it} , On fait t=t+1 et on va à l' étape 2.

Annexe 6 : Algorithme de Viterbi

L'algorithme de Viterbi sert à déterminer le meilleur chemin correspondant à l'observation. C'est à dire trouver, dans le modèle ∧, la meilleure suite d'états qui maximise la quantité :

$P(Q, O|\wedge)$

Algorithme 5.3 : Viterbi

Algorithme de Viterbi :

Pour i=1, n faire
.$\delta_1(i) \leftarrow \Pi_i b_i(O_1)$

$$\Psi_1(i) \leftarrow 0$$

Fin pour
t←2
tant que t≤T faire
j←1
tant que j≤n faire

$$\delta_t(j) \leftarrow \max_{1\le i\le n}\left[\delta_{t-1}(i)a_{ij}\right]b_j(o_t)$$

j←j+1$\psi_t(j) \leftarrow Arg \max_{1\le i\le n}\left[\delta_{t-1}(i)a_{ij}\right]$
fin tant que
t←t+1
fin tant que

$$P^* \leftarrow \max_{1\le i\le n}[\delta_T(i)]$$

.

$$q_T^* \leftarrow \max_{1\le i\le n}[\delta_T(i)]$$

t←T
Tant que t≥1 faire

$$q_T^* \leftarrow \psi_{t+1}(q_{t+1}^*)$$

t←t-1

Fin tant que

Annexe 7 : Algorithme Expectation Maximisation

L'algorithme Expectation Maximisation est un algorithme itératif permettant de ré-estimer les paramètres d'un modèle à partir du modèle précédent :

Algorithme 5.4 : Expectation-maximisation

Algorithme EM :

Définir une statistique suffisante pour estimer l'ensemble $\wedge$ des paramètres cachés

Initialiser $\wedge$

p←1

Tant que $\wedge_p \neq \wedge_{p+1}$ faire

ESTIMATION

Utiliser les observations pour calculer la statistique suffisante de $\wedge_p$.

MAXIMISATION

Calculer $\wedge_{p+1}$ comme une estimation au maximum de vraisemblance de $\wedge$ à partir des résultats de l'étape p d'estimation.

p←p+1

Fin tant que

Annexe 8 : Création des MMC pour les unités de la parole

Le signal de la parole est un vecteur de valeurs réelles. Celui-ci est obtenu par une phase d'échantillonnage qui permet d'obtenir des valeurs discrètes en utilisant une fréquence d'échantillonnage donnée. Ce signal a subi une phase de traitement pour extraire les coefficients MFCC. Cette paramétrisation permet de transformer un signal en une matrice de valeurs constituant les paramètres cepstraux connus par MFCC. Ces vecteurs sont utilisés pour créer les modèles de Markov cachés de chaque unité ou syllabe. Chaque état de ce modèle est caractérisé par la moyenne et la matrice de covariance.

Moyenne :

La moyenne est définie sur chaque colonne du matrice des coefficients MFCC. On définit la moyenne par la relation suivante :

5 $$\bar{x}_c = \frac{1}{N}\sum_{n=0}^{N-1} x_c(n) \qquad ,c = column$$

Covariance :

La matrice de covariance caractérise la variation temporelle du signal de la parole. Elle est exprimée par la relation :

$$\bar{x^2}_c = \frac{1}{N}\sum_{n=0}^{N-1} x^2{}_c(n) \qquad ,c = column$$

$$\sigma^2{}_c = \bar{x^2}_c - \left[\bar{x}_c\right]^2 \qquad ,c = column$$

Annexe 9 : Algorithme VAD pour la segmentation des signaux de la parole

L'algorithme VAD permet la détection des zones de la parole. Il commence par l'encadrement du signal audio sous forme de trames. Les N Premières trames sont utilisées pour l'initialisation des seuils. Pour chaque trame d'entrée les trois caractéristiques sont calculées :

- ✓ Energie à court terme du signal
- ✓ Planéité spectrale définie par :

$SFMdb = 10\log10\ (Gm / Am)$

- ✓ La fréquence la plus dominante dans le signal.

La trame audio est marquée comme étant une trame de parole, si plus d'une des valeurs de caractéristiques tombent sur le seuil pré-calculé. La procédure complète de la proposition de cette méthode est décrite ci-dessous :

Algorithme 5.5 : Détection des activités vocales

1- initialiser le nombre de frames Frame _Size =10ms et calculer le nombre de frames (*Num_Of _ Frames*)(pas d'intersection entre frames).
2- donner les seuils initiaux pour chaque frame :
- Premier seuil pour l'énergie (Energy_PrimThresh)
- Premier seuil pour F (F _ Pr imThresh)
- Premier seuil pour for SFM (SF _ Pr imThresh)

3- pour *i* = 1 jusqu'à *Num_Of _ Frames*
3-1- Calculer (*E*(*i*)) .
3-2- Appliquer FFT pour chaque frame.
3-2-1- Trouver F(i) = argmax(S(k)) comme la fréquence la plus dominante.
3-2-2- Calculer (*SFM*(*i*)) dans chaque frame.
3-3- Supposer que les 30 premiers frames sont du silence,
Trouver la valeur minimale de *E* (*Min _ E*) , *F* (*Min _ F*)
et SFM (*Min _ SF*).
3-4- Affecter le seuil de decision de *E* , *F* et *SFM* .
- Thresh _ E = Energy _ Pr imThresh *log(Min _ E)
- Thresh _ F = F _ Pr imThresh
- Thresh _ SF = SF _ Pr imThresh

3-5- met le compteur = 0 .
- Si ((E(i) −Min _ E) >= Thresh _ E) puis compteur + + .
- Si ((F(i) −Min _ F) >= Thresh _ F) puis *Compteur* + + .
- Si ((SFM(i) −Min _SF) >= Thresh _ SF) puis *Compteur*+ + .

3-6- Si *Compteur*> 1 noter le frame courant comme parole else Note lui comme silence.
3-7- Si le frame courant est marqué comme silence, actualiser les seuils minimums :

$$Min_E = \frac{(Silence_{count} * Min_E) + E(i)}{Silence_{count} + 1}$$

3-8- Thresh _ E = Energy _ Pr imThresh *log(Min _ E)

4- Ignorer le signal dont le silence est dans le 10 frames successives.

5- Ignorer la parole pour moins de 5 frames

Annexe 10 : Algorithme d'apprentissage quantification vectorielle

Algorithme 5.6 : Apprentissage de la quantification vectorielle

C_0 : Un dictionnaire initial à M éléments,

X : L'ensemble des L vecteurs d'apprentissage,

D : Une distance de mesure de distorsion

n=0

1- Trouver une partition de X à partir du dictionnaire C_n. A chaque vecteur de X, on associe le meilleur représentant dans le dictionnaire.
2- Calculer la distorsion moyenne pour tous les éléments de X, notés d_n :

$$d_n = \frac{1}{L}\sum_{l=1}^{L} d(X_l, \hat{X}_l)$$

3- Si la distorsion est telle que $\frac{(d_{n-1}-d_n)}{d_n} \leq \xi$, alors C_n est le dictionnaire souhaité.
4- Recalculer les centres de chaque cellule pour obtenir C_{n+1}.
5- n=n+1, retour en 1.
On prend pour d la MSE (mean square error) ou moyenne des écarts au carré.

Annexe 11 : Unicode, la norme de tri et la norme de clavier pour Tifinaghe

Code Informatique	Caractère	Code Informatique	Caractère	Code Informatique	Caractère	Code Informatique	Caractère	Code Informatique	Caractère
2D30	ⴰ	2D40	ⴀ	2D50	ⵐ	2D60	ⵠ	2D70	
2D31	ⴱ	2D41	ⴁ	2D51	ⵑ	2D61	ⵡ	2D71	
2D32	ⴲ	2D42	ⴂ	2D52	ⵒ	2D62	ⵢ	2D72	
2D33	ⴳ	2D43	ⴃ	2D53	ⵓ	2D63	ⵣ	2D73	
2D34	ⴴ	2D44	ⴄ	2D54	ⵔ	2D64	ⵤ	2D74	
2D35	ⴵ	2D45	ⴅ	2D55	ⵕ	2D65	ⵥ	2D75	
2D36	ⴶ	2D46	ⴆ	2D56	ⵖ	2D66		2D76	
2D37	ⴷ	2D47	ⴇ	2D57	ⵗ	2D67		2D77	
2D38	ⴸ	2D48	ⴈ	2D58	ⵘ	2D68		2D78	
2D39	ⴹ	2D49	ⴉ	2D59	ⵙ	2D69		2D79	
2D3A	ⴺ	2D4A	ⴊ	2D5A	ⵚ	2D6A		2D7A	
2D3B	ⴻ	2D4B	ⴋ	2D5B	ⵛ	2D6B		2D7B	
2D3C	ⴼ	2D4C	ⴌ	2D5C	ⵜ	2D6C		2D7C	
2D3D	ⴽ	2D4D	ⴍ	2D5D	ⵝ	2D6D		2D7D	
2D3E	ⴾ	2D4E	ⴎ	2D5E	ⵞ	2D6E		2D7E	
2D3F	ⴿ	2D4F	ⴏ	2D5F	ⵟ	2D6F	ⵯ	2D7F	

Tifinaghe-IRCAM de base

Autres lettres Tifinaghe-IRCAM, néotifinaghes et lettres Touarègues modernes attestées

Réservé pour un codage ultérieur

Table 5.1 : Unicode, la norme de tri et la norme de clavier pour Tifinaghe

Annexe 12 : Caractères Tifinaghe

L'alphabet Tifinaghe est composé de 33 unités, la table ci-dessous présente la liste de cet alphabet ainsi que ses prononciations :

ⵄ	ⵅ	ⵃ	ⵃⵯ	ⵄ	ⵄ	ⵄ	ⵄ	ⵄ	ⵄⵯ	ⵄ
ya	yab	yag	yagʷ	yad	yaḍ	yey	yaf	yak	yakʷ	yah
a	b	g	gʷ	d	ḍ	e	f	k	kʷ	h
[a]	[b/β]	[g/ɟ]	[gʷ]	[d/ð]	[ɗ]	[e]	[f]	[k/ç]	[kʷ]	[h]
ⵄ	ⵄ	ⵃ	ⵄ	ⵄ	ⵄ	ⵄ	ⵄ	ⵄ	ⵄ	ⵄ
yaḥ	yaɛ	yax	yaq	yi	yaj	yal	yam	yan	yu	yar
ḥ		x	q	i	j	l	m	n	u	r
[ħ]	[ʕ]	[x]	[q]	[i]	[ʒ]	[l]	[m]	[n]	[u]	[r]
ⵄ	ⵄ	ⵄ	ⵄ	ⵄ	ⵄ	ⵄ	ⵄ	ⵄ	ⵄ	ⵄ
yaṛ	yagh	yas	yaṣ	yac	yat	yaṭ	yaw	yay	yaz	yaẓ
ṛ	gh	s	ṣ	c	t	ṭ	w	y	z	ẓ
[ɾ]	[ɣ]	[s]	[s]	[ʃ]	[t/θ]	[t]	[w]	[j]	[z]	[z]

Table 5.2 : Alphabet Tifinaghe

Annexe 13 : Application java

L'interface ci-dessous permet de reconnaitre un fichier contenant un chiffre Tamazight de 1 à 199 en se basant sur les règles de construction décrites dans le paragraphe 2.8.

Figure 5.1 : Interface java pour la reconnaissance des chiffres enchainés

Annexe 14 : Système de sécurité

Un locuteur client est invité à prononcer son mot de passe, dans un premier temps le mot de passe est vérifié, si celui-ci est correcte la deuxième étape consiste en vérification du locuteur. La figure ci-dessous présente les interfaces utilisées.

Figure 5.2 : Système hybride : vérification du mot de passe et vérification du locuteur

Annexe 15 : Coefficient MFCC

Lors de la création d'un système de reconnaissance automatique de la parole, il faut ajuster les informations qui seront utilisées. Les informations contenues dans le signal analogique ne sont pas utiles que dans la reconnaissance vocale en utilisant le MMC quand il est dans la forme paramétrique discrète. C'est pourquoi on effectue une conversion du signal analogique aux paramètres de Mel. La figure ci-dessous présente un aperçu sur les étapes de mise en forme et paramétrisation du signal avant l'apprentissage des MMCs.

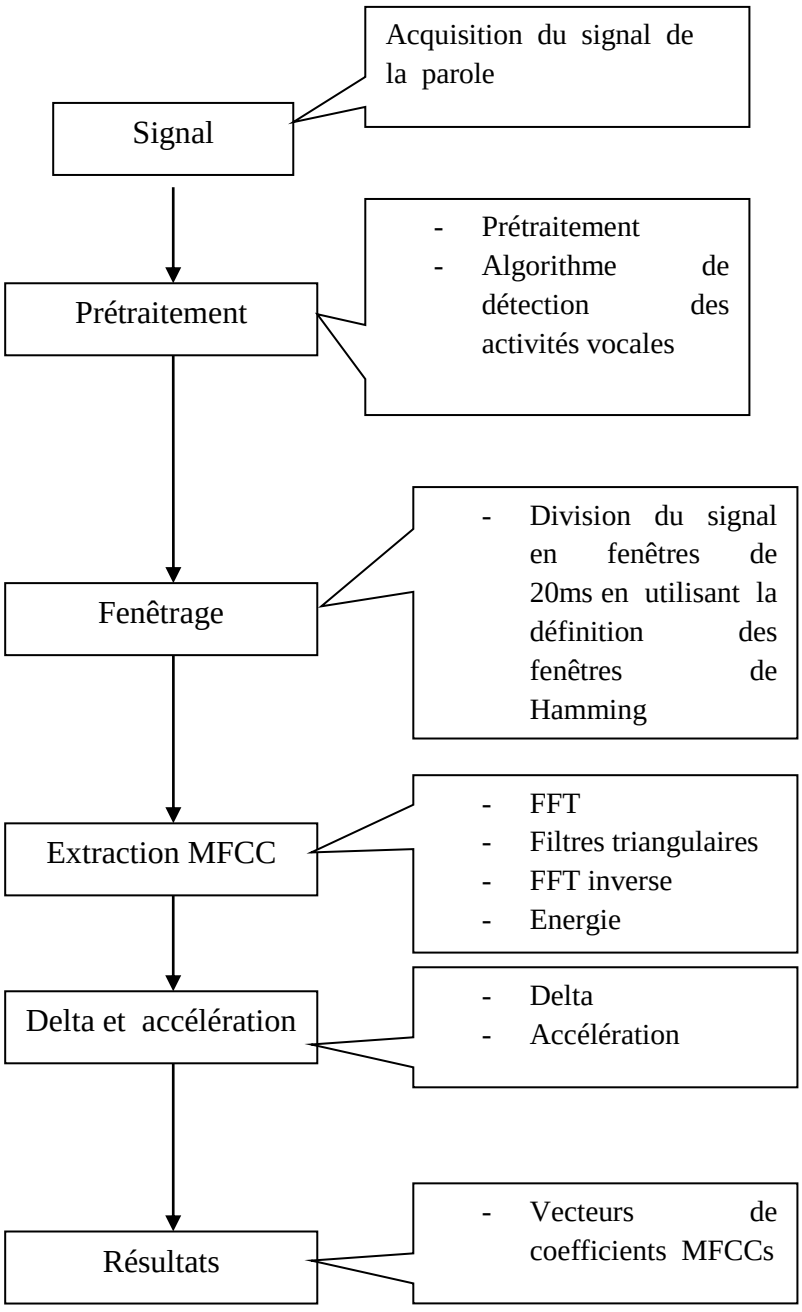


Figure 5.3 : Etapes de paramétrisation du signal de la parole

1- Signal de la parole :

Le signal analogique de la parole utilisé par le système doit être converti de l'analogique au discret x(n). La fréquence d'échantillonnage utilisée dans ce travail est f_s=16khz. La première valeur est prise à l'instant t telle que :

$$t = \frac{1}{f_s}$$

La figure ci-dessous présente un exemple d'un signal échantillonné :

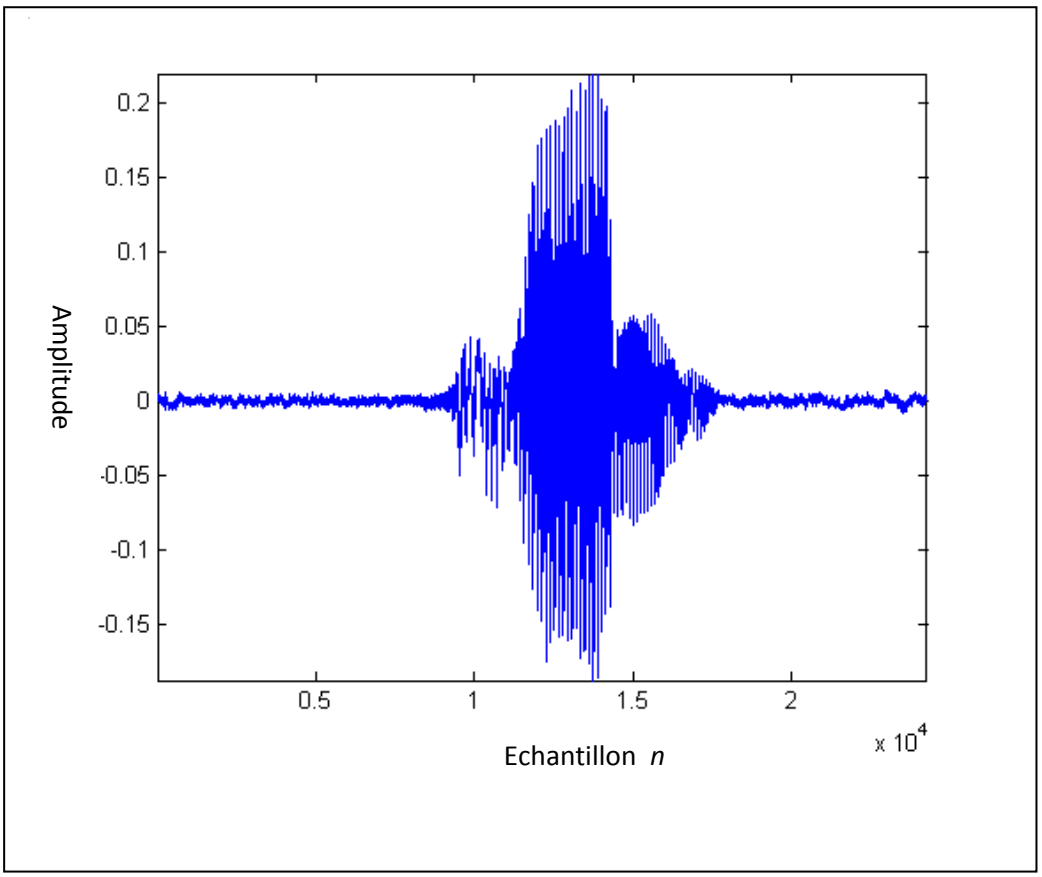


Figure 5.4 : Signal de la parole échantillonné

2. Prétraitement

Le signal de la parole doit être aplati spectralement, cette opération se fait en utilisant un filtre passe-haut qui élimine les composantes de hautes fréquences dans le signal et il est décrit par l'équation suivante :

$$h(n) = \{1, -0.95\} \qquad 5.6$$

La figure ci-dessous présente un exemple d'un signal soumis à un filtre passe-haut :

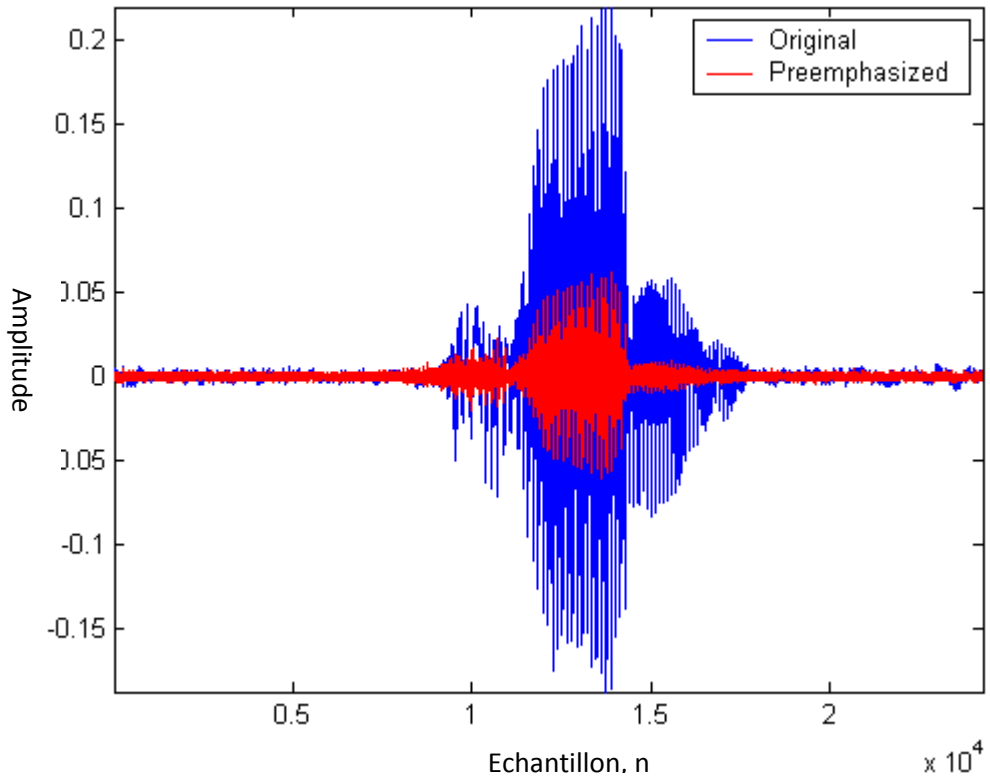


Figure 5.5 : Signal original et signal soumis à un filtre passe hout

Si x(t) le signal échantillonné et h(n) le filtre passe-haut, alors :

$$x_p = x(n) \otimes h(n) \qquad \textbf{5.7}$$

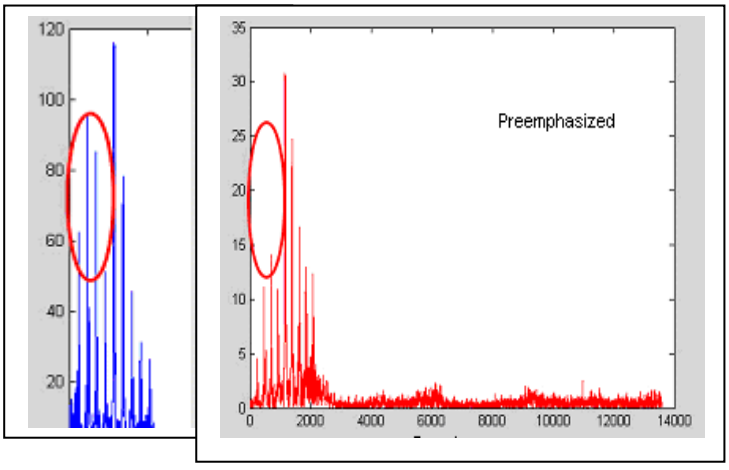


Figure 5.6 : Application du filtre passe-haut

3. Détection des activités vocales

Cet algorithme permet la segmentation du signal et l'élimination des zones de bruit, pour cela on utilise l'algorithme 2.1 dans le chapitre 2.

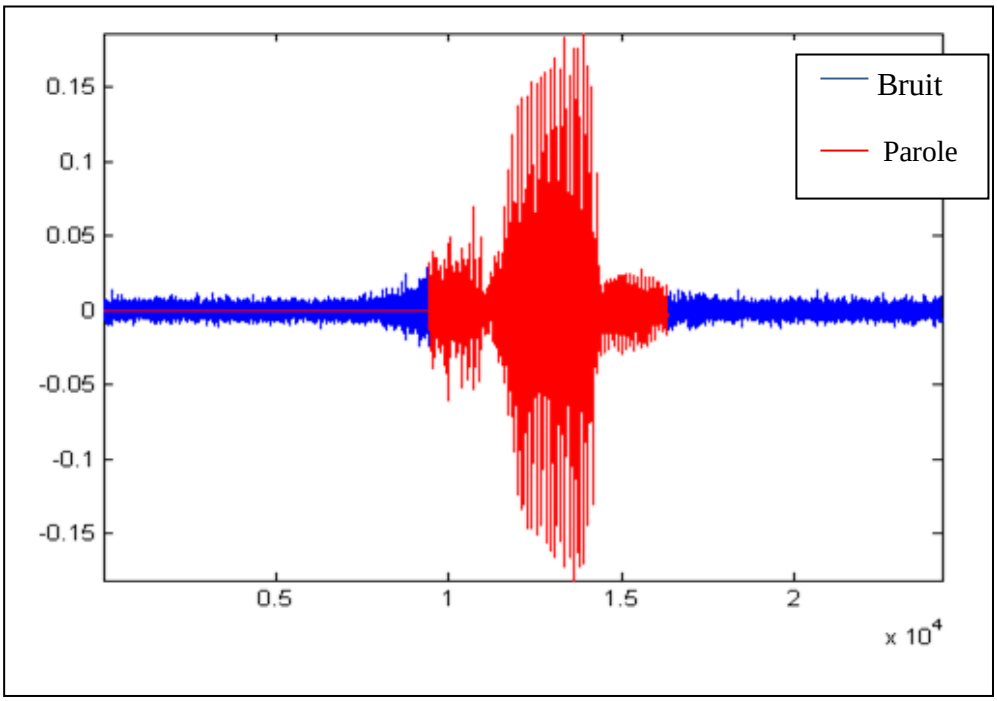


Figure 5.7 : détection des activités vocales

4. Blocage de cadres et Fenêtrage de Hamming

Le blocage de cadres consiste à diviser le signal en une matrice avec une longueur de temps appropriée pour chaque cadre. En raison de l'hypothèse selon laquelle un signal dans une trame de 20 ms est stationnaire et une fréquence d'échantillonnage à 16000Hz donnera le résultat de 320 échantillons pour chaque cadre.

$$T = \frac{1}{f} = \frac{1}{16000} = 0{,}625\text{ms}$$

$$\text{nombre}_{\text{échantion}} = \frac{20}{0{,}0625} = 320$$

Dans le cas de l'utilisation de fenêtrage un chevauchement de 62,5% va donner un facteur de séparation de 120 échantillons.

Après le blocage de cadres, le fenêtrage de Hamming est appliqué à chaque cadre. Cette fenêtre permet de réduire la discontinuité du signal au niveau des extrémités de chaque bloc.

L'équation qui définit le fenêtrage de Hamming est la suivante :

$$w(k) = 0{,}54 - 0{,}46\cos\left(\frac{2\pi k}{K-1}\right) \qquad Eq.4.3 \qquad 5.8$$

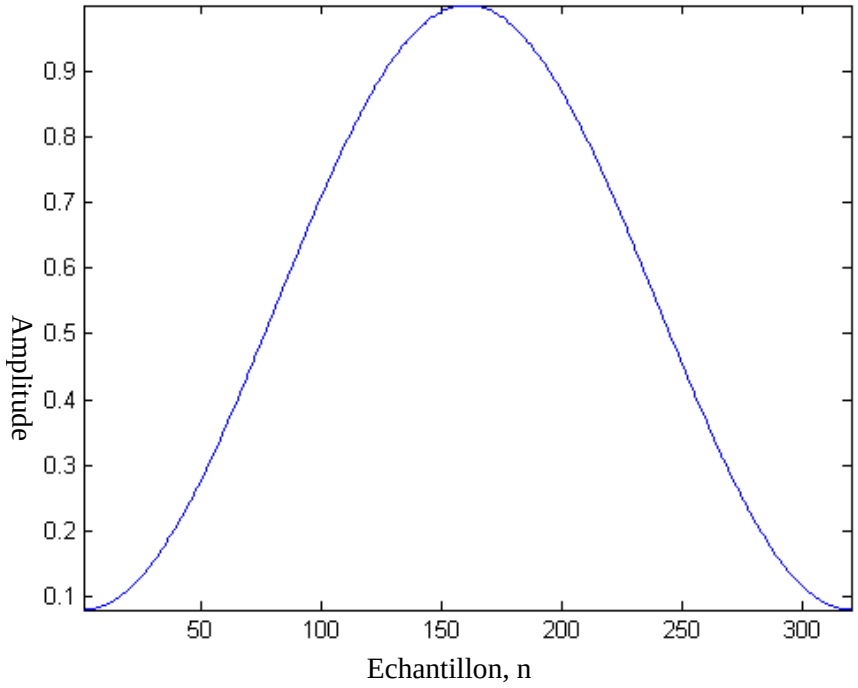


Figure 5.8 : Fenêtre de Hamming

La figure 5.9 ci-dessous présente le blocage de cadres, le block numéro 10 du mot ‘sin’ :

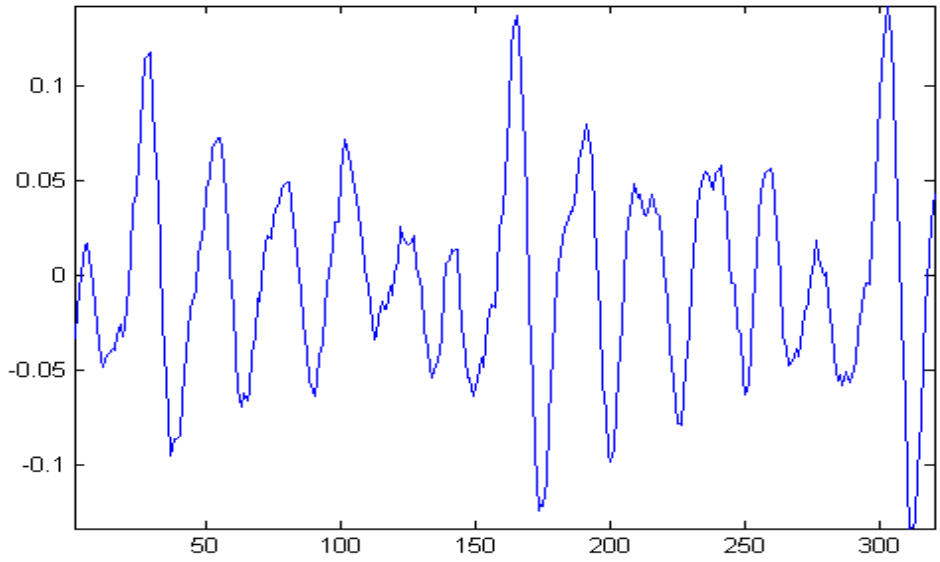


Figure 5.9 : blocage de cadres du signal de la parole (bloc 10 du mot 'sin')

L'application du fenêtrage de Hamming sur ce bloc donne l'allure sur la figure ci-dessous :

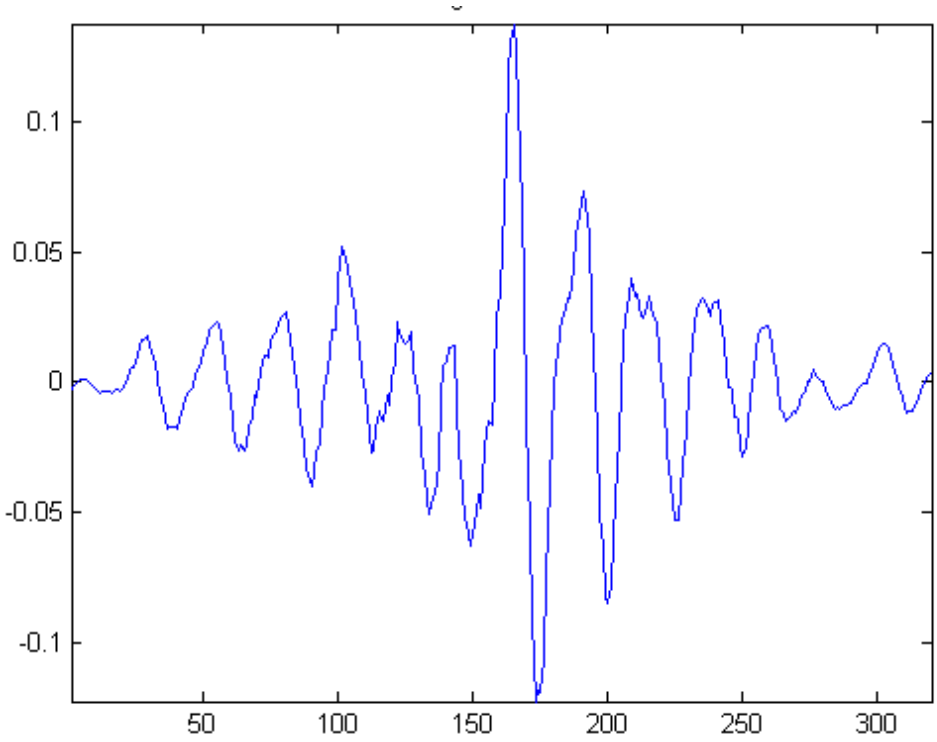


Figure 5.10 : Application du fenêtrage du Hamming au block 10 du mot 'sin'

Le résultat donne la réduction de la discontinuité à la fin du bloc.

5. Extraction des caractéristiques MFCC

5.1. FFT sur chaque bloc

On utilise 512 points pour la transformée de Fourier rapide, pour ajuster la langueur de 20ms de la fenêtre on ajoute les zéro. La figure ci-dessous présente un exemple de FFT sur une fenêtre de 20ms :

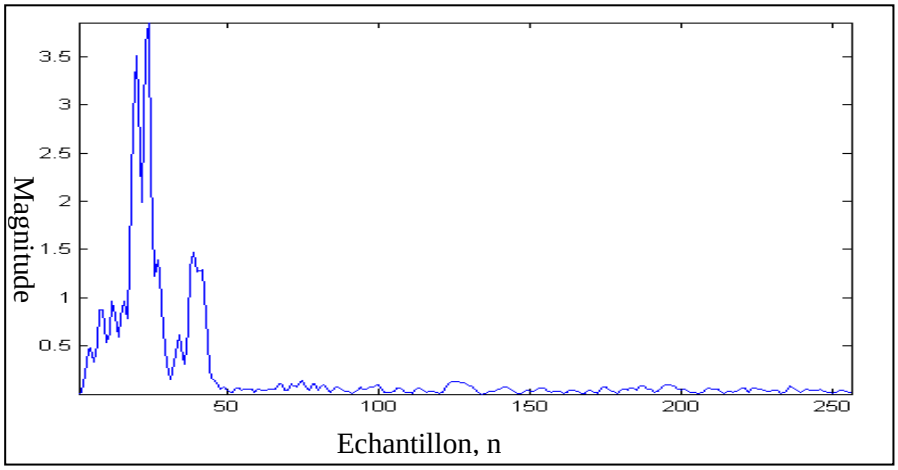


Figure 5.11 : transformée de Fourier sur un bloc de 20ms

5.2. Banc de filtre

Puisque la perception du son par l'oreille humaine n'est pas linéaire, il faut utiliser une échelle cartographique (mapping scale). L'échelle utilisée et celle de Mel. Cette échelle est un gauchissement de la fréquence de pitch mesurée à la fréquence correspondante en échelle de Mel. La conversion de l'échelle des fréquences à l'échelle de Mel est faite en utilisant les relations suivantes [4]:

$$F_{mel} = 2595 \cdot \log_{10}(1 + \frac{F_{Hz}}{700}) \qquad Eq.4.4 \qquad \textbf{5.9}$$

$$F_{Hz} = 700 \cdot (10^{\frac{F_{mel}}{2595}} - 1) \qquad Eq.4.5 \qquad \textbf{5.10}$$

Le gauchissement pratique est fait en utilisant un banc de filtres triangulaires répartis selon l'échelle de Mel qui gèrent la déformation de la fréquence entre le Hz et le Mel, la figure ci-dessous dans une illustration de ces filtres :

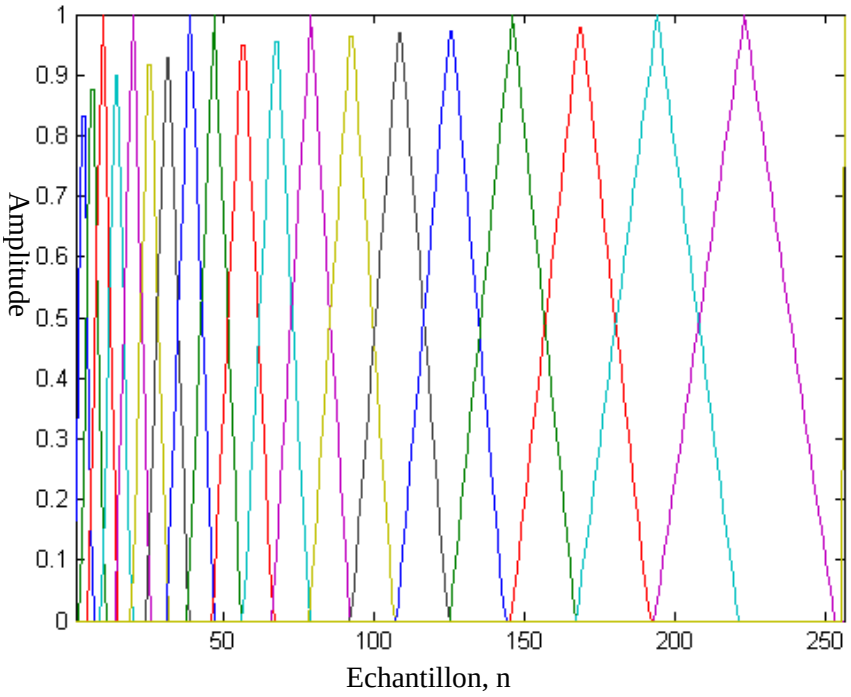


Figure 5.12 : Banc de filtres de Mel

Théoriquement, la sommation est faite pour calculer la contribution de chaque filtre. Ceci donne une matrice avec un même nombre de colonnes que l'ensemble de filtres. La transformée de Fourier de chaque frame est multipliée par chaque filtre (dans notre cas 12). Ceci donne un vecteur de 12 éléments.

5.3. Transformée de Fourier inverse (IFT)

L'application de la transformée de Fourier inverse donne directement les coefficients de Mel, elle est donnée par la relation 2.11 (chapitre II) :

$$C_k = \sum_{i=1}^{F} Log(e_i) \cos\left(\frac{\pi k(i-0.5)}{F}\right) \quad k = 1, \dots, d \qquad \textbf{5.11}$$

6. Energie du signal

Le troisième coefficient ajouté au vecteur de 12 éléments est l'énergie du signal, elle est donnée par la relation suivante :

$$E_m = \log \sum_{k=0}^{K-1} x\_windowed^{\,2}(k;m) \qquad Eq.4.9 \qquad \textbf{5.12}$$

7. Dérivée et accélération des coefficients

Les coefficients delta et l'accélération sont calculés pour accroître l'information de la perception humaine. Les coefficients delta sont calculés sur le décalage horaire, les coefficients d'accélération sont de la seconde dérivée dans le temps. Ces deux coefficients sont calculés selon les relations suivantes :

$$\delta^{[1]} = \frac{\sum_{p=-P}^{P}(c_h(n;m+p)-c_h(n;m))p}{\sum_{p=-P}^{P}p^2} \qquad Eq.4.10 \qquad 5.13$$

$$\delta^{[2]} = 2\cdot\left(\frac{\sum_{p=-P}^{P}p^2\sum_{p=-P}^{P}c_h(n;m+p)-(2P+1)\sum_{p=-P}^{P}c_h(n;m+p)p^2}{(\sum_{p=-P}^{P}p^2)^2-(2P+1)\sum_{p=-P}^{P}p^4}\right) \qquad Eq.4. \quad 5.14$$

Avec :

C_h : matrice du coefficient pour un mot donné

8. Normalisation

Les vecteurs acoustiques obtenus sont normalisés pour obtenir une moyenne nulle est une matrice de variance unité. Le vecteur des moyennes peut être calculé en utilisant la relation suivante :

$$\bar{f}_{\mu}(n) = \frac{1}{M}\sum_{m=0}^{M-1}x\_mfcc(n,m) \qquad Eq.4.14 \qquad 5.15$$

Pour normaliser les coefficients, on utilise la formule suivante :

$$\bar{f}_{\mu}(n;m) = x\_mfcc(n,m) - \bar{f}_{\mu}(n) \qquad Eq.4.15 \qquad 5.16$$

Annexe 16 : théorème de Shannon

1- Cas d'un signal sinusoïdal :

Il faut prendre au moins deux échantillons par période du signal, donc $T_e < T_0/2$. La fréquence d'échantillonnage doit être au moins deux fois supérieur à la fréquence du signal :$F_e > 2.f_0$.

2- Signaux à spectre limité :

Un signal à spectre limité est un signal dont toutes les fréquences sont comprises entre une fréquence minimale f_{min} et une fréquence maximale f_{max}.

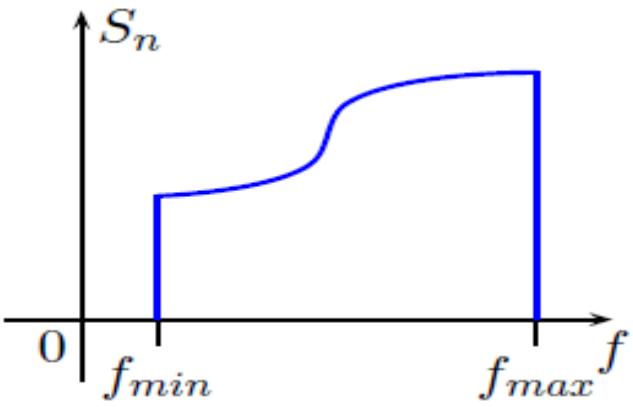


Figure 5.13 : Signal à spectre limité

Exemple : le son [20Hz ; 20kHz]

Théorème de Shannon : La fréquence d'échantillonnage doit être supérieure à deux fois la fréquence la plus élevée d'un signal à spectre limité :

$$F_e > 2.f_{max}$$

Dans le cas contraire, il y a perte d'informations et déformation du signal reconstitué.

Annexe 17 : Bibliographie personnelle

[1] - Revue Méditerranéenne de la Télécommunication vol.1 num. 2 juillet 2011- Reconnaissance de la parole Amazigh à base de l'alphabet Tifinaghe en utilisant le modèle de Markov caché.

[2] - Global journal August 2011 Vol 11 num 15, Speech recognition using HMM concerning the Moroccan dialect DARIJA.

[3] - IJEST February 2012 Volume 4 Issue 2, "AUTOMATIC SPEECH RECOGNITION SYSTEM CONCERNING THE MOROCCAN DIALECTE (Darija and Tamazight)"

[4] - IJAST sv publishers:"Password verification system based on automatic speech recognition concerning the Moroccan dialect", Vol 5, and Num 2, AUGUST 2012.

[5] - ICMCS Ouarzazate 2010/2011**:** Speech recognition using hidden Markov model (HMM) and comparison of results obtained with dynamic programming.

[6] - SITACAM Agadir 2010/2011 **:** Reconnaissance de la parole Amazigh à base de l'alphabet Tifinagh en utilisant le modèle de Markov caché.

[7] - JMIT'11 FST Beni Mellal : Système de reconnaissance automatique de la parole basé sur le modèle de Markov Caché'.

[8] - SITACAM'13 Beni Mellal : Système de vérification de mots de passe basé sur la reconnaissance automatique des dialectes Darija et Tamazight.