République Algérienne Démocratique et Populaire

Université A. MIRA de Béjaia

Faculté des Sciences Exactes
Département de Recherche Opérationnelle

Mémoire présenté pour l'obtention du diplôme de master

Spécialité : Sciences de Données et Aide à la Décision

Création et implémentation d'un data-set en Tamazight sur une plateforme de reconnaissance vocale

Présenté par :
ADRAR Lisa
SAADI Narimane

Sous la direction de : Dr. A.ZAIDI et Dr. L Asli

Défendu le 02/07/2024, devant le jury composé de :

Mr.K. ABBAS	Professeur	Président de jury	UAMB - Bejaia.
Mr.M.AZZI	M.C.B	Examinateur	UAMB - Bejaia
Mr.M.M'HAMEDI	M.C.A	Examinateur	UAMB - Bejaia.

Année Universitaire 2023 – 2024

Remerciements

Louange au bon Dieu le miséricordieux

On tient à remercier M.ZAIDI.A autant que encadreur pour son acceuil chaleureux au niveau de Centre de Recherche de la langue et la culture Amazighe, ses conseils et son aide tout au long de ce travail.

On remercie également M.ASLI.L autant que co-encadreur de ce mémoire pour son encouragement et son aide précieux Malgré ses multiples activités et missions.

On remercie infiniment les membres de jury pour leur temps,leur présence et pour leur lecture attentive pour notre travail.

On tient à adresser nos remerciements pour nos parents, nos frères et nos soeurs pour leurs soutiens et encouragements.

Nos remerciements à tous nos ami(es) et camarades et toutes personne qui nous a aidé.

Nos remerciements à tous les enseignants durant tout nos parcours du primaire à l'université.

Dédicace

À mes chers parents,

Les gardiens de mon cœur, merci pour votre amour inconditionnel et votre soutien constant. Vous m'avez inculqué des valeurs essentielles et m'avez toujours encouragé à poursuivre mes rêves.
Sans vous, ce travail n'aurait pas été possible.

À ma soeur chanez

Pour ta croyance en moi. Ton amour inconditionnel, ton aide précieuse et ton encouragement constant ont été d'un soutien inestimable tout au long de ce parcours.
À toi, qui m'as montré l'amour, la force et la douceur.

À mon frére hani

Pour ta présence constante à mes côtés et ton amour sincère.Merci pour tes encouragements et pour croire en moi même quand j'avais des doutes.

À ma chère famille et mes chères cousines,

Merci pour les moments de joie partagés, pour votre aide précieuse et pour être une source constante de motivation et de bonheur.

À mes chers binômes de stage du Centre de Recherche de Tamazight,

Ce mémoire est le fruit de notre travail acharné, de notre collaboration et de notre dévouement. Ensemble, nous avons surmonté les défis et célébré les réussites.

À mes camarde et mes amis

Je vous remercie pour votre patience et pour m'avoir aidé à avancer. Vous êtes tous pour moi comme une seconde famille.
Merci d'être toujours près de moi dans mes joies et mes peines.

L.ADRAR

Dédicace

Je dédie cet humble et modeste travail avec grand amour,sincérité et fierté :

À mes chers parents, source de tendresse et noblesse, puisse cet étape constituer pour vous un motif de satisfaction.

À mes chers fréres et soeurs source de joie, bonheur et motivation.

À tout les membres de ma famille.

À tout mes amis, tout mes professeurs.

Et à vous chers lectures.

N.SAADI

Table des matières

Table des figures

Liste d'abriviations et notations

ASR :Automatic Speech Recognition.

IA :Intelligence Artificielle.

ML :Machine Learning.

DL :Deep Learning.

NLP :Natural Language Processing.

TALN :Traitement Automatique du Langage Nature.

IVR :Interactive Voice Response.

DARPA :Defense Advanced Research Projects Agency.

SGMM :Subspace Gaussian Mixture Modèls.

RAP :Reconnaissance Automatique de La Parole.

MFCC :Mel-Scale Frequency Ceptral Coefficients.

FFT :Fast Fourier Transform.

CNN : Convolutional Neural Network.

RNN :Recurrent Neural Network.

HMM :Hidden Markov Model.

GMM :Modèle Mélange Gaussien.

DBN :Deep Belief Network.

WER :Word Error Rate.

UNESCO :Organisation des Nations Unies pour l'éducation la Science et la Culture.

AM :Modèles Acoustiques.

LDC : Linguistic Data Consortium

WFST :weighted finite-state transducers

Introduction générale

La reconnaissance automatique de la parole est l'une des technologies omniprésentes dans notre vie quotidienne, elle offre des fonctionnalités telles que la dictée de texte, les assistants vocaux et les systèmes de transcription automatisés.
Depuis son apparition en 1950, la reconnnaissance automatique de la parole a été constament améliorée par des linguistiques, phonéticiens, mathématiciens et ingénieurs en définissant des connaissances acoustiques et linguistiques nécessaires pour bien comprendre la parole humaine, en citant certain travaux réalisés soit sur la langue arabe par **BENAMMAR Ryadh**[8] dont le but était de construit un système de reconnaissance d'un vocabulaire d'une calculatrice vocale en arabe en utilisant l'outil HTK . De même, des travaux ont été menés sur la langue Amazighe en Latin par **Mustapha Kamal BENRAMDANE**[9], l'objectif principal de ce travail est de développer des modèles de reconnaissancevocale de la langue kabyle basés sur des approches d'apprentissage profond . Enfin, des travaux ont été menés en tifinagh par **Meryam Telmem, Youssef Ghanou**, [**?**] dans lesquels ils ont présenté une approche basée sur les réseaux de neurones convolutifs pour construire un système de reconnaissance automatique de la parole pour la langue amazighe.
Cependant, les perfermances atteintes ne sont pas parfaites et dépendent de plusieurs critères. Comme il est cité dans la thése de **Sethserey Sam**, Malgré les recherches et les travaux avancés dans ce domaine, certaines langues et dialectes demeurent sous-représentés, ce qui provoque des problèmes d'inclusivité et de l'accessibilité linguistique.
La langue Berbère ou bien Tamazight est considérée comme une langue autochone de l'Afrique du Nord, elle est connue par sa richesse et sa diversité linguistiques et culturelle. En revanche, à cause de cette diversité orale et du manque de ressources écrites,ces systèmes sont encore peu développés et la reconnaissance vocale de la langue Amazighe représente un défi.

Sur cette base, notre travail consiste à réaliser un système de reconnaissance vocale pour la langue Amazigh(Kabyle) à l'aide de la plateforme Kaldi, afin de faciliter l'utilisation de cette langue dans divers domaines tels que l'éducation et les technologies de communication.

Ce mémoire est constitué de quatre chapitres.
Le premier chapitre présente des notions fandametaux et préliminaires sur L'intelligence artificielle, machine learning, deep learning et ASR. Tout d'abord on définit IA, ML et DL, on parle sur leurs différents domaines d'application et leurs utilités. Puis, on s'intéresse à l'architecture des système s automatiques de la parole avec la paramétrisation des signaux et les modèles associés, en détaillant plus sur les modèles et les techniques utilisés dans ce travail.

Le deuxième chapitre se base sur la langue Amazighe. On commence par définir certaines notions linguistiques. Ensuite, on s'intéresse la langue Amazighe, son histoire et ses dialectes

notamment le Kabyle. Après, on explique le système phonologique de la langue Kabyle en citant ses cas exceptionnels. On passe enfin à la reconnaissance automatique de la langue Amazighe, ses défis et les travaux déjà faits dans ce domaine.

Le troisième chapitre parle sur le processus de récolte et de traitement des données. On commence par expliquer le choix de l'échantillon et ses paramètres.Puis, on détaille ce processus de cette collecte où on présente les appareils et les outils utilisés durant l'enregistrement, ainsi que les logiciels et les applications utilisés pour le prétraitement et le traitement des vocaux.

Le quatrième chapitre présente les différentes plateformes utilisées pour créer des systèmes de reconnaissance automatique de la parole comme sphinx et HTK, en se basant sur la plateforme kaldi. Puis on passe à l'implémetation de notre base de données sur cette platforme en donnant les commandes utilisées et en expliquant les étapes suivies.

Ce mémoire se termine par une conclusion qui synthétise les résultats obtenus, évalue les défis rencontrés et propose des perspectives pour les travaux futurs.

1

Concepts sur la Reconnaissance automatique de la parole

Introduction

La reconnaissance vocale est un domaine qui passionne au même temps le public et les chercheurs.
Le système de reconnaissance automatique de la parole (ASR) est l'une des technologies les plus importantes utilisées pour l'interaction homme-machine.
Le ASR a rendu possible pour les machines de comprendre les langues humaines.

Sommaire

1.1 Intelligence artificielle

1.1.1 Définition

L'intelligence artificielle (IA) est un processus d'imitation de l'intelligence humaine qui repose sur la création et l'application d'algorithmes exécutés dans un environnement informatique dynamique. Son but est de permettre à des ordinateurs de penser et prendre des décision comme des êtres humains [19].

1.2 Machine learning

Machine learning ou apprentissage automatique, est une branche de l'intelligence artificielle qui permet aux machines d'apprendre et de s'améliorer de manière autonome à partir de données.

1.2.1 Type de machine learning

L'apprentissage automatique classique est souvent catégorisé par la façon dont un algorithme apprend à devenir plus précis dans ses prédictions. Il existe quatre types de base d'apprentissage automatique :

Apprentissage supervisé

Dans l'apprentissage supervisé, les données fournies aux algorithmes d'apprentissage sont étiquetées et définissent les variables qu'ils souhaitent que l'algorithme évalue pour établir des corrélations.
L'entrée et la sortie de l'algorithme sont spécifiées dans l'apprentissage supervisé. Initialement, la plupart des algorithmes d'apprentissage automatique fonctionnaient avec l'apprentissage supervisé, mais les approches non supervisées sont de plus en plus populaires. Ces algorithmes sont utilisés pour plusieurs tâches :

- **Classification binaire :** Divise les données en deux catégories
- **Classification multiclasse :** Permet de choisir entre plusieurs types de réponses.
- **Assemblage :** Combine les prédictions de plusieurs modèles ML pour produire une prédiction plus précise.
- **Modélisation de Régression :** Prédit des valeurs continues.

Apprentissage non supervisé

Dans ce cas, les données n'ont pas d'étiquettes. La machine se contente d'explorer les données à la recherche d'éventuelles patterns.
Elle ingère de vastes quantités de données, et utilise des algorithmes pour en extraire des caractéristiques pertinentes requises pour étiqueter, trier et classifier les données en temps réel sans intervention humaine.
Les algorithmes de l'apprentissage non supervisé conviennent aux tâches suivantes :

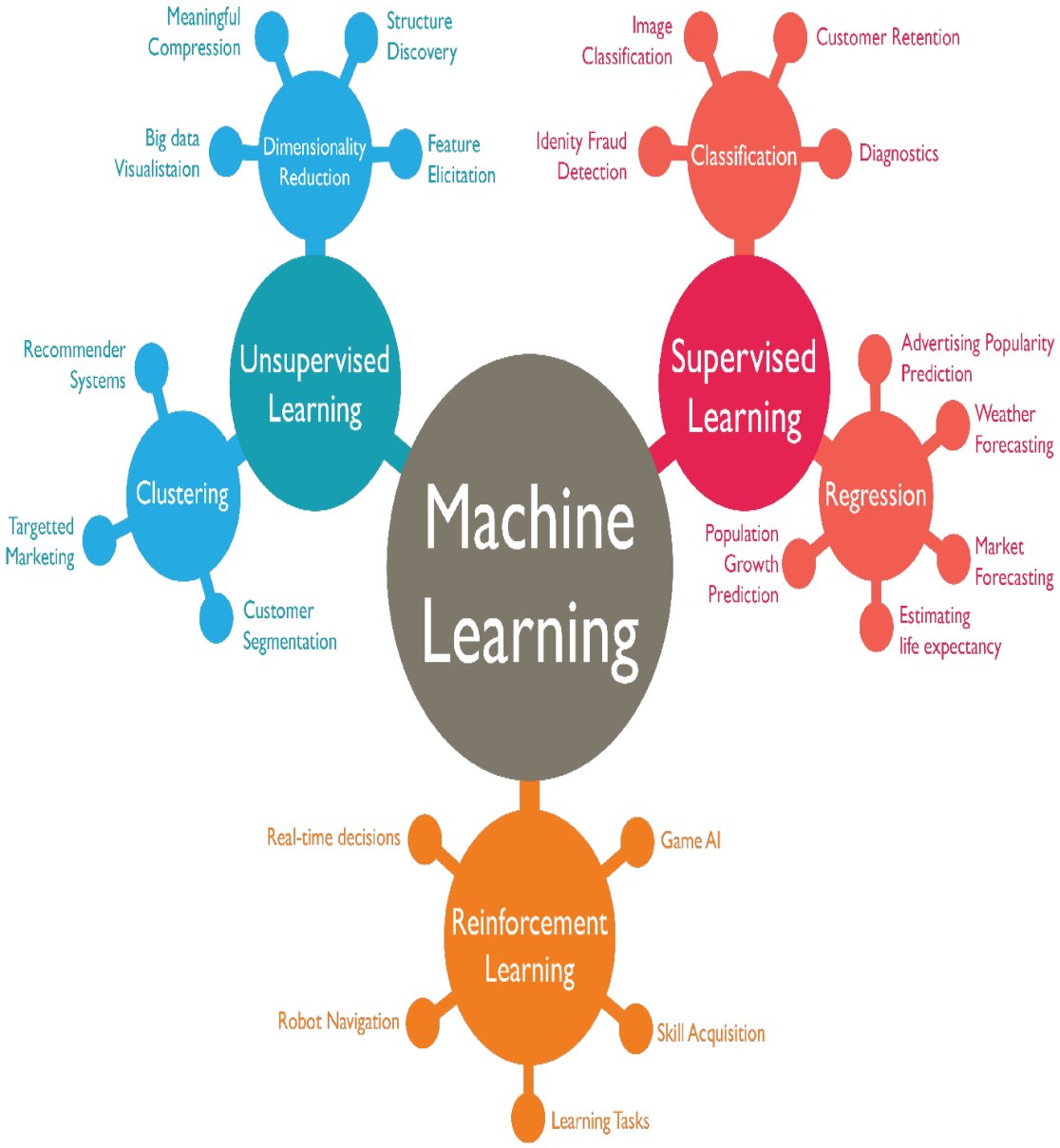

FIGURE 1.1 – Types de ML et ses domaines d'utilisations

— **Clustering(Regroupement) :** Séparer l'ensemble de données en groupes sur la base de la similarité à l'aide d'algorithmes de clustering.
— **Détection d'anomalie :** Identifier des points de données inhabituels dans un ensemble de données à l'aide d'algorithmes de détection d'anomalies.
— **Règles d'associations :** Découvrir des ensembles d'éléments dans un ensemble de données qui se produisent fréquemment ensemble à l'aide de l'exploration des règles d'association.
— **Réduction de la dimensionnalité :** Diminution du nombre de variables dans un ensemble de données à l'aide de techniques de réduction de la dimensionnalité.

Apprentissage semi-supervisé

Il fonctionne en alimentant une petite quantité de données d'apprentissage étiquetées à un algorithme. À partir de ces données l'algorithme apprend les dimensions de l'ensemble de données qu'il peut ensuite appliquer à de nouvelles données non étiquetées.
Les performances des algorithmes s'améliorent généralement lorsqu'ils s'entraînent sur des ensembles de données étiquetés. Mais l'étiquetage des données peut être long et coûteux.
Ce type d'apprentissage automatique établit un équilibre entre les performances supérieures de l'apprentissage supervisé et l'efficacité de l'apprentissage non surveillé.
Il est utilisé dans pour plusieurs domaines :

— **Étiquetages des données :** Les algorithmes entraînés sur de petits ensembles de données apprennent à appliquer automatiquement des étiquettes de données à des ensembles plus grands.
— **Détection des fraudes :** Identifie les cas de fraude lorsqu'il n'y a que quelques exemples positifs.
— **Traduction automatique :** Apprend aux algorithmes à traduire un langage basé sur un dictionnaire de mots inférieur à un dictionnaire complet.

Apprentissage par renforcement

Fonctionne en programmant un algorithme avec un objectif distinct et un ensemble de règles prescrits pour atteindre cet objectif. Un spécialiste des données programmera également l'algorithme pour rechercher des récompenses positives pour avoir réalisé une action qui est bénéfique pour atteindre son objectif ultime et pour éviter les sanctions pour avoir effectué une action qui l'éloigne de son objectif. Il est souvent utilisé dans les domaines suivants :

— **Robotique :** Il est utilisé pour entraîner des robots à effectuer des taches complexes,comme la manipulation des objets ou la navigation dans des environnement inconnu.
— **Gameplay video :** Des algorithmes ont été utilisés pour créer des agents capables d'apprendre à jouer à des jeux notamment des jeux de société comme des jeux vidéo.
— **Gestion des ressources :** Il est appliqué dans la gestion,où les agents peuvent prendre des décisions d'investissement pour maximiser le rendement.

1.3 Deep learning

1.3.1 Définition et fonctionnement

Le Deep Learning est une sous-branche du Machine Learning – qui lui même est une branche de l'intelligence artificielle (AI) – qui utilise des réseaux de neurones artificiels afin d'apprendre à partir de données. Les réseaux de neurones artificiels constituent l'un des algorithmes d'AI les plus sophistiques. Ils sont inspirées du fonctionnement du cerveau humain et sont capables d'apprendre à des modèles complexes à partir de grandes quantités d'informations.
Le système apprendra par exemple à reconnaître les lettres avant de s'attaquer aux mots dans un texte, ou détermine s'il y a un visage sur une photo avant de découvrir de quelle personne il s'agit[**?**].

1.3.2 Applications du DL

Deep learning s'avère utile dans divers secteurs :

— **Traitement d'image :** Deep learning est très efficace pour les analyses d'images. Il est par exemple employé dans l'imagerie médicale pour détecter des maladies ou dans le secteur automobile pour les voitures autonomes. Mais aussi pour le reconnaissance faciale, comme sur les smartphones ou sur Facebook.
— **Création des textes :** Le deep learning est également un atout dans la création de contenu. En effet, désormais, un ordinateur peut être capable de rédiger de manière autonome des textes ou d'effectuer des traductions. La seule condition est l'accès à une quantité de données suffisante de formation. Cela fait partie du NLP (Natural Language Processing), une branche de l'IA, qui traite automatiquement le langage humain.
— **Publicité et marketing :** Il aide à personnaliser les recommandations de produits et de contenu en fonction des préférences de chaque utilisateur. Les entreprises peuvent ainsi cibler leurs publicités de manière plus précise et offrir des expériences plus pertinentes à leurs clients.
— **Automobile et transport :** Le secteur de l'automobile et du transport bénéficie également du deep learning. Les véhicules autonomes utilisent des réseaux neuronaux pour détecter et réagir aux obstacles. Ils identifient et s'adaptent aux panneaux de signalisation et aux piétons. Cette technologie joue un rôle crucial dans la sécurité et la navigation dans les véhicules du futur.

1.3.3 Distinction entre IA,ML et DL

Certains problèmes augmentent de façon exponentielle où le temps de traitement dépend de la taille des exemples fournis. Il est possible que le problème ne puisse pas être résolu à temps à l'échelle humaine, et c'est pourquoi des techniques de traitement ont vu le jour pour vaincre cette difficulté. Comme être capable de décomposer un problème en sous-problèmes quantifiables ayant une solution.
Il y a une confusion fréquente dans le débat public entre intelligence artificielle , apprentissage automatique et apprentissage profond. Pourtant, ces notions ne sont pas équivalentes, mais sont imbriquées :
— l'intelligence artificielle englobe le machine learning, qui lui-même englobe le deep learning.

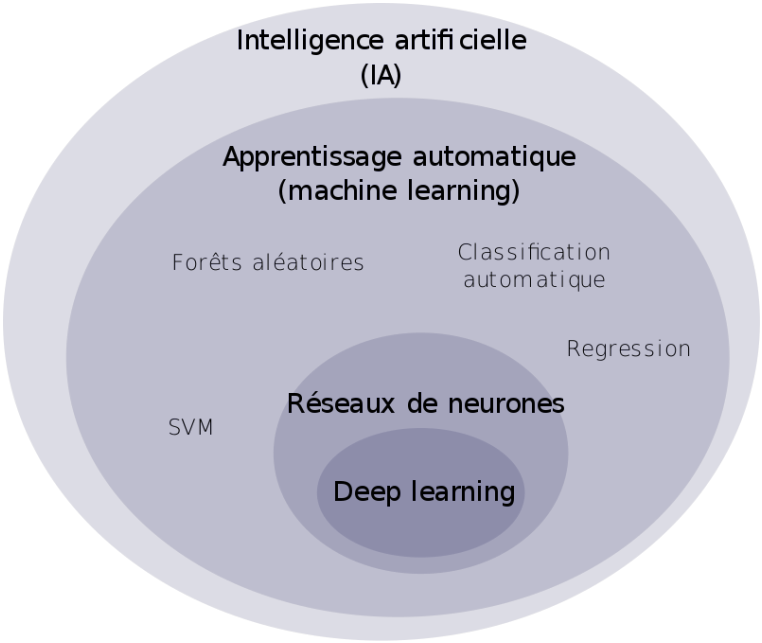

FIGURE 1.2 – Distinction entre IA,ML et DL

1.4 Traitement automatique du langage naturel

1.4.1 C'est quoi le traitement automatique du langage naturel (TALN)

C'est une branche de l'intelligence artificielle (AI) qui permet aux ordinateurs de comprendre, générer et manipuler le langage humain.
Le traitement du langage naturel combine la linguistique informatique (modélisation du langage humain basée sur des règles) avec des modèles statistiques, d'apprentissage automatique et d'apprentissage en profondeur. Ensemble, ces technologies permettent aux ordinateurs de traiter le langage humain sous la forme d'un texte ou sous forme de données vocales et d'en «comprendre »le sens complet, avec l'intention et le ressenti du locuteur ou de l'auteur.

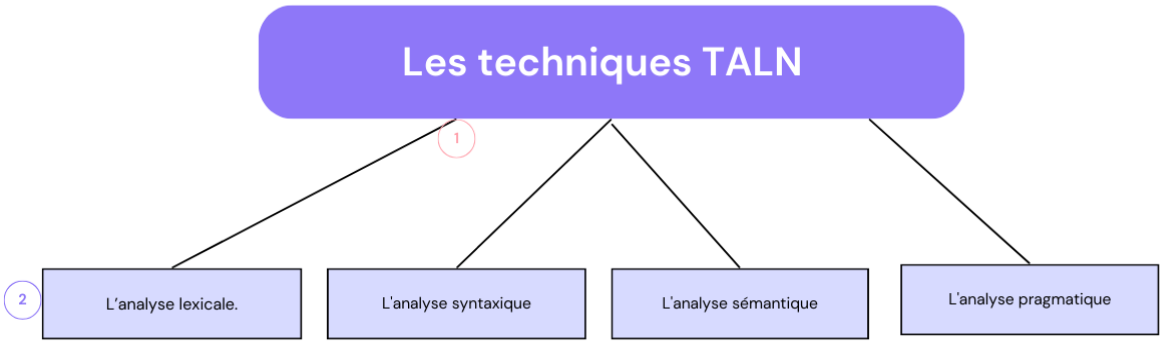

FIGURE 1.3 – Les techniques de traitement automatique du langage naturel utiliser

1.4.2 Les cas d'utilisation du TALN

Reconnaissance vocale

La reconnaissance vocale est un excellent exemple de la façon dont la TALN (PNL) peut être utilisée pour améliorer l'expérience client. C'est une exigence très courante pour les entreprises d'avoir des systèmes IVR en place afin que les clients puissent interagir avec leurs produits et services sans avoir à parler à une personne en direct. Cela leur permet de gérer plus d'appels, mais contribue également à réduire les coûts.

La traduction automatique

C'est l'une des premières applications du NLP. Ces services instantanés traduisent les contenus écrits dans diverses langues. Le plus souvent, les textes sont différents en fonction du traducteur utilisé.
Il existe quatre types sont La traduction automatique statistique , La traduction automatique basée sur les règles, La traduction automatique hybride, la traduction automatique neuronale. Ces logiciels sont Google Translate, DeepL, Microsoft Translator.

Les chatbots

Aussi appelés agents conversationnels ou assistants virtuels, tels que ceux d'Apple (Siri) et d'Amazon (Alexa) de Google (Google Assistant) et de Microsoft (Cortana), Ces programmes informatiques simulent une conversation humaine. Ils permettent d'interagir pour obtenir les réponses aux questions posées par un utilisateur. répondre correctement aux questions simples des individus.

L'analyse des sentiments(Opinion Mining)

Identifier l'opinion et le ressenti d'un individu. Autrement dit qui vient identifier les informations subjectives d'un texte pour extraire l'opinion de l'auteur. Cela permet par exemple de mesurer le niveau de satisfaction des clients vis-à-vis des produits ou services fournis par une entreprise ou un organisme.

Synthèse de la parole

Créer de la parole artificielle à partir d'un texte quelconque. Ces systèmes ont largement franchi le seuil de l'intelligibilité permettant leur utilisation.

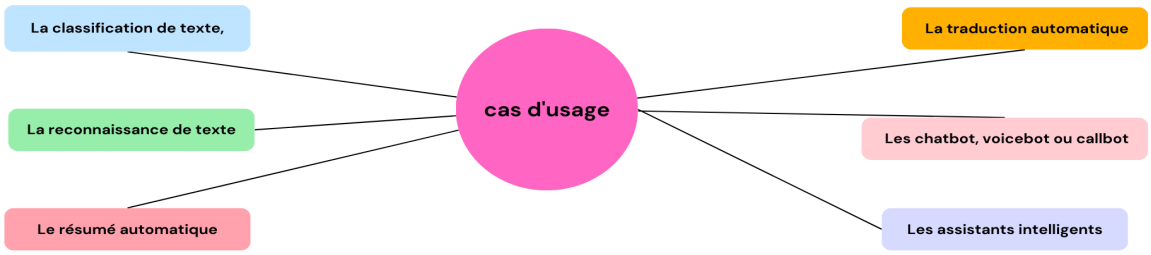

FIGURE 1.4 – autres cas d'usage plus ou moins élaboré

1.5 Reconnaissance automatique de la parole

La reconnaissance automatique de la parole(Automatic Speech Recognition) étant une branche de l'intelligence artificielle, vise principalement à convertir automatiquement le signal de parole en une séquence de mots via un algorithme implémenté sous forme de module logiciel ou matériel.
Autrement dit La reconnaissance automatique est la conversion de la voix (parole) en fichier numérique. Elle permet de décoder un signal acoustique de la parole en une suite de mots effectivement prononcés [27].

1.5.1 Historique

La premiére apparition de l'ASR était en 1950, lorsque le système a pu reconnaître 9 chiffres prononcés.

En 1970, un énorme investissement a été fait par le département de défense américain avec Defense Advanced Research Projects Agency (DARPA) dans le but de développer de la technologie militaire ,un système est developpé nommé HARPY est capable de reconnaître 1000 mots. Au meme temps, un système capable de reconnaître plusieurs locuteurs a été developpé par les laboratoires BELL.

En 1980, c'est l'apparition des réseaux de neurones et les modèles de Markov cachés qui ont permis à la reconnaissance vocale de faire un nouveau bond en avant de pouvoir reconnaitre plusieurs milliers de mots.

En 1990, l'augmentation du nombre de foyers et d'enreprises à faire l'acquisition d'un ordinateur va donner un élan aux logiciels de reconnaissance vocale comme le logiciel Dragon pour la dictée.

En 2000, c'est la commande de recherche vocale développée par Google est apparue et les corpus d'apprentissage augmentent au fur et à mesure que les requêtes des utilisateurs sont collectées.

Actuellement, Les systèmes de reconnaissance vocale jouent un rôle important, que ce soit dans la vie professionnelle ou quotidienne avec des applications comme Alexa, Google Home et Siri. La recherche des systèmes plus efficaces amène les chercheurs à utiliser des techniques différentes telles que les Subspace Gaussian Mixture Models (SGMM) et les réseaux de neurones. La recherche et le développement de ces systèmes continue toujours notamment pour les langues moins parlées qui ont le problème de manque de ressources [20].

1.5.2 Type de reconnaissance vocale

•Reconnaissance des mots isolés

Les reconnaissances de mots isolés acceptent un seul mot à la fois. Ces systèmes ont des états de ”Écoute/Non-écoute”, où ils demandent à l'orateur d'attendre entre les énoncés. ”Énoncé isolé” pourrait être un meilleur terme pour cette classe.[16]

• Reconnaissance des mots connectés

Les systèmes de mots connectés (ou plus exactement les ”énoncés connectés”) sont similaires aux mots isolés, mais permettent aux énoncés séparés d'être ”combinés” avec une pause minimale entre eux.

• Reconnaissance de la parole continue

traite de la parole où les mots sont connectés plutôt que séparés par des pauses. En conséquence, les informations de frontière inconnue sur les mots, la coarticulation, la production des phonèmes environnants et le débit de la parole affectent les performances des systèmes de reconnaissance de la parole continue. Les reconnaissances avec des capacités de parole continue sont parmi les plus difficiles à créer car elles utilisent des méthodes spéciales pour déterminer les frontières des énoncés.

• Reconnaissance de la parole spontanée

La parole spontanée a un grand nombre de définitions. À la base, cela peut être considéré comme un discours qui sonne naturellement et qui n'est pas répété. Le locuteur commence un énoncé et atteint un point où il ne peut pas trouver le mot juste ou pense mieux à un mot, et a besoin de temps pour trouver une alternative appropriée, ils répètent un mot comme une sorte de ”course” à la deuxième tentative. Un système ASR avec une capacité de parole spontanée devrait être capable de gérer une variété de caractéristiques de parole grammaticale non standard telles que des mots courir ensemble, ”pad’” et ”«al”, et même de légers bégaiements. [21]

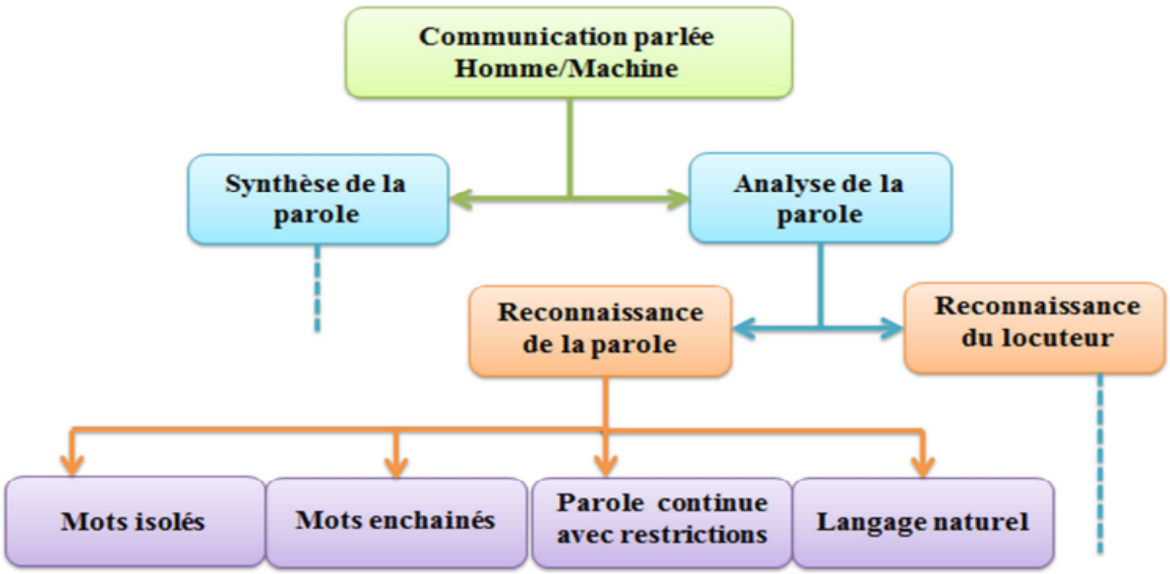

FIGURE 1.5 – Classification d'un ASR [31].

1.5.3 Problèmes liés à la reconnaissance vocale

En raison de la complexité du langage naturel et des défis qu'il pose à la compréhension et à la traduction par les machines, concevoir un système de reconnaissance vocale peut présenter divers types de difficultés et d'obstacles.

— Le bruit de fond et les échos Informations non désirées dans le signal vocal. En RAP, nous devons identifier et filtrer ces bruits du signal vocal.

Ca peut etre le son d'horloge, des voitures , le vent, un autre locuteur humain en arrière-plan, etc.

Ya aussi les echos qui sont est le signal vocal rebondi sur un objet environnant et qui arrive dans le microphone quelques millisecondes plus tard.

— Les dialectes les variantes dialectales font que les mots appartenant à une même langue soient prononcés différemment d'un locuteur à un autre.

— Langage corporel Un humain ne communique pas seulement avec la parole, mais aussi avec des parties de son corps Notamment gestes de la main, mouvements des yeux, postures, etc.

—Homophone Le problème des homophones est l'une des difficultés majeures de la reconnaissance de la parole. Ce sont les mots prononcés de la même manière, mais ont des orthographes et des significations différentes.

—La variabilité des locuteurs Cela peut se révéler de deux façons, la premiere cest quand la variabilité interlocuteur où différents locuteurs prononcent un même mot avec des voix différentes ou quand la variabilité intra-locuteur où un même locuteur prononce un mot avec des variations émotionnelles, entraînant des changements d'intonation ou de vitesse de prononciation.

Cette variabilité rend très difficile la définition d'invariants et complique la tâche dereconnaissance. Ainsi, il faut pouvoir séparer ce qui caractérise les phonèmes, de l'aspect

particulier à chaque locuteur [3].

1.5.4 Travaux déjà réalisés

De nombreuses recherches ont été menées récemment dans le domaine du traitement du signal vocal. En particulier, la technologie de reconnaissance automatique de la parole (ASR).
La reconnaissance vocale de la langue amazigh a été déjà abordée en **2023** par **Barkani, F., Hamidi, M., Laaidi, N.** nomé **Reconnaissance vocale amazighe basée sur la boîte à outils Kaldi ASR.**
Dans ce travail, ils ont proposé une nouvelle approche pour intégrer la langue amazighe, dans un système de reconnaissance vocale isolé en exploitant la plateforme open source Kaldi, leur système est capable de reconnaître les dix premiers chiffres amazighs et les dix mots isolés amazighs indispensables au quotidien [13].
Reconnaissance vocale basée sur le deep learning appliquée à la langue kabyle, réalisé par **Mustapha Kamal BENRAMDANE** en 2020, le but de ce travail est de développer des modèles de reconnaissance vocale de la langue kabyle basés sur des approches d'apprentissage profond [9].
Automatic Speech Recognition est un travail fait par **[Li Deng] et [Dong Yu]** en **2013** qui a été publié sur "IEEE/ACM Transactions on Audio, Speech, and Language Processing".
Cet article Propose une évaluation approfondie des technologies actuelles de reconnaissance vocale.
En se concentrant sur les techniques, les modèles et les applications de la reconnaissance vocale en anglais. Il couvre également les défis et les perspectives futures de ce domaine en constante évolution.
En outre, bien que cet article principalement axé sur la reconnaissance vocale en anglais, mais il offre connaissances transférables qui pourraient être utiles pour d'autres langues, telles que le Uyghur, Kazakhs et les Kirghizes, soulignant ainsi sa valeur dans un contexte plus large de recherche en reconnaissance vocale.
Reconnaissance automatique de la parole pour des langues peu dotées est un travail sur la langue khmer fait par **Viet Bac Le**. Ce manuscrit présente leur méthodologie qui vise à développer et adapter rapidement un système de reconnaissance automatique de la parole continue pour une nouvelle langue peu dotée [22].
Automatic speech recognition system with pitch dependent features for Punjabi language on KALDI toolkit réaliser par par **Jyoti Guglani, A.N. Mishra**,Dans cet article, l'efficacité du système ASR de la langue pendjabi a été rapportée sur la boîte à outils Kaldi en termes de WER[15].
Concernant la langue arabe elle y a eu déjà mal de travaux réalisés en comparent au kabyle , **Souadkia Abdelhak** dans son trvaille nommé **Reconnaissance automatique de la parole arabe : Approche évolutionniste**. Ce travail vise à réaliser un système de reconnaissance automatique de la parole arabe par un algorithme génétique [4].

1.6 Système de la reconnaissance vocale

Le Système automatique de la reconnaissance de la parole se compose généralement de deux composantes : La paramétrisation du signal et les modèles mis en œuvre dans ce système.

1.6.1 Analyse de signal

L'enregitrement vocal se fait généralement avec des appareils comme le dictaphone, le microphone, pendant cet enregistrement plusieurs paramètres peuvent influencer sur la capture de son telle que le type de l'apareil utilisé, la distance entre la source et l'apareil et le bruit environnant.

Analyse MFCC

MFCC (Mel-Scale Frequency Ceptral Coefficients) captent les caractéristiques essentielles d'un signal vocal, cette analyse implique plusieurs traitements sur le signal audio, comme illustre dans la figure 1.6 .
Le signal acoustique continu est segmenté en trames de N échantillons, avec un pas d'avancement de M trames ($M < N$). Les valeurs couramment utilisées pour M et N sont respectivement 10 et 20. Le calcul de MFCC implique plusieurs étapes présentées ci-dessous :

— **Préaccentuation :** Pour la préaccentuation du signal on applique l'équation de différence de premier ordre aux échantillons $x(n)$ [8].

$$x^(n) = x(n) - kx(n-1) \qquad \text{avec} \qquad 0 < n < N-1$$
$$\text{et } k \text{ coefficient de préaccentuation.} \qquad 0 < k < 1$$

— **Fenêtrage :** On considère $w(n)$ comme une fenêtre où $0 < n < N-1$ et N est le nombre d'échantillons dans chacune des trames, donc on aura le signal x_a comme résultat du fenêtrage [8].

$$x_a = x(n)w(n) \qquad \text{avec} \qquad 0 < n < N-1$$

Plusieurs types de fenêtrages peuvent être utilisés comme le fenêtrage triangulaire, Hann et Hamming. Nous allons définir le plus utilisé c'est celui de Hamming.

$$w(n) = \begin{cases} 0.54 - 0.46\cos(\frac{2\pi.n}{N-1}) & 0 \le n \le N-1 \\ 0 \qquad\qquad sinon. \end{cases}$$

— **Tranforme de Fourrier rapide :** Connu sous le nom Fast Fourier Transform (FFT), permet de passer de domaine temporelle au domaine fréquentiel, les valeurs obtenues sont appelées les spectres [8].

$$x[k] = \sum_0^{N-1} x_a[n](\exp \frac{-2j\pi}{N} kn) \qquad 0 \le n \le N-1$$

— **Filtrage sur l'échelle de Mel :** Dans l'échelle de mesure Mel, la correspondance est approximativement linéaire sur les fréquences au-dessous de 1kHz et logarithmique sur les fréquences supérieures à celle-ci. Cette relation est donnée par la formule suivante :

$$m = 2595 \log_{10}(1 + \frac{f}{700})$$

Le logarithme de l'énergie de chaque filtre est calculé selon l'équation :

$$S[m] = \ln[\sum k = 0^{N-1} X_a[k] H_m[k]] \qquad 0 < m \leq M$$

— **Calcul de Cepstre sur l'échelle de Mel :** Pour obtenir Le cepstre sur l'échelle de fréquence Mel, on calule la transformée cosinus discrète du logarithme de la sortie des M filtres [8].

$$c[n] = S[n] \cos \pi n(m - 1/2)/M \qquad \leq n < M$$

Le premier coefficient, $c[0]$, représente l'énergie moyenne dans la trame de la parole.
$c[1]$ reflète la balance d'énergie entre les basses et hautes fréquences.
Pour $i > 1$, $c[i]$ représente des détails spectraux de plus en plus fins.

— **Calcul de coefficient MFCC :** À travers les dérivées des coefficient Δ_c et les dérivées secondes $\Delta\Delta_c$ on peut calculer les changements temporels dans le spectre (c) [8].

$$\begin{pmatrix} c_k \\ \Delta c_k \\ \Delta\Delta c_k \end{pmatrix}$$

Dans le traitement vocal, Les MFCC sont les coefficients qui constituent collectivement le Cepstrum de fréquence de Mel (MFC), une représentation du spectre de puissance à court terme du son est basée sur la transformation linéaire en cosinus d'un spectre de puissance logarithmique sur une échelle de fréquence logarithmique Mel non linéaire.
Les MFCC sont linéairement espacés sur l'échelle de fréquence Mel qui se rapproche beaucoup de la réponse du système auditif humain.
Cette représentation du signal sonore extrait des caractéristiques discriminantes qui permettent d'atteindre une classification du son avec une bonne précision.

1.6.2 Modèle de langage

Un modèle de langage a donc pour but d'estimer la probabilité a priori de toutes les séquences de mots qu'il est possible de construire à partir du lexique. Pour ce faire, il peut s'appuyer sur différentes sources d'informations, comme par exemple des règles syntaxiques ou sémantiques, ou encore des statistiques issues de gros volumes de données. Nous nous concentrerons ici sur les modèles de langages les plus utilisés dans la reconnaissance vocale [24].

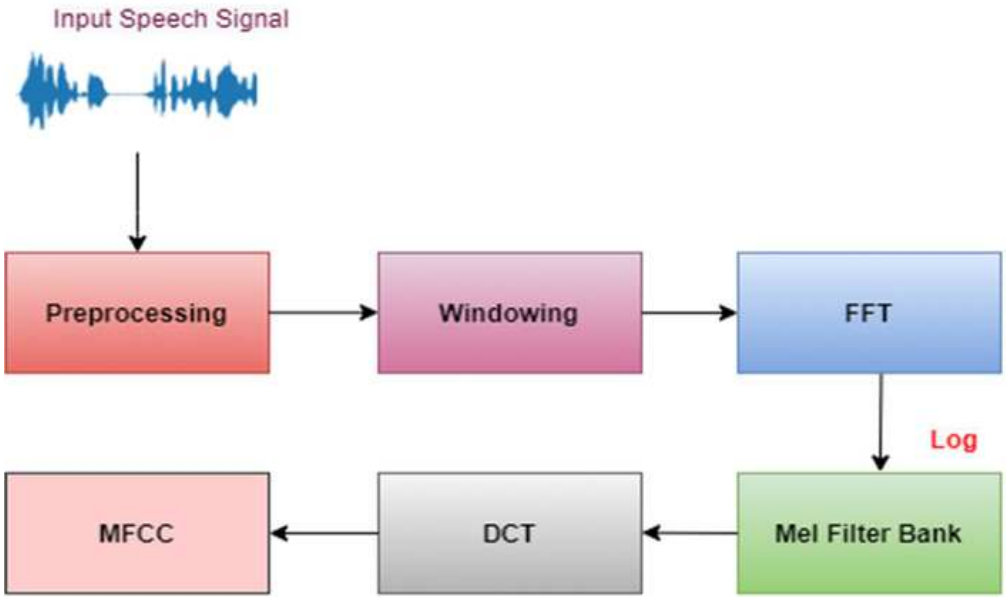

FIGURE 1.6 – Analyse MFCC

Modèle de langage statistique n-gramme

Ce modèle a été proposé par **Jelinek (1976)**. L'idée à la base des modèles de langage n-gramme est que la probabilité d'apparition d'un mot peut être estimée à partir des $(n-1)$ mots précédents.

$$P(W) \approx P(W_i|h_i^n)$$

avec h_i^n l'historique de taille $(n-1)$ du mot W_i

$$P(W) \approx \prod_{i-1}^{M}(W_i|h_i^n)$$

.

Modèle des Réseaux de neurones convolutifs

Connu sous le nom Convolutional Neural Network, représentent un grand avancement dans la reconnaissance d'images. Ils sont le plus souvent utilisés pour analyser les images visuelles, ils sont aussi utilisés dans la reconnaissance vocale. Les CNN ont été créés à partir d'un réseau de neurones formé de plusieurs couches. Chaque couche est formée de neurones qui sont connectés aux neurones de la couche suivante. En général, on les divise en trois couches :

— **Couche de convolution :** Elle est constitué des filtres, chaque filtre représente tableaux de valeurs, donc la couche convolutif prend en entrée une image et produit des features maps en sortie.

— **Couche de pooling :** Agit comme une couche de réduction, elle reçoit des feature maps, divise l'images en blocs, réduit la dimension en gardant les caractéristiques importantes.

— **Couche fully-connected :** C'est la dernière couche des réseaux convolutifs, elle

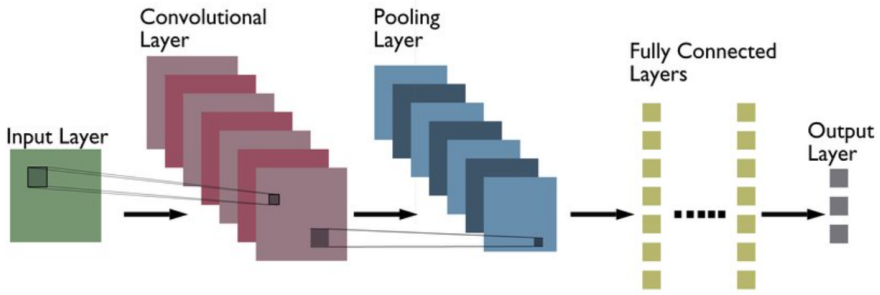

FIGURE 1.7 – Architecture de CNN

combine les features précédentes pour classifier l'image.

Modèle des Réseaux de neurones récurrents

Les RNN ont été proposés dans les années 80 (Rumelhart, et al., 1986 ; Elman, 1990 ; Werbos,1988) pour modéliser les séries temporelles. Cette structure est similaire aux réseaux de neurones traditionnels, mais ces réseaux comportent des cycles dans leur graphe (Pascanu, et al., 2013). Ces cycles permettant au réseau d'entretenir une information en l'envoyant à lui- même. Ces réseaux sont largement utilisés pour la traduction automatique et la reconnaissance de la parole pour extraire des caractéristiques acoustiques .
Les RNN se compose aussi de trois couches :

— **Couche d'entrée :** Elle reçoit l'entrée de réseau neuronal le traite et le transmet vers le couche prochaine.
— **Couche intermédiaire :** Standardise les différentes fonctions d'activation, les poids et les biais, de sorte que chaque couche cachée possède les mêmes paramètres.
— **Couche de sortie :** Elle est responsable de la prédiction et la classification finale.

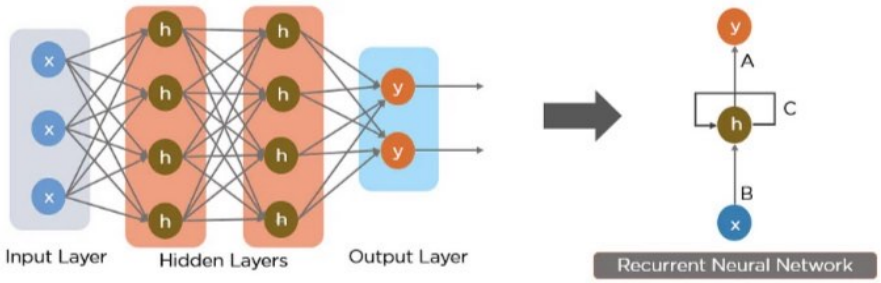

FIGURE 1.8 – Architecture de RNN

1.6.3 Modèle de prononciation

Dans un système de reconnaissance automatique de la parole, on a besoin d'une ressource lexicale qui fait le lien entre les unités acoustiques et les mots, il s'agit du lexique de prononciation. Deux lexiques sont nécessaires : un lexique utilisé lors de la

phase d'apprentissage du modèle acoustique, il contient les prononciations de tous les mots qui se trouvent dans la transcription textuelle des données acoustiques du corpus d'apprentissage. Un lexique de reconnaissance qui contient tous les mots qui pourront être reconnus par notre système.
Ces mots doivent être classés en fonction d'un poids qui définit leur importance pour le domaine d'application. Ce poids $C(w_i)$ est calculé dans le cas de k documents selon l'équation suivante

$$C(w_i) = \sum_k \lambda_k C_k(w_i)$$

avec $C_k(w_i)$ est la fréquence du mot w_i dans le document $k$et λ_k le poids de chaque documents.
Une approche simple consiste à utiliser la taille du texte comme le poids de chaque document dans l'interpolation linéaire. Une autre approche consiste à optimiser ces poids à l'aide d'un corpus de validation.
Les approches statistiques sont basées sur l'idée qu'avec suffisamment d'exemples, on peut prédire la prononciation d'un nouveau mot. Diverses techniques de modélisation ont été adaptées pour ce problème, notamment les réseaux de neurones[24].

1.6.4 Modèle acoustique

Modèles génératifs GMM-HMM

Aujourd'hui les Modèles de Markov caché sont les plus utilsés dans le système de reconnaissance vocale, ils ont été introduits par **Jelink et Baker 1975, 1976**. Chaque unité est modélisée par HMM (Hidden Markov Model), cette unité peut être des mots dans le cas des petits lexiques ou bien des phonèmes dans le cas des grands lexiques.
Dans un modèle de Markov caché, à chaque instant t, une transition de l'état actuel i à l'un de ses états connectés j est exécutée avec une probabilité a_{ij}. Chaque état i peut émettre une observation acoustique o_t avec une probabilité $b_i(o_t)$. Mathématiquement, un modèle de Markov caché se décrit par les paramètres suivants :

— Le nombre N des états du modèle.
— Les probabilités de transition d'états a_{ij}. Ces probabilités sont représentées par une matrice A de taille $N * N$
— La densité de probabilité d'observation associée à l'état i, c'est-à-dire l'estimation des probabilités $b_i(o_t) = P(o_t|S_i)$ avec $1iN$.
— La distribution de probabilité d'être à un état à l'instant initial, $\pi = (\pi_i)avec1iN$

Pour la modélisation caractéristique de Modèle de Markov on utilise le modèle Mélange Gaussien (GMM) qui est developpé en 1980, les GMM sont des distribution flexibles, sont mieux adaptés a la modélisation de la parole.
Les modèles GMM-HMM ont deux hypothèses, la première considère qu'un état est conditionnelement independant de tous les autres états,la deuxième concerne les émissions acoustiques, qui sont conditionnellement independantes des autres émissions [14].

Modèle Hybrides DNN-HMM

Ce modèle est apparue en 1980, il est appliqué avec le modèle GMM en 1999**Hermansky 1999**, en 2006 l'un des premiers travaux basés sur DNN est proposé par **Hinton 2006** sous le nom Deep Belief Network (DBN). Ce modèle est plus efficace que le modèle GMM-HMM. Parmi ses avantages est la capacité de modèliser des représentations plus complexes et non flexibles, par contre la complexité de la procédure d'entrainement reste toujours un inconvénient [14].

Modèle End-to-end

En cherchant une solution pour la complexité du modèle DNN-HMM, le modèle End-to-end est apparu, tout d'abord il permet de se débarrasser de pré-alignement les données avec un autre modèle. Ensuite avec un modèle unique il permet d'unifier le processus d'entraînement et dans le but de simplicité, il agit au niveau de caractères pas au niveau phonèmes [14].

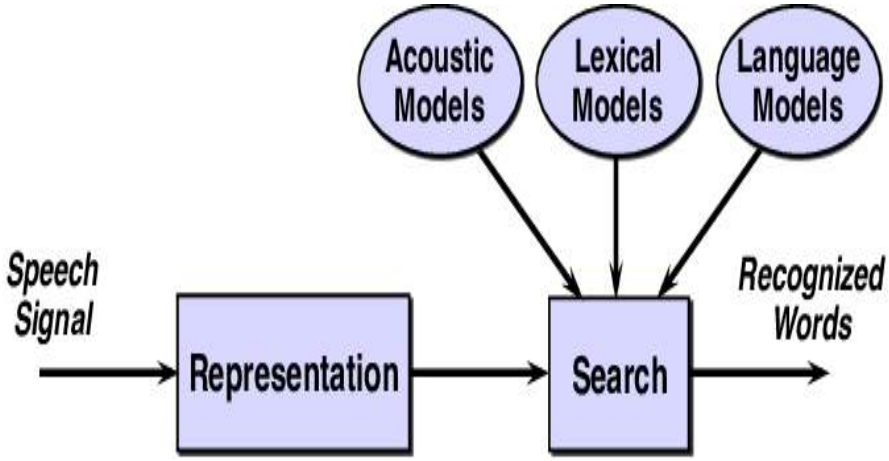

FIGURE 1.9 – Modèles de langage

1.7 Processus de reconnaissance vocale

1.7.1 Algorithme de Décodage

Le décodage consiste à rechercher le chemin optimal dans le graphe de toutes les phrases possibles. La phrase est construite comme un enchaînement de mots, eux-mêmes constitués de phonèmes chacun modélisés par un petit automate états acoustiques. Le processus de reconnaissance automatique de la parole détermine la séquence de mots $\hat{W}$ la plus vraisemblable, soit une séquence d'observations acoustiques O.

$$\hat{W} = MaxP(W|O) = Max\frac{P(O|W)P(W)}{P(O)}$$

$P(O|W)$ est la probabilité acoustique et $P(W)$ est la probabilité linguistique.

1.7.2 Évaluation de performance

L'évaluation de la reconnaissance de la parole est donnée par le critère du taux d'erreur de mots WER (Word Error Rate) qui mesure le rapport du nombre d'erreurs de reconnaissance sur le nombre total de mots prononcés.). Il y a trois types d'erreurs de reconnaissance : substitution (S), omission (O) et insertion (I0. La vérification des erreurs de reconnaissance nécessite une transcription de référence, qui est généralement créée manuellement.

$$WER = \frac{S+O+I}{N_{mots}}$$

1.7.3 Mesure de confiance

L'algorithme de décodage est un processus probabiliste qui recherche la meilleure adéquation entre un signal audio et une séquence de mots.
Cependant, cette adéquation risque de contenir des erreurs de reconnaissance même dans les meilleures conditions d'application.
Ces erreurs de reconnaissance peuvent être liées au modèle de langage (le plus souvent à cause des mots hors-vocabulaire), au modèle de prononciation (prononciation non-définie pour un mot connu) ou au modèle acoustique (bruits inconnus, locuteurs trop différents).
L'algorithme de reconnaissance affiche simplement le message décodé, mais il ne peut pas dire de lui-même (de manière certaine) si le message contient des erreurs, ni où se trouvent ces erreurs.
Les trois types de mesures de confiance les plus courantes sont celles fondées sur une combinaison des paramètres prédictifs **[Wessel et al. 1999 ; Zhang et Rudnicky 2001]**, sur le rapport de vraisemblance **[Rose et al. 1995]** ou sur la probabilité a posteriori **[Rueber 1997 ; Wessel et al. 2001]**.
Les mesures de confiances basées sur le rapport de vraisemblance utilisent deux hypothèses :
H_0 si le message et son modèle acoustique sont corrects.
H_1 si le message et son modèle acoustique ne sont pas bons.
Les mesures de confiance a posteriori sont calculées par la formule suivante :

$$P(w|O) = \frac{P(O|w)P(w)}{P(O)}$$

Pour une bonne mesure de confiance, il faut que le produit entre la probabilité du modèle acoustique $P(O|w)$ et la probabilité du modèle de langage $P(w)$ soit bien normalisé par la probabilité acoustique $P(O)$.

Conclusion

Dans ce chapitre, on a pu explorer les concepts fondamentaux de IA, ML, DL leur application dans l'ASR, on a pu présenter les différentes technologies utilisées pour entraîner les modèles de la reconnaissance automatique dans le but de comprendre et de d'interpréter le langage humain à partir des données vocales.
Malgré les techniques avancées qui permettent d'effectuer des tâches plus complexes et plus précises, l'ASR rencontre toujours des difficultés et des défis concernant la généralisation de certaines langues et dialectes comme la langue Amazighe.
Dans le prochain chapitre on va procéder à la richesse de la langue Amazighe, ses aspects linguistiques, son importance culturelle et historique ainsi que son interaction avec les nouvelles technologies.

2
La langue Amazighe

Introduction

Tamazight, la langue maternelle de millions de personnes, a été classée parmi les langues menacées de disparition par l'UNESCO, elle a longtemps été définie en tant que langue orale à usage quotidien seulement.
Le Tifinagh, l'écriture utilisée par les Amazighs pour noter leur langue, a été conservée seulement par les Touaregs du Sahara Algérien, Libyen, Malien et Nigérien.
Néanmoins, cet alphabet n'a pas contribué à la préservation de ce patrimoine historique et culturel oral si riche. Malgré son caractère d'oralité, cette langue a survécu, à l'écart du reste du monde et des autres cultures, cloîtrée dans différentes aires disséminées à travers le Maghreb et d'autres contrées, mais toujours présente – peut-être justement grâce à cette oralité.

Sommaire

2.1 Language et parole

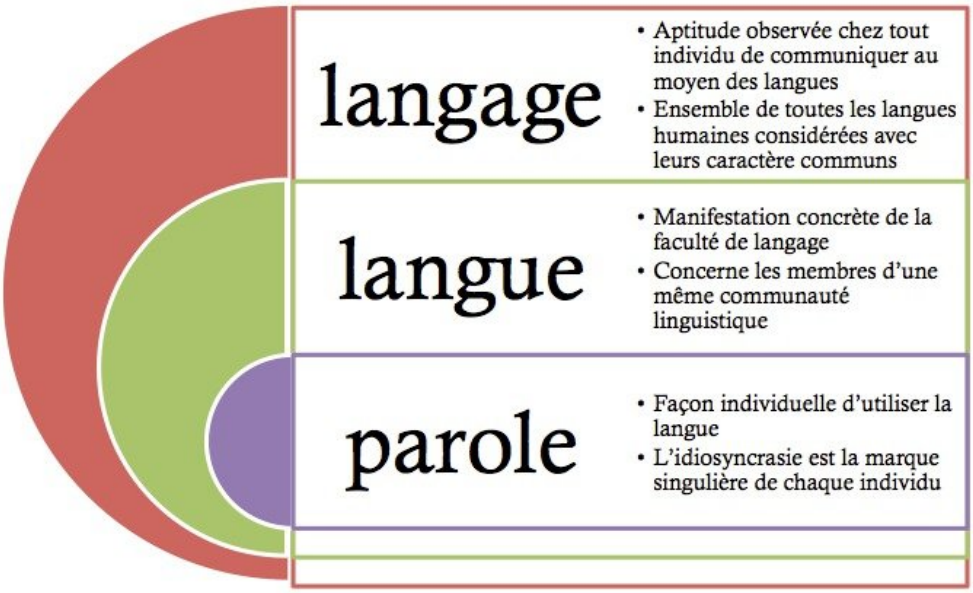

FIGURE 2.1 – Langue,Language et Parole

2.2 Phonétique et phonologie

— **Phonétique** : C'est une branche linguistique qui étudie les phonèmes.
— **Phonologie** : C'est la science qui étudie les phonèmes non en eux-mêmes mais quant à leur fonction dans la langue.
— **Phonéme** : C'est une unité minimale du langage porteuse d'une signification linguistique[12].

2.3 Tamazight Définitions et Géographie

La langue Amazighe ou bien (Tamazight en berbère), est considérée comme une langue autochtone de l'Afrique du Nord, elle couvrait à l'origine l'ensemble de l'Afrique du Nord et du Sahara. Elle est devenue minoritaire à la suite d'un lent processus d'arabisation linguistique de l'Afrique du Nord consécutif à la conquête arabe et à l'islamisation, puis à l'arrivée de populations Arabes nomades venues du Moyen-Orient.
Tamazight est dispersée en îlots d'importance très variable, de quelques milliers à plusieurs millions d'individus sur un territoire immense. Les principaux pays concernés sont le Maroc (40% de la population) et l'Algérie (25%). En dehors des Touaregs, dispersés sur cinq pays de la zone saharo-sahélienne (Niger, Mali, Algérie, Libye, Burkina-Faso), il existe des groupes berbères en Libye (10 %), en Tunisie (1%), en Égypte (Siwa) et en Mauritanie [6].

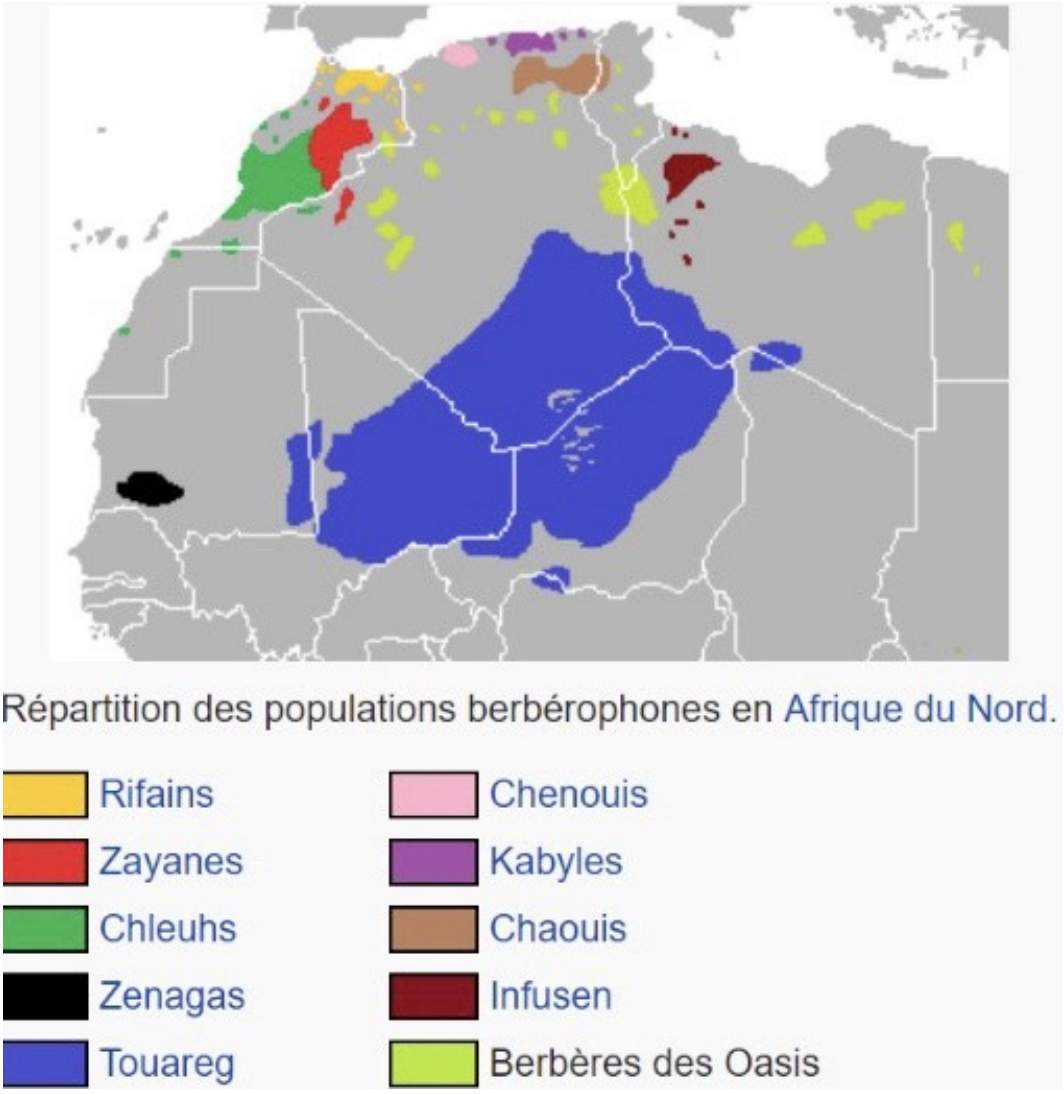

FIGURE 2.2 – Carte de la langue Berbère

2.4 Historique

Longtemps restées dans l'ombre, les origines de la langue Amazighe ont pris une nouvelle direction dans les derniers temps. Les revendications politico-indépendantistes ont laissé la place à des recherches scientifiques approfondies qui éclairent l'histoire d'un jour nouveau et de façon irrévocable.
Les dernières recherches sur les Imazighen touchent trois domaines des sciences humaines : l'anthropologie, la génétique et la linguistique, dans notre cas on se base sur la linguistique.
Selon les recherches linguistiques les origines de la langue Amazigh appartiennent à la famille des langues dites chamito-sémitiques. Selon les linguistes, l'Afro-asiatique viendrait d'Afrique orientale entre 10 000 et 17 000 ans.
D'après **Salem Chaker**, universitaire et professeur de linguistique berbère à l'INALCO, Les connaissances linguistiques se sont affinées. On sait que l'ancien libyque a un alphabet et une écriture, les tifinagh (pluriel de tafinek) qui ont été conservés par les Touareg . Les recherches ont encore progressé sur ce sujet.
Selon **Malika Rached** préhistorienne "On pourrait admettre que les inscriptions associées aux Paléoberbères du Tassili sont donc les plus anciens témoignages de l'écriture libyque en Afrique du Nord et qu'elles peuvent se situer vers 1300-1200 ans avant J.-C. Or, nous sommes en plein Sahara central, bien loin du domaine phénicien et carthaginois. "L'ensemble des données historiques anciennes s'ajoutent aux nouvelles aires des sciences humaines et brossent un tableau plus clair et plus scientifiquement précis des Berbères d'Afrique du nord.
L'époque du changement est peut-être déjà là. En 1996, la réforme de la Constitution Algérienne fait du Berbère l'une des composantes de l'identité nationale. En 2011, le Royaume du Maroc a promu le berbère deuxième langue nationale dans sa nouvelle Constitution. Le Tamazight est maintenant enseigné dans les écoles marocaines à l'aide de l'alphabet tifinagh. Il l'est également dans certaines wilayas d'Algérie où il a été introduit.[36]

2.5 Les variantes de langue Amazighe

Tamazight connaît plusieurs variations, comme le kabyle , le chaoui , Tachelhit, le Tarifit, Touareg, le Chenoua, Mozabite , et bien d'autres, ne se limitent pas à être de simples outils de communication. Elle sont également l'identite de la langue Amazighe.

— **Kabyle :** C'est la première variété Amazigh en nombre de locuteurs en Algérie, Située dans le nord du pays, elle est parlée par les kabyles dans les wilayas suivantes : Tizi Ouzou, Bejaïa, massif du Djurjura, Alger, Bouira, Sétif, Boumerdès, Jijel, Bordj Bou Arreridj . [35]
— **Chaoui :** Appelé aussi l'Aurasien, pratiqué par les chaouis du grand Aurès qui sont (Batna, khenchela, Biskra, Oum El Bouaghi, Souk-Ahras, Tébessa et Ain M'lila).
— **Chenoua :** Cette variante est parlée dans la région de Chénaoua, en Algérie.
— **Mozabite :** pratiqué par les Mozabites dans le nord du Sahara algérien (Ghardaïa etles autres villes ibadhites).
— **Touareg :** Dans plusieurs pays d'Afrique du Nord, notamment au Mali, au Niger, en

Algérie, au Burkina Faso et en Libye. Employé au sud d'Algérie par les Touarègues qui sont connus sous l'appellation de (les hommes bleus) qui se trouve dans :(Hoggar, Tassili, Aïr, Adrar des Iforas).

— **Tarifit :** C'est la variante de l'Amazighe parlée dans la région du Rif, située dans le nord du Maroc.

2.6 Kabyle

Le kabyle est un dialecte du berbère parlé en Kabylie, une région qui se situe au nord de l'Algérie, plus exactement au nord-est d'Alger. Il partage actuellement le terrain linguistique avec une variante du français maghrébin parlé en Kabylie et avec l'arabe sous ses deux formes. L'influence de ces langues les unes sur les autres est constatée dans plusieurs études linguistiques (Boucherit, 1987 ; Cheriguen,Merzouk, 2014 ; Lahrouchi, 2018). Le kabyle englobe plusieurs parlers qu'on classe en groupes relativement homogènes. Selon Nait-Zerrad K. (2001) [26], On peut envisager, quatre (04) groupes linguistiques, que l'on peut encore subdiviser.

— **Extrême occidental :** par les variétés de Boghni, Draa-El-Mizan, Tizi-Ghennif, ...
— **Occidental :** parlé dans les régions d'Ait Aissi, Ath-Yenni, Ath-Yirathen, Ath-Mangellat.
— **Oriental :** parlé dans les régions d'Ath Abbas et Mlikech (oriental ouest), Ath-Aidel et Ath-Khiar (oriental centre), Ath-Slimane (Oriental est).
— **Extrême oriental :** parlé dans les régions suivante d'Aokas, Ath-Smail et Melbou [17].

2.7 Ecriture la langue Kabyle

Il y a trois façons d'écrire une langue, quelle que soit sa complexité phonétique et sa syntaxe :

— **L'écriture phonétique :** C'est l'outil qui permet d'écrire toutes les occurrences d'une expression orale d'une langue donnée [32].
En Amazighe le mot **tawwurt** qui signifie **porte** s'écrit comme suivant :

/tawwurt/, /taggurt/, /tabburt/, /tappurt/...

— **L'écriture phonologique :** le système phonologique d'une langue, Lorsqu'on supprime toutes les différences phonétiques non pertinentes dans cette langue.
Dans lexemple precedent tawwurt, phonologiquement on éliminera les différences non pertinentes, on ecrit : /tawwurt/

— **L'écriture orthographique :** L'écriture orthographique suppose un choix immuable de transcription d'une langue. Un mot s'écrira toujours de la même manière.
Dans le mot tawwurt on va définir une seule et unique écriture qui est /tawwurt/ [32].

L'alphabet kabyle utilise l'alphabet latin adapté pour la variation phonologique de la langue. Les premières transcriptions étaient plus phonétiques que phonologiques, à titre d'exemple, Mouloud Mammeri a construit 46 lettres pour le berbère.

Depuis des décennies, les linguistes, au fil des études, ont allégé cet alphabet afin de ne prendre que les variations phonologiques.[2]

2.7.1 Système phonologique

Les Amazighologues ont défini le système phonologique suivant :

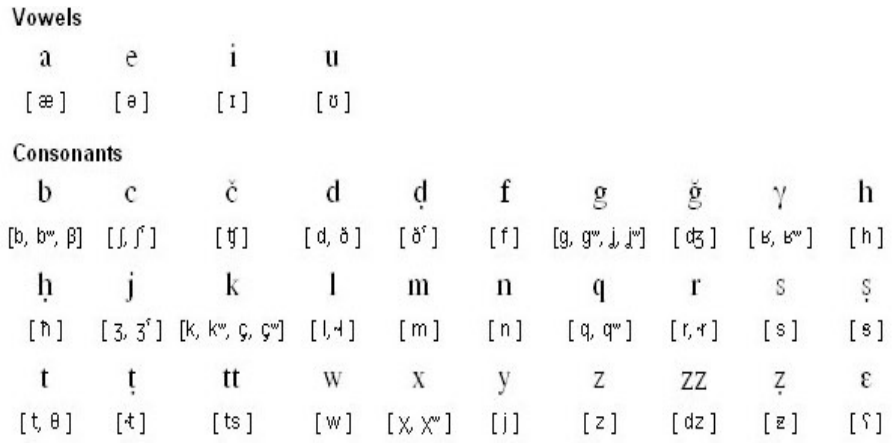

FIGURE 2.3 – Système phonologique

2.7.2 Système vocalique

Le vocalisme Berbère(Kabyle) se résume dans le triangle vocalique suivant :

FIGURE 2.4 – Triangle vocalique

On distingue trois voyelles de base /a/, /u/ et /i/
Deux semi voyelles /y/ et /w/.
Voyelle neutre /e/.

2.7.3 Système consonantique

Le système consonantique repose sur la tension, l'emphase et la labiovélarisation.

Spirante et Occlusives

Entre **/akal/** avec un /k/ spirant et **/akal/** avec un /kk/ occlusive, il y a une variation phonétique mais pas de différence de sens. Cette différence n'est pas introduite dans la transcription phonologique, on écrit un seul **/akal/** [32].

Les empathiques

Les phonèmes différenciés par l'emphase introduisent une variation de sens, par exemple :
/adar/ (rang) avec (d spirant ou occlusive) et **/aḍar/** (pied) avec d (empathique), il y aune différence de sens.
La différence d'emphase implique la différence de sens, donc lors de la transcription phonologique on fait attention de marquer l'emphase [32].

Labio-vélaire

Les labio-vélaires sont ignorés car elles sont des réalisations régionals $/g^w/$ $/k^w/$ [10].
Exemple :

— porte : tabburt/taggurt/tawwurt/ tappurt (/pp/ variante propre féminin)
on écrit **tawwurt**

— cuir : eww/ebb_o /egg_o on écrit **eww**

2.8 ASR appliqué dans la langue Amazighe

2.8.1 Défis rencontrés

Variabilité linguistique

La langue Amazighe est connue par ses dialectes variés ce qui explique la difficulté de création des modèles adaptés à toutes ces variations [25].

Manque de ressources linguistiques

le manque de transcription et de corpus annotés pour entrainer les modèles de reconnaissance automatique de la parole [25].

Modélisation acoustique

La phonologie unique de la langue Amazighe rend la modélisation des ses phonèmes et accents plus complexe [25].

Sensibilte aux contexte culturelle

Pour réaliser un système de reconnaissance précis il faut bien comprendre le contexte culturel et les particularités de cette langue [25].

2.8.2 Travaux réalisés

Malgré toutes ces défis rencontrés, plusieurs recherches ont été faites sur la reconnaissance automatique de la langue Amazighe en utilisant différentes techniques comme l'apprentissage automatique et les réseaux de neurones dans le but de la préservation culturelle et l'accessibilité numérique. On cite quelques-uns :

— «**The convolutional neural networks for Amazigh Speech Recognition System**» réalisé par [Meryam Telmem, Youssef Ghanou, Moulay Ismail, University Meknes, Morocco], ils ont présenté une approche basée sur les réseaux neuronaux convolutionnels pour construire un système de reconnaissance automatique de la parole pour la langue amazighe. Ce système est construit avec TensorFlow et utilise le mel frequency cepstral. Ce système est construit avec TensorFlow et utilise le coefficient (MFCC) pour extraire des caractéristiques. Afin de tester l'effet du sexe et de l'âge du sur la précision du modèle, le système a été entraîné et testé sur plusieurs ensembles de données.
— «**Comparative study of Amazigh Speech Recognition Systems based on different toolkits and approaches**», réalisé par [Safâa EL OUAHABI,Sara EL OUAHABI et Mohamed ATOUNTI, University Oujda, Morocco], L'objectif de cette étude est d'évaluer et de contraster les performances de différentes approches ASR appliquées à la langue Amazighe. Les techniques de modélisation markovienne, y compris les modèles de Markov de Markov cachés avec une distribution de mélange gaussien, les réseaux neuronaux convolutifs, la taille du vocabulaire, et enfin le choix de l'algorithme.
— «**Investigation Amazigh speech recognition using CMU tools**», réalisé par [Hassan Satori Fatima ElHaouss, International Journal of SpeechTechnology], L'objectif de cet article est de décrire le développement d'un système automatique continu indépendant du locuteur. Le système conçu est basé sur les outils Sphinx de l'Université Carnegie Mellon.

2.8.3 Contribution de ASR au développement de la langue Amazighe

L'ASR joue un rôle important dans la revitalisation et le developpement de la langue Amazighe, on cite ci-dessous quelques utilités de ASR :

— **Éducation et apprentissage :** Faciliter l'apprentissage de la langues et l'intércation avec les systèmes éducatifs qui reconnaissent la langue.
— **Reconnaissance internationale :** La présence de la langue Amzighe dans le domaine numérique lui permet d'être reconnue au niveau international.

— **Accessibilité :** Avec la communication par la voix, la langue Amazighe sera plus accessible pour les personnes handicapées visuellement.
— **Promotion culturelle :** Avec des application de poesie et de narration on peut promouvoir la langue Amazighe.
— **Revitalisation de la langue :** ASR est élément essentiel pour la préservation de la langue Amazighe.

Présentation de centre

Le CRLCA (Le centre de recherche en langue et culture amazighes) de Béjaïa, est un établissement publique à caractère scientifique et technologique à vocation sectorielle a été crée par le décret exécutif n° 17-95 du 26 février 2017 . Son organisation interne a été faite par l'arrêté interministériel du 12 juin 2019 portant organisation interne du centre de recherche en langue et culture amazighes.
Le centre a pour mission la réalisation des programmes de recherche scientifique dans les différents domaines de la langue et culture Amazighes.
En résumé, le Centre de Recherche en Langue et Culture Amazighes de Béjaïa joue un rôle central dans la promotion et le développement de l'amazighité à travers ses activités de recherche, de formation et de débat public sur ces questions.

2.9 Historique et statut

— Le CRLCA a été créé en 1991 au sein de l'Université de Béjaïa, d'abord en tant que département, puis promu en institut en 1996 avant de redevenir un département en 1999.
— Depuis la mise en place du système LMD, le département propose une licence, trois masters (géographie linguistique, civilisation et didactique) et deux doctorats

mission de CRLCA

Le centre a pour mission la réalisation des programmes de recherche scientifique dans les différents domaines de la langue et culture Amazighes.
A ce titre, le centre est chargé :

En matière de langue Amazighe

— De mettre en œuvre des projets de recherche dans les domaines des sciences et techniques du langue appliqués à la langue Amazighe dans toutes ses variétés ;
— De réaliser des travaux de recensement, de rationalisation, d'adaptation et de production de la terminologie scientifique et technique ;
— De participer à la prospection, à la sélection, à l'acquisition et à la diffusion des blogs lexique et la documentation à caractère pédagogique, scientifique et technique ;
— De développer des méthodes et techniques de traduction en vue de répondre aux besoins du système éducatif, de formation et de recherche ;
— D'exécuter des recherche théoriques et appliquées sur le développement de la langue et de la linguistique Amazighes, en liaison avec les institutions et établissements concernés.

En matière de culture Amazighe

— De recenser les as et coutumes et les pratiques culturelles et culturelles ;

— De transcrire et de valoriser les expression de la culture Amazighe ;
— De reconstituer le patrimoine immatériel ;
— D'étudier la pratique et l'interpretation des cultures orales et leur transmission à travers les étapes historique ;
— De réaliser des recherches sur l'évolution de la culture Amazighe dans toutes les étapes.

2.10 Organigrame dentreprise :

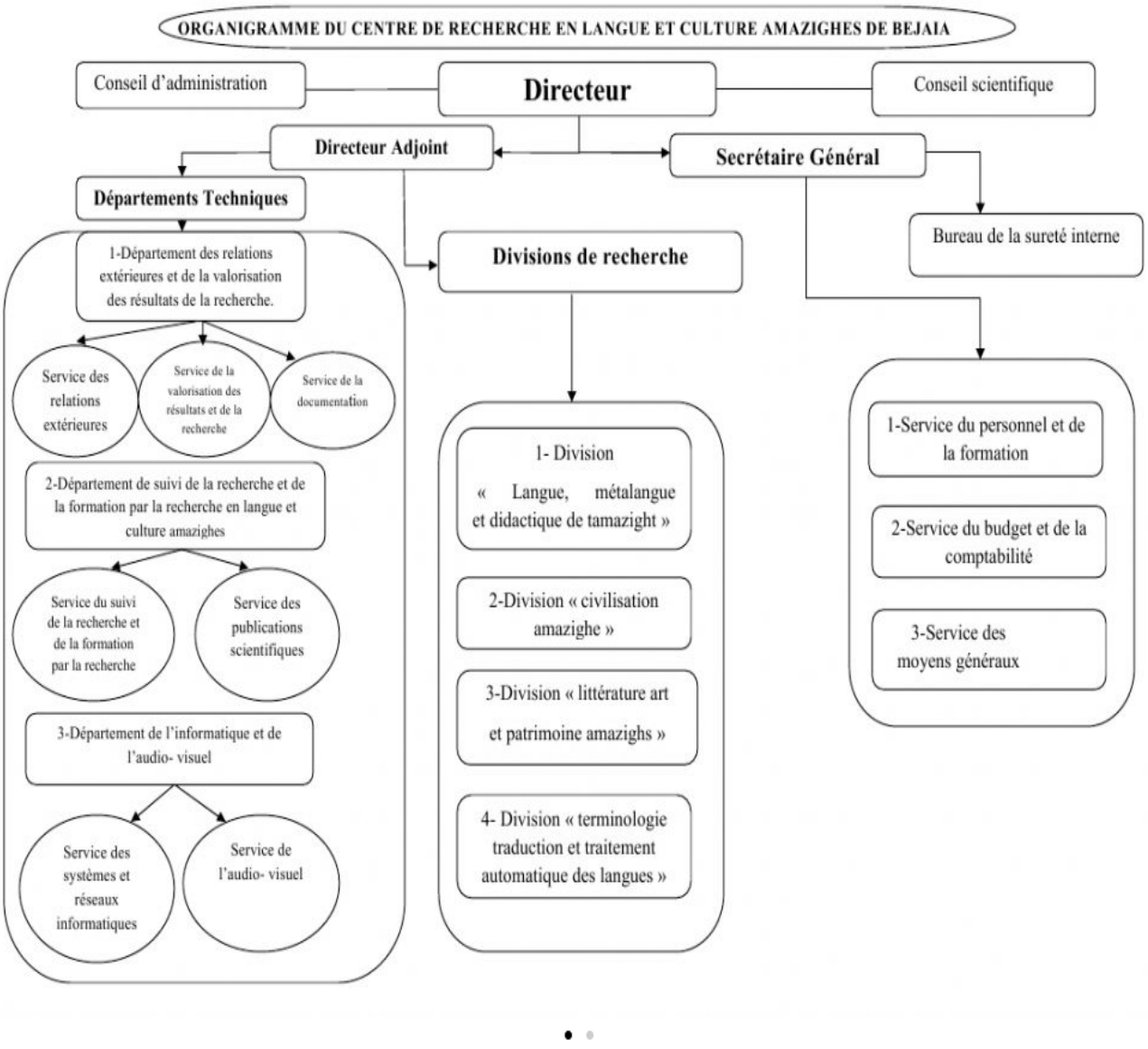

FIGURE 2.5 – Organigrame dentreprise.

Conclusion

Dans ce chapitres on a pu présenter la langue Amazighe comme un moyen de communication, un gardien de traditions millénaires et véhicule de l'histoire berbère.
Comme on a pu parler des ses aspets linguistiques et des dialectes variés et notamment le kabyle. Puis on a cité certains travaux réalisés dans le reconnaissance automatique de parole pour la langue Amazighe dans le but de préserver et promouvoir cette langue .
Dans le prochain chapitre, on va présenter l'une des étapes les plus importantes dans ASR qui est la récolte et le Prétraitement des données.

3

Collecte et traitement des données

Introduction

La collecte des données est une étape cruciale dans la recherche et l'analyse de données, elle permet d'obtenir les informations nécessaires et pertinentes qui seront analysées et traitées pour répondre à une hypothèse ou bien une problématique posée.

Sommaire

3.1 Échantillonnage et Corpus

3.1.1 Échantillonnage

Un échantillonnage est la phase qui consiste à sélectionner les individus que l'on souhaite interroger au sein de la population de base.

3.1.2 Types de l'échantillonnage

Aléatoires ou Représentative

— **Aléatoire :** Elle repose sur le principe de sélection au hasard ou aléatoire à partir d'une population ciblée.

— **Représentative :** Elle repose sur un choix arbitraire des unités, c'est l'enquêteur qui choisit les unités et non le hasard.

Pour une diversité linguistique des locuteurs on a utilisé le type aléatoire où le choix des participant est au hasard.

Quantitatives ou Qualitatives

— **Quantitative :** Elle permet de prouver des faits. L'échantillon doit être le plus représentatif possible. Plusieurs méthodes sont utilisées pour réaliser cette étude comme le sondage et le questionnaire.
— **Qualitative :** Lors d'une étude qualitative l'objectif n'est pas de vérifier sur le terrain des hypothèses mais de comprendre des phénomènes en profondeur. Pour réaliser cette étude on peut utiliser la recherche documentaire ,observation et l'analyse des discours.

Comme notre travail se base sur l'analyse et et le traitement de grandes quantité des vocaux, donc on a utilisé la méthode quantitative. Cette dernière permet d'améliorer la précision et l'efficacité de la reconnaissance vocale.

3.1.3 corpus

Un corpus est un ensemble de données langagières qui sont sélectionnées d'après des critères explicites pour servir d'échantillon de langage.

3.1.4 Types d'un corpus

— **Corpus arboré :** est un ensemble de données linguistiques dans lequel chaque phrase est annotée et analysée selon sa structure syntaxique.
Ces corpus sont souvent utilisés pour comprendre la structure grammaticales des langues.
— **Corpus parallèle :** est composée de textes dans deux langues ou plus, ces corpus sont souvent utilisés pour la traduction automatique.
— **Corpus phonologique (vocal) :** est un ensemble d'enregistrement vocaux ou audio, ces corpus sont utilisés pour la reconnaissance vocale et l'analyse de la parole.

3.1.5 Éléments d'un corpus

— **Support de production des données :** Cela fait références au dispositifs et aux outils utilisés, ça peut être des enregistrements à l'aide d'un dictaphone ou bien des transcriptions manuscrites...
— **Nature des documents constitutifs :** Ça peut être des articles,transcription des discours ou autres documents.
— **Forme et nature des annotations :** La nature des annotation dépends des objectifs de corpus.
— **Représentativité :** Un corpus représentatif inclut une diversité de textes ou d'enregistrement vocaux qui captures différentes variations linguistiques.

3.2 Processus de collecte des données

3.2.1 Préparation des phonèmes et des phrases

La première étape à effectuer consiste à identifier les lettres et les phonèmes de la langue Amazighe (Kabyle),, qui s'efforce de préserver et de documenter cette langue.
Ce travail est fait par le centre de recherche la langue et la culture Amazighe. Ils ont associé pour chaque phonème un exemple de Noms (15) et de verbes (15) contenant ce phonème dans différentes positions (début,milieu,fin).
Des exemples de phrases ont ensuite été créés ,our chaque nom et verbe sélectionné afin de montrer comment les utiliser dans leur contexte.
Ces phrases aideront à comprendre comment les phonèmes sont utilisés dans un discours complet et varié.
Ceci est essentiel pour l'analyse du langage et la reconnaissance vocale.
En tout, ce corpus se compose de 37 phonèmes. Avec 30 phrases par phonème, cela donne un total de 1 110 phrases.
Chaque liste de phrases (associée à un phonème) est répétée cinq fois pour améliorer la robustesse de l'analyse et garantir des résultats fiables.
Ces répétitions permettent de vérifier la constance et la précision des modèles de reconnaissance vocale développés.
Le corpus final contenant toutes ces itérations constituera une ressource précieuse pour les chercheurs travaillant sur la langue kabyle, fournissant une base solide pour développer des techniques de reconnaissance vocale précises et efficaces.

3.2.2 Les paramètre et réglage

Il nous faut un corpus bien spécifique, c'est pour cette raison on a pris en considération 4 paramètre essentiels

Âge : Pour la rubrique âge, nous avons essayé de toucher toute les tranches d'âge possibles pour avoir des résultats plus vastes, la parole varie également avec l'âge, pour des raisons à la fois générationnelles et physiologiques.

Région : En ce qui concerne les régions on a opté pour 3 villes essentielle qui sont Bejaia (997enregistrement), Tizi Ouzou (36enregistrement) et Bouira(72enregistrement), et on a visé toute les régions de Béjaia de l'est au west (Tazmalt , Akbou , Sidi aich, Amizour, Béjaia ville , Tichy, Aokas, Soukltenin, Aithsmail , Kherrata).

Sexe : Les femmes et les hommes ont des différentes voix, c'est pour ça on a pris les deux sexe.

Statut sociale : Nous avons inclus plusieurs statuts sociales (étudiants, enseignants, agents, etc.),pour capturer une diversité de styles de parole.

Le choix de ces paramètres vise à garantir que le corpus soit représentatif et inclusif, ce qui reflète les variations linguistiques et sociales de la langue Kabyle.

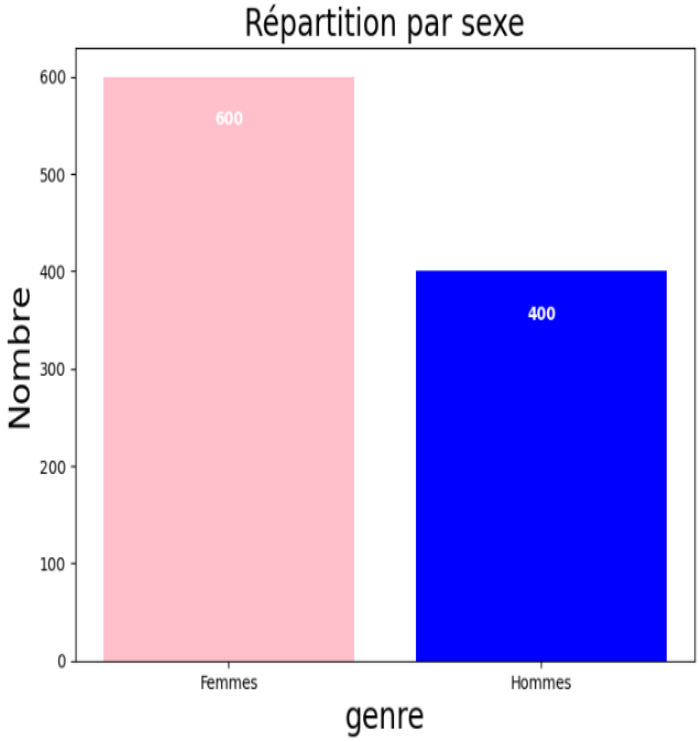

FIGURE 3.1 – statistiques des femmmes et homme dans notre base de données.

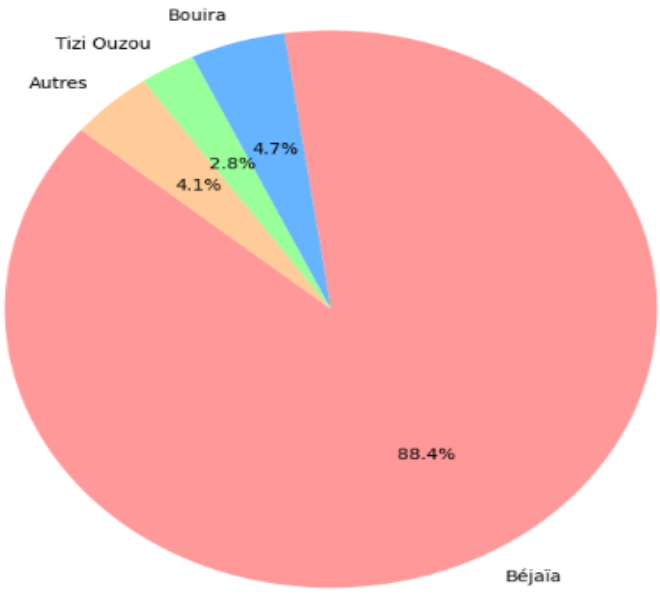

FIGURE 3.2 – statistiques des villes dans notre base de données.

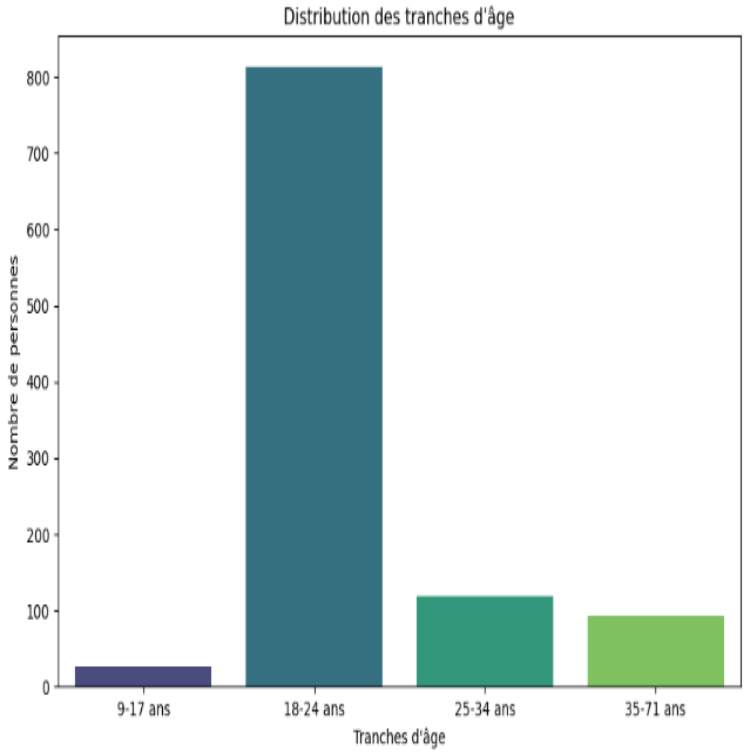

FIGURE 3.3 – statistiques des tranches d'âges.

3.2.3 Choix des équipements et du terrain

Pour nous assurer que notre collecte est effectuée avec précision et efficacité, nous avons choisi des zones représentatives de la diversité linguistique et sociale de la population kabyle. Cela est essentiel pour capturer pleinement les nuances de cette culture linguistique riche et variée.

— **Formation des équipes :** Formation de quatre équipes de deux personnes où chaque binôme travaille dans une zone précise.

— **Région :** On a choisi différentes zones de la ville, en commençant par le milieu universitaire qui rassemble différentes tranches d'âge et régions kabyles, que ce soit de la même ville ou venant d'autre villes, puis passant au centre-ville où il y a une diversité d'âge et de statut social.

3.2.4 Enregistrement vocale avec dictaphone

Pour l'enregistrement des voix, on a utilisé un dictaphone numérique fourni par le centre de recherche de la langue et la culture Amazighe.

Un dictaphone numérique Il s'agit d'un appareil multimédia qui permet d'enregistrer des sons et les stocker directement sur une carte SD Mini SD. Parmi ses avantages, on trouve la très bonne qualité de captation sonore et la facilité de le transporter.

FIGURE 3.4 – Dictaphone numérique

Ce qui en fait un choix parfait pour enregistrer sur le terrain

3.2.5 constitution de base de données

GLIDE est une plateforme sophistiquée, sans nécessiter de connaissances en codage, qui permet de créer une application puissante et riche en fonctionnalités qui n'a pas d'expérience en matière de codage.
Elle utilise une interface visuelle pour ajouter des éléments, des images, des boutons et personnaliser l'apparence de l'application. Glide se connecte à des bases de données telles que Google Sheets et Airtable pour récupérer les données et les utiliser pour créer des applications.
En résumé, Glide est un outil puissant et facile à utiliser pour créer des applications mobiles et Web sans codage.
Idéal pour les entreprises, les entrepreneurs et les enseignants qui souhaitent créer des applications rapidement et efficacement.

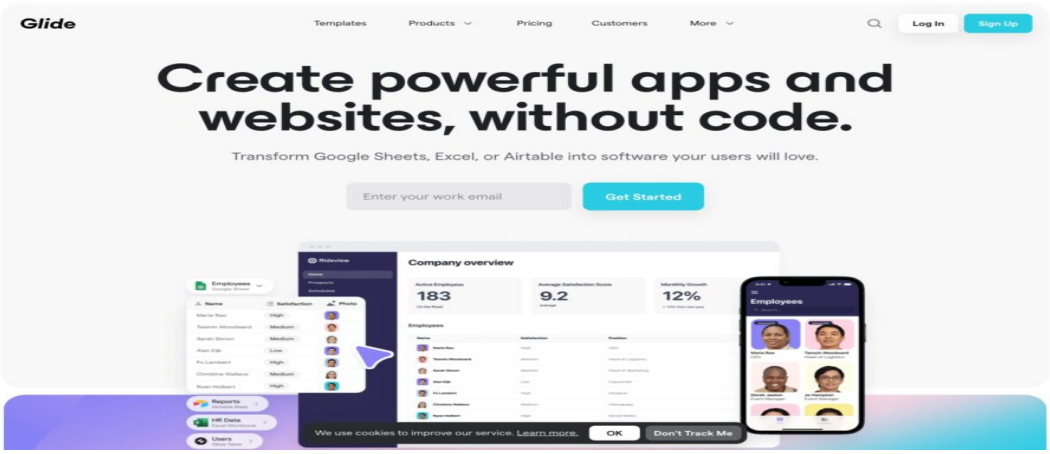

FIGURE 3.5 – Logiciel Glide

3.2.6 Enregistrement sur la base de données

L'application TALN Avec l'application TALN-RV-DATA créée par le centre de recherche de la culture Amazighe, on a pu enregistrer les données de chaque participant (Âge,sexe,région et statut sociale).

3.3 Prétraitement des données

Le Prétraitement des données ou bien en anglais Data Preprocessing est une étape essentielle qui consiste à nettoyer, trier et regrouper des données pour qu'elle soit prête à traiter correctement.

3.3.1 Prétraitement (Segmentation)

La phase de segmentation joue un rôle très important dans les systèmes de reconnaissance vocale et nécessite un intérêt particulier, de notre part on a traité les enregistrements manuellement en utilisant le logiciel Audacity et ça en supprimant les bruits , parasites et en éliminant les enregistrements vocaux de mauvaise qualité. Cela avait pour but d'améliorer la clarté des enregistrements, ce qui est essentiel pour une analyse précise.
Ensuit Nous les avons segmentés en plusieurs parties : le phonème, le mot et la phrase qui correspond.
Après chaque segmentation, nous effectuons des tests par écoute afin de s'assurer que la segmentation a été bien faite et pour vérifier la précision de la segmentation, corrigeant ainsi les éventuelles erreurs.
Pour bien organiser la base de données, on a attribué chaque enregistrement vocale a un identifiant unique qui est associé a des critères d'âge, sexe, région et statut sociale.
L'attribution d'identifiants uniques facilite le suivi et l'analyse des enregistrements, ce qui permet une gestion plus efficace des données.

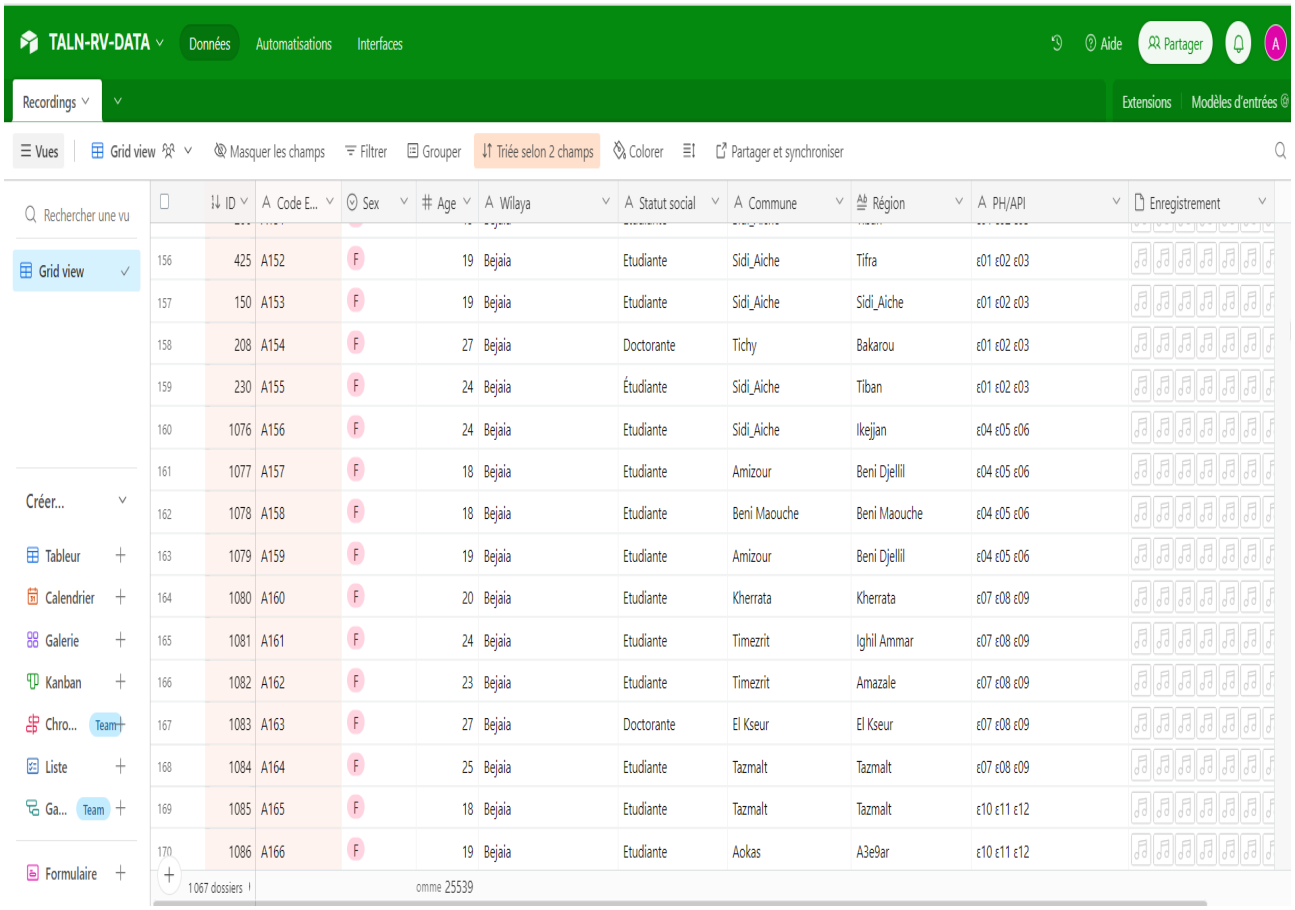

FIGURE 3.6 – La base de données TALN

Audacity :

Audacity est le logiciel gratuit et open source d'enregistrement et d'édition audio le plus populaire au monde, crée par Roger [**B. Dannenberg (d)**] et [**Dominic Mazzoni (d)**]. Ce logiciel polyvalent permet l'enregistrement de sons numériques et l'édition de sources audio-numériques sous différents formats, le rendant idéal pour les chercheurs et tous ceux qui souhaitent faire preuve de créativité avec le son.
Grâce à ses nombreuses fonctionnalités et à son interface conviviale, Audacity est un outil précieux pour travailler avec des données audio.

Ses fonctionnalités

— **Enregitrement et édition audio :** Avec Audacity on peut enregitrer et éditer des pistes audio, enregistrer des voix et des podcasts.

— **Formats des fichiers :** Importer et exporter des fichiers audio sous différents formats.

— **Analyse audio :** Pour visualiser des fréquences audio on utilise des spectogrammes.

— **Multi-pistes :** On peut faire des projets multi-pistes avec audacity.

FIGURE 3.7 – Logiciel Audacity

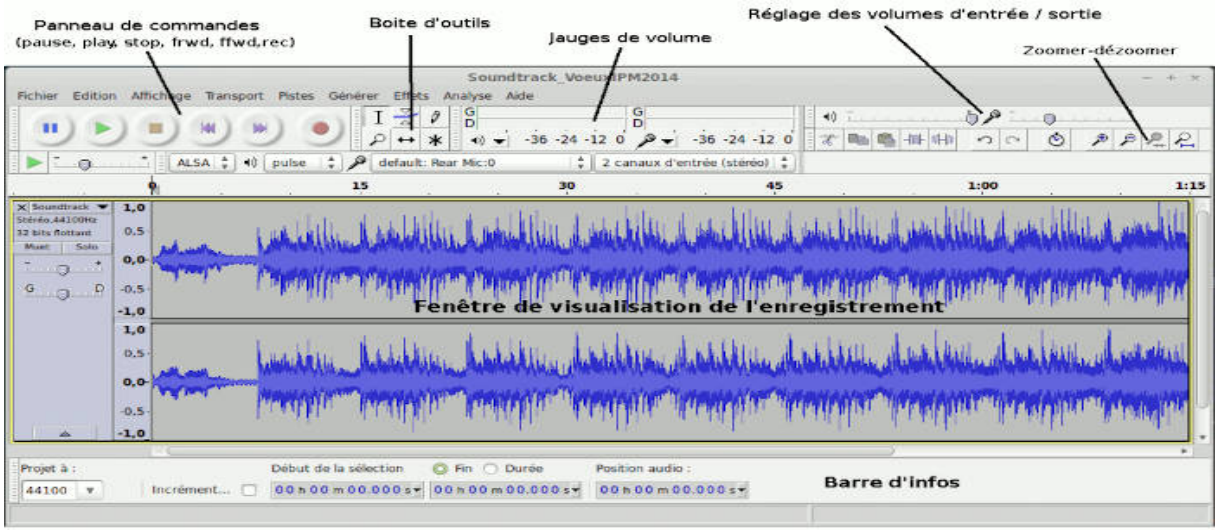

FIGURE 3.8 – Barre d'outils d'audacity

3.3.2 Étiquetage et création d'un fichier csv

L'étiquetage ou bien Data lebelling en anglais, c'est connecter des informations à des divers points des données pour que les algorithmes puissent comprendre la signification.

Type d'étiquetage

— **Vision par ordinateur :** Pour catégoriser des images et identifier des objets dans des images, les algorithmes de machines learning ont toujours besoin des étiquetes des ces données.
— **Traitement de langage naturel :** Les étiquetes NLP sont utilisées par des applications textuels afin d'identifier des mots, des phrases.
— **Traitement audio :** Étiqueter les sons du contenu audio permet au algorithmes d'apprentissage automatique de reconnaitre les sons.

Avantage d'étiquetage

— **Précision des données :** La méthode utilisée pour étiqueter a un impact sur la précision des résultats obtenus.
— **Qualite :** Étiqutage est souvent utilisé pour améliorer la qualité des applications d'apprentissage automatique.
— **Decouverte des opportunites commercials :** Étiqutage précis avec des analyses permet au entreprises de définir des opportunites génératrices.

Pour organiser et structurer nos données de manière lisible et accessible, nous avons créé un fichier de type CSV sur LibreOffice, avec trois colonnes : la première contient le code vocal, la deuxième contient sa transcription et la troisième contient le chemin ce dernier.
Notre base de données contient au total 9513 vocaux.
Dans notre cas, on a bien utilisé le typ3 d'étiquetage de traitement audio, ce qui signifie que chaque enregistrement a sa propre transcription.

A1 | Identifiant

	A	B	C
5306	651_Č03_3	Čeɛčeɛ	C:\Users\pc\Desktop\TALN\G082_Č03_3.wav
5307	651_Č03_4	Černen-d kra yidrimen ɣer lqaɛa	C:\Users\pc\Desktop\TALN\G082_Č03_4.wav
5308	651_Č03_5	Tečaɛčiɛ fell-as qrib tečča-tt	C:\Users\pc\Desktop\TALN\G082_Č03_5.wav
5309	652_Č03_1	č	C:\Users\pc\Desktop\TALN\G083_Č03_1.wav
5310	652_Č03_2	Černen	C:\Users\pc\Desktop\TALN\G083_Č03_2.wav
5311	652_Č03_3	Čeɛčeɛ	C:\Users\pc\Desktop\TALN\G083_Č03_3.wav
5312	652_Č03_4	Černen-d kra yidrimen ɣer lqaɛa	C:\Users\pc\Desktop\TALN\G083_Č03_4.wav
5313	652_Č03_5	Tečaɛčiɛ fell-as qrib tečča-tt	C:\Users\pc\Desktop\TALN\G083_Č03_5.wav
5314	652_Č05_1	č	C:\Users\pc\Desktop\TALN\G083_Č05_1.wav
5315	652_Č05_2	Čefčef	C:\Users\pc\Desktop\TALN\G083_Č05_2.wav
5316	652_Č05_3	Čenčen	C:\Users\pc\Desktop\TALN\G083_Č05_3.wav
5317	652_Č05_4	Kul ɛmal ičefčif	C:\Users\pc\Desktop\TALN\G083_Č05_4.wav
5318	652_Č05_5	Teddu deg ubrid axelxal-is yečenčun	C:\Users\pc\Desktop\TALN\G083_Č05_5.wav
5319	655_Č05_1	č	C:\Users\pc\Desktop\TALN\G084_Č05_1.wav
5320	655_Č05_2	Čefčef	C:\Users\pc\Desktop\TALN\G084_Č05_2.wav
5321	655_Č05_3	Čenčen	C:\Users\pc\Desktop\TALN\G084_Č05_3.wav
5322	655_Č05_4	Kul ɛmal ičefčif	C:\Users\pc\Desktop\TALN\G084_Č05_4.wav
5323	655_Č05_5	Teddu deg ubrid axelxal-is yečenčun	C:\Users\pc\Desktop\TALN\G084_Č05_5.wav
5324	655_Č06_1	č	C:\Users\pc\Desktop\TALN\G084_Č06_1.wav
5325	655_Č06_2	Senčew	C:\Users\pc\Desktop\TALN\G084_Č06_2.wav
5326	655_Č06_3	Beččeč	C:\Users\pc\Desktop\TALN\G084_Č06_3.wav
5327	655_Č06_4	Senčaweɣ iyuzad leqbayel	C:\Users\pc\Desktop\TALN\G084_Č06_4.wav
5328	655_Č06_5	Ibeččeč akk lxebz-nni d imersulen	C:\Users\pc\Desktop\TALN\G084_Č06_5.wav
5329	657_Č05_1	č	C:\Users\pc\Desktop\TALN\G085_Č05_1.wav
5330	657_Č05_2	Čefčef	C:\Users\pc\Desktop\TALN\G085_Č05_2.wav
5331	657_Č05_3	Čenčen	C:\Users\pc\Desktop\TALN\G085_Č05_3.wav
5332	657_Č05_4	Kul ɛmal ičefčif	C:\Users\pc\Desktop\TALN\G085_Č05_4.wav
5333	657_Č05_5	Teddu deg ubrid axelxal-is yečenčun	C:\Users\pc\Desktop\TALN\G085_Č05_5.wav
5334	657_Č06_1	č	C:\Users\pc\Desktop\TALN\G085_Č06_1.wav
5335	657_Č06_2	Senčew	C:\Users\pc\Desktop\TALN\G085_Č06_2.wav
5336	657_Č06_3	Beččeč	C:\Users\pc\Desktop\TALN\G085_Č06_3.wav
5337	657_Č06_4	Senčaweɣ iyuzad leqbayel	C:\Users\pc\Desktop\TALN\G085_Č06_4.wav

FIGURE 3.9 – Création d'un fichier CSV

Conclusion

Dans ce chapitre, on a abordé les étapes essentielles de la collecte et du traitement des données nécessaires à la réalisation de notre étude. La qualité et la représentativité des données sont cruciales pour garantir la fiabilité des résultats obtenus. Donc, on a commencé par détailler les méthodes d'échantillonnage et les types de corpus utilisés. Ensuite, on a expliqué les processus de collecte commençant par la préparation des phonèmes, le choix de paramétrage puis l'enregistrement, le choix des équipes et des régions.
Par la fin, nous sommes passés au prétraitement des données et aux logiciels utilisés.

Ces étapes sont fondamentales pour préparer une base de données robuste, indispensable pour les analyses ultérieures et le développement d'applications telles que la reconnaissance vocale.

Dans le prochain chapitre, nous aborderons les différentes étapes de l'implémentation de cette base de données ainsi que le plateforme utilisée.

4

Réalisation d'un système de reconnaissance de la parole

Introduction

Pour construire des systèmes de reconnaissance vocale, plusieurs plateformes sont disponibles telles que Sphinx, Common Voice et Kaldi.

Kaldi est une plateforme de reconnaissance vocale qui gère la reconnaissance dans plusieurs langues en utilisant des modèles acoustiques spécifiques à chaque langue. Dans notre chapitre, nous explorons les caractéristiques générales des plateformes de reconnaissance vocale.
Ensuite, nous nous concentrerons spécifiquement sur la plateforme Kaldi. Nous donnerons un aperçu détaillé sur son architecture, ses fonctionnalités et la manière dont notre jeu de données a été intégré et utilisé pour l'entraînement, et l'évaluation des modèles de reconnaissance vocale.
Chaque composant du système de reconnaissance vocale pris en charge par cette boîte à outils sera également illustré, comme l'extraction des caractéristiques, la modélisation acoustique et le décodage.

Sommaire

4.1 Plateformes de reconnaissance vocale

Il existe plusieurs librairies pour le domaine de RAP, parmi ces librairies : HTK, Sphinx, Kaldi, ASR de Matlab, Java Speech, ISIP.

4.2 La plateforme kaldi

Kaldi est une boîte à outils de reconnaissance vocale composée d'une bibliothèque, de programmes en ligne de commande et de scripts pour la modélisation acoustique.
Kaldi est un outil open-source de reconnaissance vocale écrit en C++. L'objectif de Kaldi est de fournir du code flexible et "facile" à comprendre. Il est alors possible de modifier le code source pour répondre à des besoins spécifiques et rajouter des fonctionnalités.
Kaldi déploie plusieurs décodeurs pour évaluer les modèles acoustiques (AM) de Kaldi. Il existe naturellement d'autres alternatives open-source telles que Sphinx ou HTK, toutes inférieures à Kaldi en terme de performance et de support.
C'est pour les raisons citées ci-dessus, que Kaldi a été retenu pour le développement du système de reconnaissance vocale [17].

4.2.1 Sphinx

Sphinx est une librairie de reconnaissance vocale gratuitement téléchargeable, avec la possibilité de modifier le code source, il a la capacité d'implémenter des systèmes avec un large vocabulaire, indépendants du locuteur [17].

4.2.2 HTK

Hidden Markov Model Toolbox (HTK) est une boîte à outils dédiée aux Modèles de Markov Cachés est principalement utilisée pour la reconnaissance de la parole. Elle se compose d'un ensemble de modules et d'outils disponibles gratuitement et téléchargeables à partir du site. HTK est implémenté en langage C et il s'exécute en ligne de commande. Il est capable de mettre en oeuvre un grand vocabulaire, indépendamment du locuteur et est applicable sur n'importe quelle langue. La documentation sur HTK est très riche avec des exemples pratiques [5].

4.2.3 Simon

Simon est un logiciel de reconnaissance vocale technologiquement avancé et très flexible.
Le logiciel offre une personnalisation de haut niveau pour toutes les applications et peut donc être utilisé avec tous les systèmes dans lesquels la reconnaissance vocale est requise.

La technologie derrière Simon comprend les bibliothèques KDE, ainsi que HTK et CMU SPHINX. Le logiciel est disponible en open source et gratuitement pour les systèmes d'exploitation Windows et Linux.

4.2.4 Matlab

Matlab inclut une boîte à outil (toolbox) incluant des algorithmes d'apprentissage artificiel basés sur les modèles de Markov cachés et des algorithmes de détection des séquences temporelles hors ligne et en ligne [5].

4.2.5 VoxForge

VoxForge a été mis en place pour recueillir la parole transcrite pour une utilisation dans les moteurs de reconnaissance de la parole Open Source (Speech Recognition Engines "SRE") tels que ISIP, HTK, Julius et Sphinx[29].

4.2.6 Pourquoi kaldi

Kaldi est une plateforme de pointe en matière de reconnaissance automatique de la parole, offrant des performances élevées et une personnalisation avancée des modèles. Que ce soit pour la transcription audio, la recherche en traitement du langage naturel, ou la création d'assistants vocaux, Kaldi est un choix puissant pour les projets de reconnaissance vocale.

4.2.7 Installation

La platforme Kaldi fonctionne généralement sur linux.

Tout d'abord, on doit installer Git. Ouvrez le terminal et tapez la commande suivante,l'installation peut prendre plusieurs heures.

```
git clone https ://github.com/kaldi-asr/kaldi.git kaldi –origin upstream cd kaldi
```

Pour une mise à jour de kaldi on utilise la commande suivante.

```
mira@mira-Inspiron-3521 : /Desktop/kaldi sudo apt upgrade
```

Aprés la mise à jour générale on doit installer les libraies suivantes tools, extras et src en suivant les étapes et les commandes suivantes :

tools
mira@mira-Inspiron-3521 :/Desktop/kaldi cd tools/ accéder à tools
mira@mira-Inspiron-3521 :/Desktop/kaldi/tools extras/check dependencies.sh
verifier les installation dependantes
mira@mira-Inspiron-3521 :/nproc vérifier le nombre de processeurs
mira@mira-Inspiron-3521 :/Desktop/kaldi/tools make pour la compilation
mira@mira-Inspiron-3521 :/Desktop/kaldi/tools make -j 12

extras
mira@mira-Inspiron-3521 :/Desktop/kaldi/tools extras/install irstlm.sh

Aprés l'installation de tous les packages dans tools et extras on passe au src.

src
mira@mira-Inspiron-3521 :/Desktop/kaldi/tools cd ../src/ accéder a src.
mira@mira-Inspiron-3521 :/Desktop/kaldi/src ./configure pour la configuration
mira@mira-Inspiron-3521 :/Desktop/kaldi/src make depend
mira@mira-Inspiron-3521 :/Desktop/kaldi/src make pour la compilation

4.2.8 Fonctionnement général de Kaldi

Les composants de la plateforme Kaldi sont essentiels pour son fonctionnement dans la reconnaissance automatique de la parole (RAP). Voici une vue simplifiée des différents composants de Kaldi :

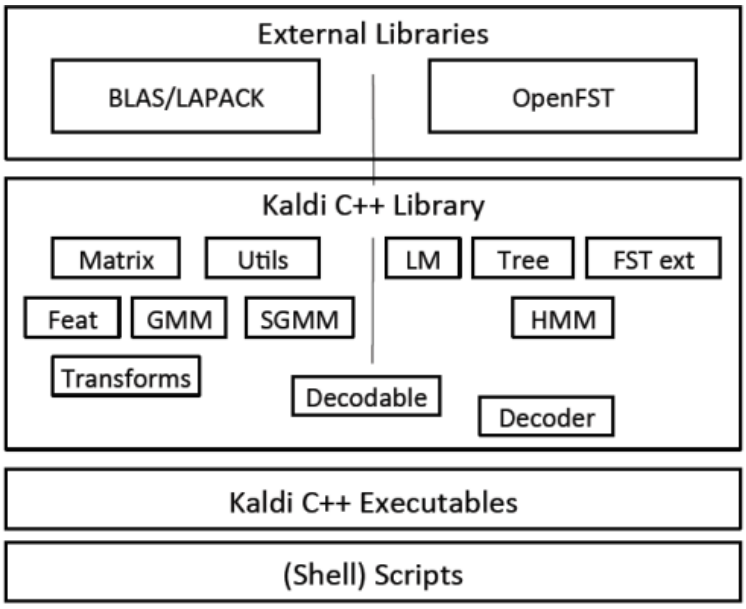

FIGURE 4.1 – Les différentes composantes de Kaldi [30].

Kaldi Library : Les bibliothèques externes sont utilisées dans le code C++ de Kaldi

pour implémenter les diverses briques technologiques.
Kaldi Executables : Ces briques, une fois assemblées à partir des bibliothèques externes, donnent lieu à des fonctionnalités opérationnelles pour la reconnaissance vocale.
Scripts Shell : Les fonctionnalités générées par les Kaldi Executables permettent de composer le système de reconnaissance vocale, facilitant ainsi l'utilisation de la plateforme dans des applications pratiques.
External Libraries : Le finite-state framework de OpenFst, une bibliothèque essentielle pour Kaldi . "Basic Linear Algebra Subroutines" (BLAS) et "Linear Algebra PACKage" (LAPACK) pour le calcul algébrique [11].

En résumé, les Kaldi Libraries sont utilisées pour implémenter les algorithmes de reconnaissance vocale, offrir une flexibilité et une personnalisation dans le développement de systèmes, et bénéficier d'une mise à jour constante et du support de la communauté pour les tâches de reconnaissance vocale automatique.

4.2.9 Utilisations courantes

Kaldi est utilisé dans un large éventail d'applications, notamment :
Transcription audio : Il est capable de transcrire de l'audio en texte, ce qui est utile pour les services de transcription automatisée.
Assistants vocaux : Kaldi peut être intégré dans des assistants vocaux pour permettre des interactions vocales avec des systèmes informatiques.
Recherche en Traitement du Langage Naturel : Il est utilisé dans la recherche académique pour développer de nouveaux modèles de reconnaissance vocale et d'autres applications en traitement du langage naturel.
Systèmes de commande vocale : Kaldi peut être utilisé pour développer des systèmes de commande vocale pour des applications telles que la domotique.

4.2.10 Structure de la plateforme Kaldi

Après l'installation de la boîte à outils Kaldi, on trouve que son répertoire contient l'arborescence suivante :
kaldi-trunk/
cmake/
docker/
egs/
misc/
scripts/
src/
tools/
windows/
README.md

egs : Contient plein d'exemples des scripts qui vous permet de construire rapidement des systèmes ASR depuis plus de 30 corpus vocaux populaires. On note que la majorité des corpus appartiennent à : Linguistic Data Consortium « LDC » et qui ne sont pas

gratuits .
src : Contient le code source de Kaldi (Kaldi Libraries et Kaldi Executables).
Misc : Contient des outils supplémentaires, il est optionnel pour le fonctionnement correct Kaldi.
outils : Contient des composants utiles et des outils externes .
windows : Contient des outils pour exécuter Kaldi sous Windows.

4.3 Processus de réalisation d'un ASR avec Kaldi

4.3.1 Préparation des données

Pour commencer avec la plateforme Kaldi on doit fournir ces fichiers :
fichier wav : Ce fichier contient les audio enregitrés.
fichier texte : Ce dernier contient la transcription de chaque audio.
dictionnaire : Ce dictionnaire contient les phonèmes et les sons.

4.3.2 Extraction des caractéristiques

Après la constitution de la base de données, kaldi utilise la méthode Mel-Frequency-Cepstral (MFCC) pour extraire les caractéristiques nécessaires et convertir le signal audio de l'analogique au numérique.

Les paramètres de MFCC sont choisis selon les données qu'on possède.

La normalisation des données vient après l'extraction des données, c'est une étape qui permet d'améliorer les résultats en éliminant les redondances .
Dans cette étape, Kaldi utilise Cepstral Mean and Variance Normalization (CMVN).

4.3.3 Modèle acoustique

Avant de passer à l'entraînement du modèle acoustique on doit passer par l'alignement et l'entraînement des monophones et des triphones.

Le modèle monophone est un modèle acoustique qui ne contient pas d'informations contextuelles sur les phonèmes précédents ou suivants, il passe par les deux étapes suivantes :
— Entraînement des modèles associés à chaque phonème grâce à l'algorithme Expectation Maximization. On obtient ensuite la transcription phonétique du signal audio utilisé.
— Alignement de la transcription avec le signal audio. Cette étape permet d'associer un phonème à un son.
Le modèle Triphone est introduit pour prendre en compte la variabilité de la prononciation des phonèmes dans leur contexte et ainsi améliorer grandement les performances de reconnaissance.

L'entraînement du modèle triphone suit les mêmes étapes que le modèle monophone.

Le modèle HMM-DNN

Les réseaux de neurones profonds (DNN) sont le dernier sujet en matière de reconnaissance vocale. Depuis 2010 environ, de nombreux articles ont été publiés dans ce domaine, et certaines des plus grandes entreprises comme Google, Microsoft commencent à utiliser les DNN dans leurs systèmes de production.
Dans kaldi, il y a trois bases de code distinctes pour les réseaux neuronaux profonds qui sont [1] :

— **nnet1 :** Se trouve dans les sous-répertoires de code nnet/ et nnetbin/, et est principalement maintenu par Karel Vesely.
La configuration de Karel (nnet1) prend en charge la formation sur une seule carte GPU, ce qui permet à la mise en œuvre plus simple et relativement facile à modifier[1].
— **nnet2 :** Est situé dans les sous-répertoires de code nnet2/ et nnet2bin/, et est principalement maintenu par Daniel Povey (ce code était à l'origine basé sur une version antérieure du code de Karel, mais il a été largement réécrit).
La configuration de Dan (nnet2) est plus flexible dans la façon dont vous pouvez vous entraîner : elle prend en charge l'utilisation de plusieurs GPU[1].
— **nnet3 :** Se trouve dans les sous-répertoires de code nnet3/ et nnet3bin/.
La nouvelle configuration, nnet3 (nnet3), est la plus utilisé dans les travaux récents [1].

Le Modèle HMM-GMM

le modèle GMM-HMM est un modèle statistique qui combine les modèles de Markov caché (HMM) et de mélange gaussien (GMM). Dans un modèle GMM-HMM, chaque état caché est modélisé par un mélange gaussien, ce qui permet de modéliser des distributions de probabilité complexes pour les observations. Les transitions entre les états sont modélisées par des matrices de transition, comme dans les modèles HMM classiques [34].

4.3.4 Modèle de langage

Le modèle de langage est généré par SRILM - The SRI Language Modeling Toolkit. Cet outil permet la création de modèles de langage statistiques. Il suffit de lui fournir du texte représentatif de l'application cible (par exemple les transcriptions du train set) ainsi que l'ordre du n-gram désiré. Il se charge de calculer les probabilités log d'apparition de séquences de k mots avec $k \in [1, n]$. Le résultat est donné sous la forme d'un fichier (.arpa), qui sera transformé par la suite au format (.fst) par l'intermédiaire d'un script Kaldi.

4.3.5 Décodeur

Pour représenter les divers composants du système de reconnaissance (HMM, modèles de dépendance du contexte, grammaire), Kaldi utilise des weighted finite-state transducers (WFST).
Lors du décodage, le décodeur explore le graphe HCLG pour trouver la meilleure séquence de mots correspondant aux observations acoustiques.
H(HMM) : contient le modèle de markov caché. Ses symbols d'entrée sont des identifiants de transition et les symbols de sortie sont des contextes dependent de phonèmes.
C(Contexte) : Ses symbols de sortie sont des phonèmes et les symbols d'entrée sont des contextes dependant des phonèmes. On peut dire qu'elle représente la dependance contextuelle.
L(lexique) : Ses symboles d'entrée sont des phonèmes et les symboles de sortie sont des mots. Elle contient les trasncriptions phonèmiques.
G(Grammaire) : Représente le modèle de langage.Ses symbols d'entrée et de sortie sont identiques qui sont des mots ou bien des unités lexicals.

Des techniques d'optimisation comme la déterminisation et la minimisation du graphe WFST permettent de réduire la complexité du décodage [23].

4.3.6 Évaluation

L'évaluation de la reconnaissance de la parole est très importante pour comparer les modèles, le critère le plus utilisé est Word Error Rate (WER).
On calcule le WER des divers modèles à partir des phrases du test set. Il existe trois types d'erreurs de reconnaissance : substitution (S), omission (O) et insertion (I), comme il est expliqué dans le prmier chapitre.

4.4 Application sur notre base de données

4.4.1 Nettoyage des données

Nettoyage des données, ou bien data cleaning est une étape qui permet de corriger les données altérés et inexacte, dans le but d'avoir un ensemble de données cohérent et améliorer la qualité des ces données 4.2.

FIGURE 4.2 – Nettoyage des données

4.4.2 Préparation des données

Tout d'abord, on a commencé par créer un projet dans Kaldi qu'on a nommé 'mon projet kabyle'.
Pour implémenter notre base de données csv, nous avons créé un répertoire **mon data** dans **kaldi egs** et ouvrir deux dossier.
wav : Contient les fichiers audio
texte : Contient la transcription de chaque audio.

FIGURE 4.3 – Résultat après l'implémentation

Pour afficher les résultats de l'implémentation on utilise la commande suivante :

FIGURE 4.4 – Commande d'affichage

FIGURE 4.5 – Affichage des wav.scp

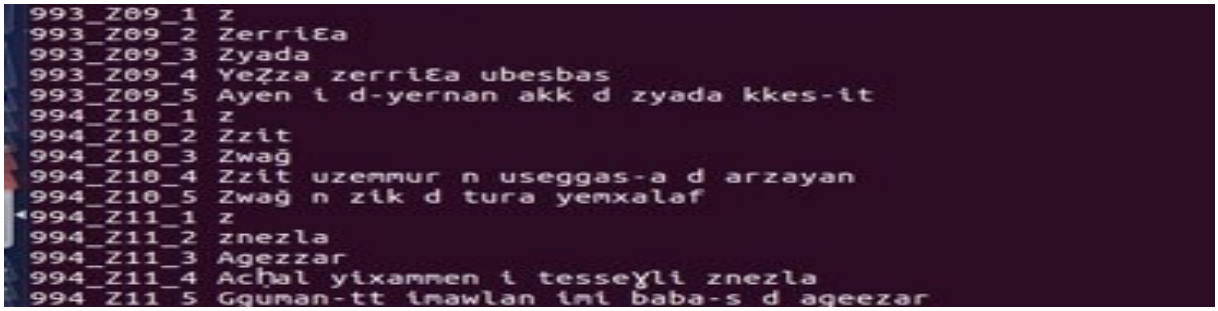

FIGURE 4.6 – Affichage des text

Conversion de la base de données

Après la vérification des formats des enregistrements de notre base de données avec kaldi on a trouvé des enregistrements stéreo, mono et multicanals. Alors que kaldi accepte seulement le format mono donc on a converti notre base de données vers le mono avec un bash(script) qu'on a crée dans kaldi.

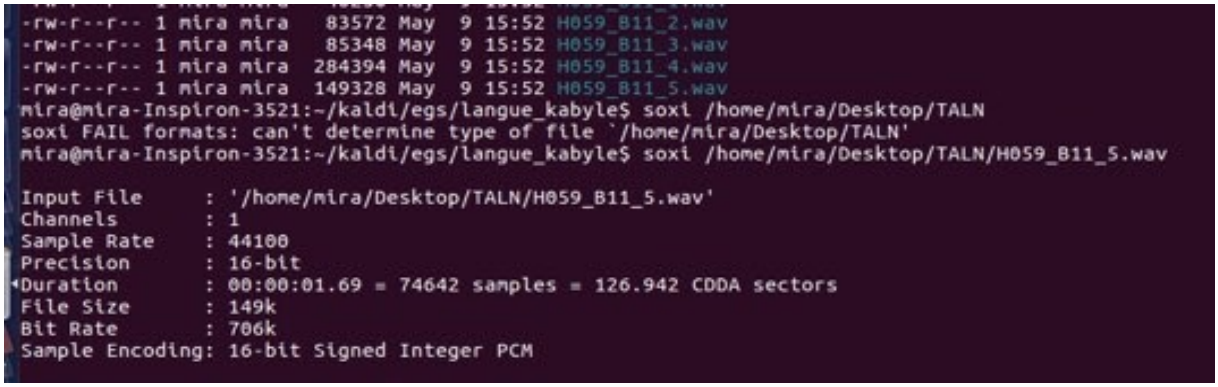

FIGURE 4.7 – Affichage de formats des enregistrements

Correction de base de données

On a crée un bash pour corriger l'extension (.wav) qui est genérer en majuscule et en miniscule lors la création de la base de données et reconnu comme erreur par kaldi.

4.4.3 Extraction des caractéristiques

Une fois la base de données est constituée, on doit extraire des caractéristiques acoustiques du signal de la parole en utilisant MFCC (Mel Frequency Cepstral Coefficients).

Paramétres de configuration :

Tout d'abord on va créer un fichier **conf** dans notre projet qui contient les paramétres de MFCC qui convient notre modèle.

-sample-frequency=44100 la fréquence d'échantillonnage des audio. -num-ceps=13 Le nombre de coefficients cepstraux a extraire. -snip-edges=false si false, eviter de tronquer le fenêtres. -use-energy=false si false,l'énergie n'est l'un caractéristiques. -window-type=hamming Minimiser le bruit.

Application MFCC

Pour calculer le MFCC, on utilise la commande suivante : 4.8
Les résultats sont stockés dans un fichier matriciel (.ark) créé automatiquement après la fin des calculs.4.9.

FIGURE 4.8 – Calculer le MFCC

```
mira@mira-Inspiron-3521:~/kaldi/egs/mon_projet_kabyle$ ls
conf  kaldi_data  mfcc  path.sh  steps
mira@mira-Inspiron-3521:~/kaldi/egs/mon_projet_kabyle$ cd mfcc
mira@mira-Inspiron-3521:~/kaldi/egs/mon_projet_kabyle/mfcc$ ls
mfcc.ark  mfcc.scp
mira@mira-Inspiron-3521:~/kaldi/egs/mon_projet_kabyle/mfcc$
```

FIGURE 4.9 – Résultats de MFCC

La normalisation

Pour appliquer la normalisation, on utilise les commandes suivantes :

```
mira@mira-Inspiron-3521:~/kaldi/egs/mon_projet_kabyle$ compute-cmvn-stats scp:mfcc/mfcc.scp ark,scp:mfcc/mfcc_cmvn.ark,mfcc/mfcc_cmvn.scp
compute-cmvn-stats scp:mfcc/mfcc.scp ark,scp:mfcc/mfcc_cmvn.ark,mfcc/mfcc_cmvn.scp
LOG (compute-cmvn-stats[5.5.1126~1-8c451]:main():compute-cmvn-stats.cc:171) Done accumulating CMVN stats for 0 utterances; 0 had errors.
mira@mira-Inspiron-3521:~/kaldi/egs/mon_projet_kabyle$ cd mfcc
mira@mira-Inspiron-3521:~/kaldi/egs/mon_projet_kabyle/mfcc$ ls
cmvn.ark  mfcc.ark  mfcc.scp  mfcc_cmvn.ark  mfcc_cmvn.scp
mira@mira-Inspiron-3521:~/kaldi/egs/mon_projet_kabyle/mfcc$ cd ..
mira@mira-Inspiron-3521:~/kaldi/egs/mon_projet_kabyle$ apply-cmvn --norm-means=true --norm-vars=true scp:mfcc/mfcc_cmvn.scp scp:mfcc/mfcc.scp
ark:- | copy-feats ark:- ark,scp:mfcc/norm_mfcc_cmvn.ark,mfcc/norm_mfcc_cmvn.scp
copy-feats ark:- ark,scp:mfcc/norm_mfcc_cmvn.ark,mfcc/norm_mfcc_cmvn.scp
apply-cmvn --norm-means=true --norm-vars=true scp:mfcc/mfcc_cmvn.scp scp:mfcc/mfcc.scp ark:-
LOG (apply-cmvn[5.5.1126~1-8c451]:main():apply-cmvn.cc:159) Applied cepstral mean and variance normalization to 0 utterances, errors on 0
LOG (copy-feats[5.5.1126~1-8c451]:main():copy-feats.cc:143) Copied 0 feature matrices.
mira@mira-Inspiron-3521:~/kaldi/egs/mon_projet_kabyle$ cd mfcc
mira@mira-Inspiron-3521:~/kaldi/egs/mon_projet_kabyle/mfcc$ ls
cmvn.ark  mfcc.ark  mfcc.scp  mfcc_cmvn.ark  mfcc_cmvn.scp  norm_mfcc_cmvn.ark  norm_mfcc_cmvn.scp
mira@mira-Inspiron-3521:~/kaldi/egs/mon_projet_kabyle/mfcc$
```

FIGURE 4.10 – La normalisation

4.4.4 Création de Modèle acoustique

Préparation des données linguistiques

Cette étape est cruciale avant l'entraînement monophone et l'alignement, pour bien garantir correctement la correspondance entre la transcription et les données audio, elle consiste à créer des dictionnaires lexicaux, grammaticaux, surtout phonètiques.
Dans notre cas, on a crée dans le fichier 'mon-data' un dossier **dict** qui contient les dictionnaires suivants :

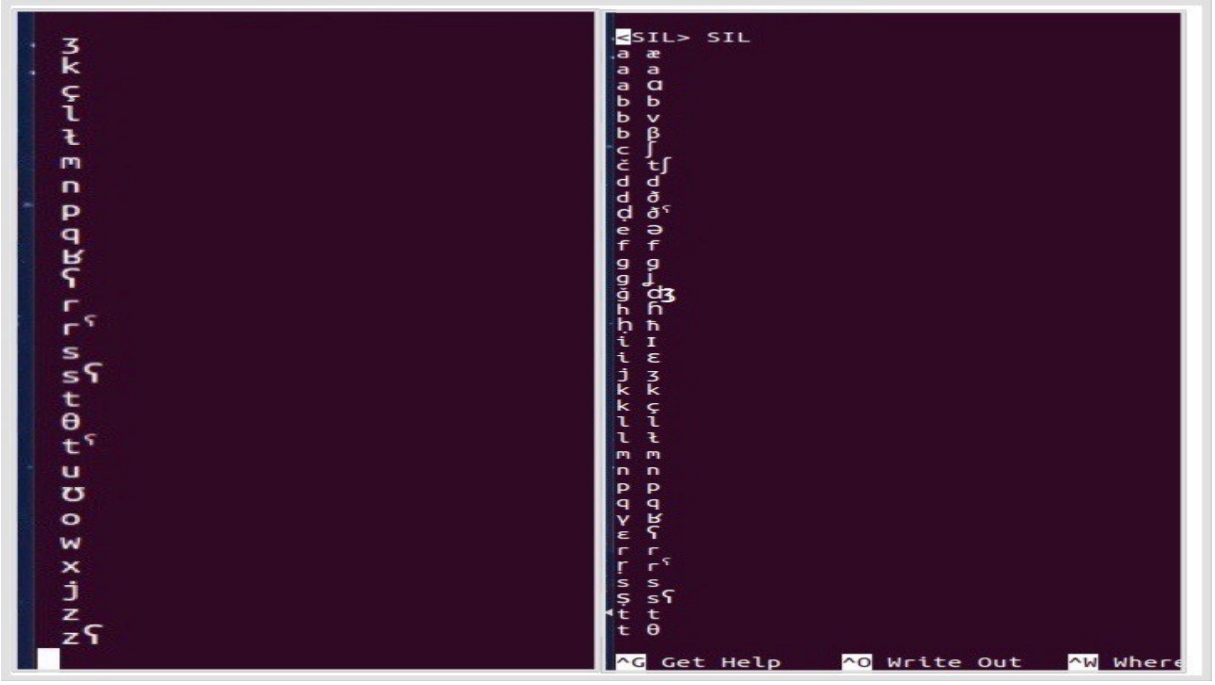

FIGURE 4.11 – Partie des dictionnaires crées

— **lexicon.txt :** Contient les lettres et les phonèmes 4.11.
— **silence-phones.txt :** Comme nous n'avons pas de lettres silencieuses, ce fichier est vide et contient uniquement 'SIL'.
— **nonsilence-phones.txt :** Contient uniquement les phonèmes 4.11.
— **lexiconp.txt :** Ce dictionnaire est initialement vide et nous le remplissons avec 'SIL'.

Après l'exécution, le système a géneré d'autres dictionnaires automatiquement à partir de ceux qu'on a créé 4.14.

```
mira@mira-Inspiron-3521:~/kaldi/egs/mon_projet_kabyle/mon_data$ cd ..
mira@mira-Inspiron-3521:~/kaldi/egs/mon_projet_kabyle$ utils/prepare_lang.sh  mon_data/dict "<OOV>" mon_data/local/lang mon_data/lang
utils/prepare_lang.sh mon_data/dict <OOV> mon_data/local/lang mon_data/lang
Checking mon_data/dict/silence_phones.txt ...
--> reading mon_data/dict/silence_phones.txt
--> text seems to be UTF-8 or ASCII, checking whitespaces
--> text contains only allowed whitespaces
--> mon_data/dict/silence_phones.txt is OK

Checking mon_data/dict/optional_silence.txt ...
--> reading mon_data/dict/optional_silence.txt
--> text seems to be UTF-8 or ASCII, checking whitespaces
--> text contains only allowed whitespaces
--> mon_data/dict/optional_silence.txt is OK

Checking mon_data/dict/nonsilence_phones.txt ...
--> reading mon_data/dict/nonsilence_phones.txt
--> text seems to be UTF-8 or ASCII, checking whitespaces
--> text contains only allowed whitespaces
--> mon_data/dict/nonsilence_phones.txt is OK

Checking disjoint: silence_phones.txt, nonsilence_phones.txt
--> disjoint property is OK.
```

FIGURE 4.12 – Préparation linguistique partie 1

```
Checking mon_data/dict/lexicon.txt
--> reading mon_data/dict/lexicon.txt
--> text seems to be UTF-8 or ASCII, checking whitespaces
--> text contains only allowed whitespaces
--> mon_data/dict/lexicon.txt is OK

Checking mon_data/dict/lexiconp.txt
--> reading mon_data/dict/lexiconp.txt
--> text seems to be UTF-8 or ASCII, checking whitespaces
--> text contains only allowed whitespaces
--> mon_data/dict/lexiconp.txt is OK

Checking lexicon pair mon_data/dict/lexicon.txt and mon_data/dict/lexiconp.txt
```

FIGURE 4.13 – Préparation linguistique partie 2

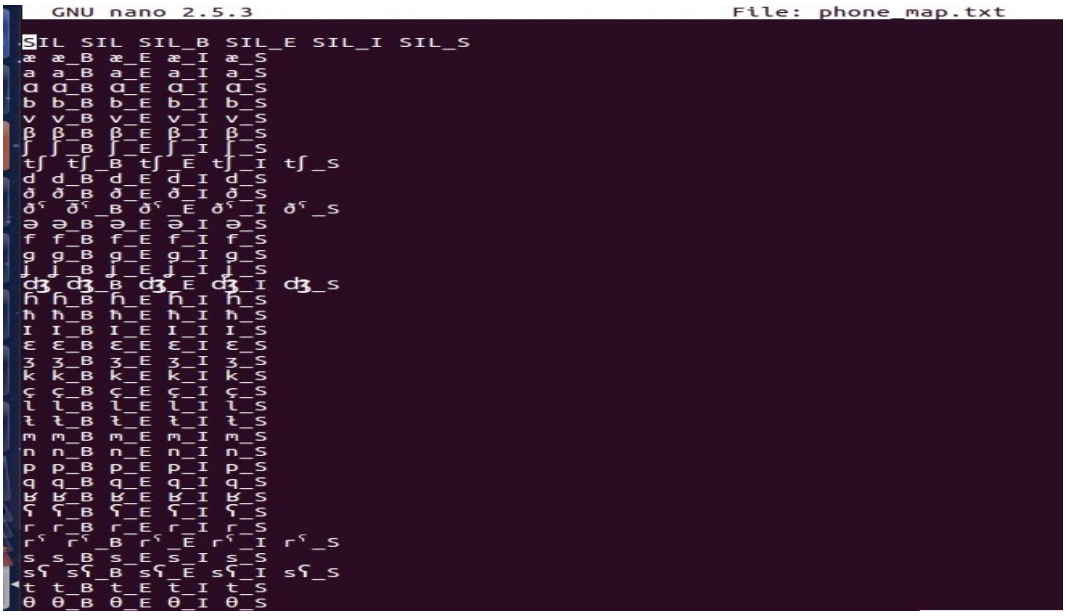

FIGURE 4.14 – Dictionnaire géneré aprés la préparation linguistique

Création des locuteurs fictifs(utt2spk)

La création de fichiers utt2spk fictifs (Utterance to speaker) consiste à créer un fichier qui établit la correspondance entre les identifiants des enregistrements(utterance) et les identifiants des locuteurs (speakers).
Ce fichier peut être généré manuellement ou bien automatiquement par Kaldi, comme on a plus de 9000 vocaux donc on a utilisé cette commande pour le créer automatiquement 4.15.

FIGURE 4.15 – Création des locuteur fictifs

Entraînement monophone et Alignement

Pour appliquer l'entraînement et l'alignement monophone on utilise les commandes suivantes, malheuresement à cette étape on a eu le problème de reconnaitre le fichier MFCC comme fichier vide.

Entraînement monophone
mira@mira-Inspiron-3521 :/kaldi/egs/langue kabyle steps/train mono.sh –cmd ”run.pl” –nj 4 data/train data/lang exp/mono

Alignement
mira@mira-Inspiron-3521 :/kaldi/egs/langue kabyle align si . sh – nj nj – $-cmd$”train cmd ” data / train data / lang $\exp_d ir/mono$

Entraînement Triphone et Alignement

Pour passer a l’entraînement et l’alignement triphone on utilise les commandes suivantes :

entraînement triphone
mira@mira-Inspiron-3521 :/kaldi/egs/langue kabyle steps / train deltas . sh – cmd ” train cmd ” numLeavesTri1 numGaussTri1 data/train data/lang exp dir / mono ali exp dir / tri1

alignement
steps/ align si . sh – nj nj – cmd ” train cmd ” – use - graphs true data/ train data/ lang exp dir/ tri1 exp dir/ tri1 ali

Entraînement de Modèle DNN

4.4.5 Création de Modèle de language

La commande utilisée pour la Création de Modèle de language est la suivante :

utils/prepare lang.sh data/local/dict ”UNK” data/ local/lang data/lang ngram-count-order lm order-write- vocab local/tmp/vocab-full.txt-wbdiscount-text data/text .txt-lm local/tmp/lm.arpa arpa2fst–disambig-symbol =0 – read-symbol-table=lang /words.txt local/tmp/ lm.arpa lang/G.fst

4.4.6 Décodage

Pour appliquer le décodage on utilise la commande suivante :

steps/nnet3/decode.sh – nj 4 cmd run.pl exp/nnet3/tdnn sp/graph data/test exp/nnet3/tdnn sp/decode test

4.4.7 Évaluation

Pour l’évaluation de modèle on utilise la commande suivante :

```
steps/score kaldi.sh data/test exp/nnet3/tdnn sp/graph exp/nnet3/tdnn sp/decode test
```

Conclusion

Ce chapitre nous a offert une introduction complète aux plateformes de reconnaissance vocale, avec un accent particulier sur Kaldi et son utilisation dans le cadre de ce projet.

Conclusion générale

Dans le cadre de ce travail, on a exploré la reconnaissance automatique de la parole qui est un domaine de recherche necéssaire pour la communication homme-machine, la langue Amazighe (Kabyle) connue par sa richesse et sa diversité linguistiques et les platformes de reconnaissances vocale principalement Kaldi.

Notre objectif était de combiner ces trois aspects pour créer un système de reconnaissance vocale pour la langue Kabyle en se basant sur une étape cruciale qui était la création d'une base données qui contient des enregistrements (phonémes, mots et phrases), des transcriptions et des annotations précises.
Cela a impliqué plusieurs étapes importantes, à commencer par la première étape cruciale de ce processus, qui est la collecte des données, qui doit être effectuée avec le plus grand soin pour garantir la qualité et la représentativité des enregistrements.

la création du dataset a débuté par la préparation minutieuse des phonèmes et des phrases en kabyle afin de capturer la richesse et la diversité linguistique de la langue. Cette approche méthodologique a assuré la validité des résultats et a ouvert la voie à des analyses approfondies.
Des équipes, équipées de dictaphones numériques de haute qualité, ont recueilli des données claires et précises, qui ont ensuite été traitées avec Audacity, nommées et organisées dans un fichier CSV pour faciliter les analyses ultérieures.

Ce projet a été une initiative complexe et instructive, l'installation et la configuration de kaldi ont constitué les premiers défis majeurs, nécessitant l'installation de librairies spécifiques, une configuration détaillée, et des mise a jour fréquentes pour assurer la stabilité de systéme. Ces étapes préliminaires ont été cruciales et ont pris un temps considérable, mais elles ont été essentielles pour préparer le terrain pour les phases de développement ultérieurs.

Notre travail a atteint une étape critique c'est l'entraînement monophone à cause des problémes techniques avec les données MFCC, qui étaient correctement générées mais identifiées comme vide par le systéme. Cependant, plusieurs obstacles ont été surmonté avec succés tels que la gestion des caractéres spéciaux de l'alphabet berbére (Kabyle), la conversion de la base de données en format mono qui est une format spécifique pour kaldi corrections des chemins dans les fichiers CSV et la normalisation des fichiers audio.

Ces obstacles, bien que surmonté, ont souligné la complexité inhérente à la reconnaisance de la parole dans une langue moins présentée technologiquement, mettant en

lumière l'importance de la persévérance et de l'innovation dans le domaine de la linguistique computationnelle.

Enfin,le progrés et les accomplissements réalisés contribuent de maniére significative à la recherche en reconnaissance de la parole pour la langue Kabyle et ouvrent la vois a des futures recherches et améliorations.

Bibliographie

[1] https ://kaldi-asr.org/doc/dnn.html.

[2] ADJED, F. Vers une normalisation du kabyle : Alphabet.

[3] A.HACINE-GHARBI. *Sélection de paramètres acoustiques pertinents pour la reconnaissance de la parole.* PhD thesis, Université d'Orléan ,Université Ferhat Abbas-Sétif ., 2012.

[4] A.SOUADKIA. Reconnaissance automatique de la parole arabe : Approch évolutionniste. Mémoire de master, Université de Guelma, 2010.

[5] BAHI, H., AND FARIHIA, H. *Etude comparative entre les librairies de reconnaissance vocale*. PhD thesis, Université Badji Mokhar (UBMA), Annaba, Algérie., 2014.

[6] BAUDE, O. Article : Langues et cité.

[7] BEN, Y. O. *Livre : cours de machine learning*.

[8] BENAMMAR, R. *Traitement Automatique De La Parole Arabe Par Les HMMs*. PhD thesis, Université de Tlemcen., 2012.

[9] BENRAMDANE, M. K. Reconnaissance vocale basée sur le deep learning appliquée à la langue kabyle. Mémoire de master, École nationale Supérieure d'Informatique., 2020.

[10] CHAKER, S. Livre sur le berbere de kabylie. *Encyclopédie Berbère* (2004), 4055–4066p.

[11] COTSAFTIS, C. Mise en oeuvre de techniques d'intelligence artificielle pour la reconnaissance vocale. Mémoire de master, Uensta bretagne, 2020.

[12] DRYGAJLO, D. A. *TRAITEMENT DE LA PAROLE*. PhD thesis, Institut de Traitement des Signaux,Lausane, 2003.

[13] F.BARKANI, M.HAMIDI, N. O. H., AND K.SATORI. Article : Amazigh speech recognition based on the kaldi asr toolkit. *International Journal of Information Technology* (2023), pages 3533–3540.

[14] GELIN, L. *Reconnaissance automatique de la parole d'enfant apprennat lecture en salle de classe*. PhD thesis, Universite de Toulouse 03 Paul sebastier, Fevrier 2022.

[15] GUGLANI, JYOTI, M. A. Automatic speech recognition system with pitch dependent features for punjabi language on kaldi toolkit. *Applied Acoustics 167* (2020), 107386.

[16] HAMMADECHE, A. H., AND TAKI, M. *Reconnaissance automatique de la parole arabe continue.* PhD thesis, Université Saad Dahleb, Belida 1, 2019.

[17] HAMZA, F. *Approche Hybride pour la Reconnaissance de la Parole*. PhD thesis, Université Badji Mokhar (UBMA), Annaba, Algérie., 2018.

[18] HEBA, A. *Reconnaissance automatique de la parole à large vocabulaire : des approches hybrides aux approches End-to-End*. PhD thesis, Université SAAD DAHLAB de BLIDA., 2021.

[19] IMRANI, N. E., AND AHADRI, O. *L'impact de L'intelligence artificielle sur le monde de travail*.

[20] JUANG, B., AND RABINER, L. R. *Livre sur Automatic Speech Recognition – A Brief History of the Technology Development*. 08 2004.

[21] K.AMAR, AND N.HAMMOU. Reconnaissance vocale du genre basée sur l'apprentissage profond. Mémoire de master, UNIVERSITE ABDELHAMID IBN BADIS - MOSTAGANEM, 2023.

[22] LE, V. B. *Reconnaissance automatique de la parole pour des langues peu dotées.* PhD thesis, UNIVERSITÉ JOSEPH FOURIER - GRENOBLE 1 ., 2006.

[23] LIN, S. S. Optimisation du graphe de décodage d'un système de reconnaissance vocale par apprentissage discriminant. Memoire master, 2007.

[24] MENACER, M. A. *Doctorat de l'Université de Lorraine (mention informatique)*. PhD thesis, Université de Lorraine, November 2020.

[25] MEZAGHANI, K., AMMAR, O., AND ALLAH, A. Automatic speech recognition for amazigh. *Langues et cité* (2019).

[26] NAIT-ZERRAD, K. *Livre sur Esquisse d'une classification linguistique des parlers berbères*.

[27] N.ZERARI. *INTÉGRATION D'UN MODULE DE RECONNAISSANCE DE LA PAROLE AU NIVEAU D'UN SYSTÈME AUDIOVISUEL - APPLICATION TÉLÉVISEUR.* PhD thesis, Université Batna 2 – Mostefa Ben Boulaïd., 2021.

[28] OUAHABI, S. S., AND ATOUNT, M. Conference sur comparative study of amazigh speech recognition systems based on different toolkits and approaches. In *E3S Web of Conferences* (2023).

[29] OUALID, D. Reconnaissance automatique de la parole arabe par cmu sphinx 4. Mémoire de master, UNIVERSITÉ FERHAT ABBAS – SÉTIF 1.

[30] POVEY, D., AND GHOSHAL, A. The kaldi speech recognition toolkit. In *Livre sur IEEE 2011 workshop on automatic speech recognition and understanding* (2011), IEEE Signal Processing Society.

[31] RAMDANI, R. Commande vocale d'une plateforme mobile. Mémoire de master, Université SAAD DAHLAB de BLIDA, 2017.

[32] SAHKI, H. *Livre sur LA LANGUE BERBERE*, vol. 138p. Decembre 1998.

[33] SATORI, H., AND ELHAOUSSI, F. Article :investigation amazigh speech recognition using cmu tools. *International Journal of Speech Technology* (2014).

[34] S.GAGNON. Modèles de markov cachés à haute précision dynamique. Mémoire de master, UNIVERSITÉ DE SHERBROOKE, 2016.

[35] SOLTANI, I. *L'usage de la langue maternelle dans l'enseignement/apprentissage du FLE "cas des élèves de 2ème année moyenne" C.E.M Ghassira Batna.* PhD thesis, Université Mohamed Khider de Biskra ., 2019.

[36] SORAND, C. *Livre sur Berbérité : En quête de l'identité amazighe de la Préhistoire à la chute de Carthage*. Le Lys Bleu Éditions, 2023.

Résumé

Dans le cadre de ce travail, notre objectif était de développer un système de reconnaisssance de la parole pour la langue kabyle, en utilisant la platforme kaldi et en se basant sur une étape cruciale qui est la création d'une base des données. Cette approche méthodologique est l'étape la plus difficile et importante pour assurer la validité des résultats et ouvre la voie à des analyses approfondies de la richesse linguistique.
L'installation et la configuration initiales de kaldi ont été particulièrement exigantes, impliquant l'intégration de librairies spécifiques et des mises a jour continue pour assurer la fonctionnalité du système.
Malgré les progrès significatifs et les techniques utilisées pour assurer une representstivité de tout le système phonétique berbère (kabyle), on a rencontré un obstacle majeur lors de l'entrainement monophone. Les données MFCC, bien que correctement générées.
Les défis rencontrés et les solutions apportées ont mis en lumières la complexité de la reconnaissance de la parole pour une langue moins représentée ainsi que le manque flagrand de data-set, nous soulignons l'importance de création de data-set et de corpus spécialisés pour la langue Amazighe. Les fondations établies et les connaissances acquises constituent une base solide pour les recherches futures et l'amélioration de la reconnaissance de la parole.

Mots clés : Intelligence artificielle, Apprentissage automatique, Reconnaissance automatique de parole(RAP), Langue Kabyle, Corpus vocale, Apprentissage profond, Réseaux de Neurones Récurrents.

Abstract

As part of this work, our objective was to develop a speech recognition system for the Kabyle language, using the Kaldi platform and based on a crucial step which is the creation of a database. This methodological approach is the most difficult and important step to ensure the validity of the results and opens the way to in-depth analyzes of linguistic richness.
The initial installation and configuration of kaldi was particularly demanding, involving the integration of specific libraries and continuous updates to ensure the functionality of the system.
Despite the significant progress and the techniques used to ensure representativeness of the entire Berber (Kabyle) phonetic system, we encountered a major obstacle during monophone training. MFCC data, although correctly generated.
The challenges encountered and the solutions provided have highlighted the complexity of speech recognition for a less represented language as well as the blatant lack of data sets, we underline the importance of creating data sets and specialized corpora for the Amazigh language. The foundations established and the knowledge gained provide a solid basis for future research and improvement of speech recognition.

Keywords : Artificial Intelligence, Machine learning, Automatic system recognition(ASR), Kabyle language, deep learning, Vocal corpus, Recurrent Neural Networks.