الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne démocratique et populaire

وزارة التعليم العالي والبحث العلمي
Ministère de l'enseignement supérieur et de la recherche scientifique

جامعة سعد دحلب البليدة
Université SAAD DAHLAB de BLIDA

كلية التكنولوجيا
Faculté de Technologie

قسم الإلكترونيك
Département d'Électronique

Mémoire de Master

Référence PFE : ST4

Filière : Télécommunication

Spécialité : Systèmes des télécommunications

Présenté par

Hamchaoui Rayane

&

Youssari Melissa

Système de reconnaissance du locuteur basé sur les réseaux RNN et GTCC

Proposé par :

Dr BOUTALEB Nassima

Année universitaire : 2024/2025

Dédicaces

Avec un énorme plaisir, un cœur ouvert et une immense joie que je dédie ce travail :

À ceux qui sont ma lumière, ma force et mon refuge…

À ma famille,

À mes parents Chahra et MOH piliers de mon existence,
Merci pour votre amour pur, vos sacrifices silencieux et vos prières constantes. Vous avez semé en moi les graines de la persévérance et de l'espérance. Ce mémoire est l'aboutissement de vos efforts autant que des miens.

À mon frére Anis : *Merci pour ton affection, ta discrétion, ton soutien tranquille mais toujours présent.*

À mes petites nièces, Ania , Melina : *Vous apportez toujours de la joie dans ma vie. Je vous aime très fort.*

À mes sœurs Riheb , Rachel
Ceux qui m'ont soutenu, écouté, conseillé, relevé dans les moments de doute. Merci pour votre amitié sincère, vos encouragements, et vos mots justes qui ont souvent remplacé le silence pesant de la fatigue.

À ma meilleure amie Ahlem

Merci d'être cette présence rare qui apaise, qui comprend, qui soutient. Ton amitié est un trésor dans ma vie, et ton amour une douceur qui me porte.

À Rayane, ma sœur d'une autre mère et ma meilleure amie
Merci d'être toujours là, dans les bons comme dans les moments difficiles. Ta présence et ton soutien ont tout rendu plus facile et plus beau.

À la famille de Rayane
Merci d'avoir élevé une personne aussi exceptionnelle. Rayane est non seulement une amie précieuse, mais aussi un vrai soutien dans ma vie. Je vous suis reconnaissante pour tout ce qu'elle est.

À tous ceux qui ont cru en moi quand moi-même j'hésitais,
Ce travail vous est dédié, avec une gratitude que les mots ne suffisent pas à exprimer.

MELISSA

Dédicaces

Tout au long de ce parcours, vers l'aboutissement de ce chapitre si important de ma vie, certaines personnes ont joué un rôle essentiel — à vous, je dédie ce travail.

À ma mère et à mon père, pour leur amour inépuisable, leur confiance absolue et leur présence constante, dans les joies comme dans les tempêtes. Rien n'aurait été possible sans vous.

À mon frère Ryadh, même loin, tu restes ce modèle de sagesse et de force tranquille sur lequel je peux toujours compter.

À ma petite sœur Razane, pour ton soutien instinctif, ton humour, ta tendresse — tu sais toujours être là au bon moment, sans même que je le demande.

À mes sœurs de cœur : Riheb, Fayrouz, Rachel et Narimen, vos présences lumineuses, vos conseils spontanés et vos messages de minuit ont donné du sens à bien plus que ce mémoire.

À ma famille dans son ensemble, votre bienveillance spontanée, vos petits gestes et votre présence sont un cadeau que je chéris. Ainsi, qu'à ma tante Leila, ma seconde maman. Toujours là, sans jamais attendre qu'on lui demande, avec une attention douce, des mots justes et une présence réconfortante. Ton amour et générosité ont compté plus que tu ne le crois.

À Melissa particulièrement, ma meilleure amie depuis la maternelle, ma confidente de toujours, ma binôme dans cette aventure et tant d'autres. Ton soutien, ta loyauté et ta force m'ont accompagné depuis l'enfance. Ce mémoire est autant le tien que le mien. Merci d'être restée à mes côtés, hier, aujourd'hui, et pour longtemps encore.

Et enfin, à moi-même. Pour avoir tenu bon malgré les doutes, pour avoir continué même quand l'énergie manquait, pour avoir cru en ce rêve. Ce chapitre est une victoire, et je me la dois aussi.

RAYANE

Remerciements

Nous voudrons commencer par remercier ALLAH le tout puissant de nous avoir donné la foiet de nous avoir permis d'en arriver là.

Nous exprimons notre parfaite reconnaissance et nos sincères remerciement au Dr. N. BOUTALEB pour avoir dirigé ce travail avec rigueur et bienveillance. Sa disponibilité, la qualité de ses conseils, son accompagnement attentif et son soutien constant ont grandement contribué à la réalisation de ce mémoire. Nous lui sommes profondément reconnaissantes pour l'intérêt qu'elle a porté à ce projet et pour l'ensemble des efforts qu'elle y a consacrés.

Nous exprimons nos sincères remerciements aux membres du jury : Madame Yahiaoui ainsi que Monsieur Habib, pour l'honneur qu'ils nous font en acceptant d'évaluer ce travail. Nous

portons notre gratitude envers leur engagement consacré l'examen de ce travail.

Enfin, merci à toutes les personnes qui ont contribué de près ou de loin à la réalisation de ceprojet

Résumé

Ce travail de fin d'étude porte sur la reconnaissance automatique du locuteur à l'aide des réseaux de neurones récurrents (RNN) combinés aux coefficients cepstraux Gammatone (GTCC) pour l'extraction des caractéristiques audio. L'objectif est de développer un système capable d'identifier efficacement un individu à partir de sa voix, en exploitant les capacités de modélisation temporelle des RNN. Le modèle a été entraîné et évalué sur un corpus vocal prétraité avec les GTCC, montrant des performances prometteuses en termes de classification des locuteurs. Cette approche ouvre la voie à des applications dans la sécurité biométrique, la personnalisation vocale et les systèmes intelligents.

Mots-clés : Reconnaissance du locuteur, RNN, GTCC, Classification vocale, Apprentissage profond.

Abstract

This final year project addresses automatic speaker recognition using Recurrent Neural Networks (RNN) combined with Gammatone Cepstral Coefficients (GTCC) for audio feature extraction. The goal is to develop a system capable of accurately identifying individuals based on their voice, by leveraging the temporal modeling strength of RNNs. The model was trained and evaluated on a vocal dataset processed with GTCC, yielding promising results in speaker classification. This approach opens opportunities for applications in biometric security, voice-based personalization, and intelligent systems.

Keywords: Speaker recognition, RNN, GTCC, Voice classification, Deep learning.

الملخص

يتناول هذا المشروع في سنته النهائية التعرف التلقائي على المتحدثين باستخدام تقنيات متقدمة لاستخراج ميزات الصوت. الهدف هو تطوير نظام قادر على التعرف بدقة على الأفراد استنادًا إلى أصواتهم، من خلال الاستفادة من القدرة على نمذجة التسلسل الزمني في البيانات الصوتية. تم تدريب النموذج وتقييمه باستخدام مجموعة بيانات صوتية معالجة، وقد أسفر ذلك عن نتائج واعدة في تصنيف المتحدثين. يفتح هذا النهج آفاقًا لتطبيقات في مجال الأمن البيومتري، والتخصيص ألصوتي والأنظمة الذكية.

الكلمات المفتاحية : التعرف على المتحدث، الشبكات العصبية التكرارية، معاملات غاماتون السِبسترالية، تصنيف الصوت، التعلم العميق.

Liste des acronymes

CNN : Convolutional Neural Network

DCT : Discrete Cosine Transform

GRU : Gated Recurrent Unit

GTCC : Gammatone Cepstral Coefficients

IA : Intelligence Artificielle

LSTM : Long Short-Term Memory

MFCC : Mel-Frequency Cepstral Coefficients

RAL : Reconnaissance Automatique de Locuteur

RAP : Reconnaissance Automatique De la Parole

RNA : Réseau Neurone Artificiel

RNN : Recurrent Neural Network

VAL : Vérification Automatique du Locuteur

MLT : Multi-Language Text

WAV : Waveform Audio File Format

FR-FA : False Rejection - False Acceptance

PLP : Perceptual Linear Prediction

Table des matières

INTRODUCTION GENERALE.....01

Chapitre 1 Fondements théoriques de la reconnaissance automatique du locuteur.....02
1.1 INTRODUCTION 02
1.2 La parole humaine 02
1.2.1 Définition 02
1.2.2 Production de parole 02
1.2.3 Identification du locuteur:.....03
1.2.4 Paramètres du signal de parole 04
1.3 La reconnaissance 05
1.4 La reconnaissance de parole 05
1.5 La reconnaissance automatique de locuteur 06
1.6 Domaine d'application de la RAL.....07
1.7 Vérification Automatique du Locuteur.....07
1.8 Les paramètres MFCC.....08
1.9 Les paramètres GTCC.....08
1.10 Comparaison entre MFCC et GTCC.....09
1.11 conclusion.....10

Chapitre 2 Réseaux RNN GTCC pour la reconnaissance du locuteur 11
2.1 INTRODUCTION 11
2.2 Intelligence artificielle 11
2.2.1 DEFINITION 11
2.3 DEEP LEARNING.....11
2.3.1 Définition.....11
2.3.2 Domaines d'applications de deep learning 11
2.4 Apprentissage automatique 12
2.5 Types d'apprentissage.....12
2.5.1 Apprentissage supervisé 12
2.5.2 Apprentissage non-supervisé 13
2.5.3 Apprentissage par renforcement.....14
2.6.Neurone biologique.....15
2.7. Réseaux de neurones artificiels (RNA) 16
2.8 Fonctionnement des réseaux de neurones artificiels.....16
2.9 Types de réseaux de neurones artificiels.....17
2.10.Réseaux de neurones récurrents.....19

2.10.1 Les types de réseaux de neurones récurrents 19
2.10.2 Architecture de RNN.....20
2.10.3 Fonctionnement des RNN.....21
1-Astuce de la fenêtre glissante et Connexions Récurrentes.....22
2-Modélisation des Dépendances Temporelles.....22
2.10.4 Comparaison entre RNN et CNN.....22
2.11 Les réseaux Gammatone cepstral coefficients GTCC.....23
2.11.1 Fonctionnement des GTCC.....23
Conclusion 24

Chapitre 3 Travaux d'expérimentations et résultats..........25
3.1 Introduction.....25
3.2 Langage python.....26
3.3. DEMARCHE METHODOLOGIQUE.....26
3.4 Collecte de la base de données.....27
3.5.Conception des modèles d'identification du locuteur.....28
3.5.1. Importation de la bibliothèque 29
3.5.2. Extraction des caractéristiques audio.....29
3.5.3. Construction du modèle RNN.....30
3.5.4. Construction du modèle CNN 31
3.5.5. Entraînement et évaluation du modèle 31
3.6. Optimisation des paramètres du modèle 33
3.7. Evaluation du modèle d'identification 35
3.7.1.Analyse des taux de perte et précision.....36
3.7.2.Analyse de la matrice de confusion et taux de test 38
3.8. Amélioration du modèle d'identification.....43
3.8.1. Analyse des taux de perte et précision 43
3.8.2. Analyse de la matrice de confusion.....46
3.8 Conclusion.....50

CONCLUSION GENERALE.....51

Bibliographie

Liste des figures

Figure 1.1. Organes de production de la parole humaine.....03

Figure1.2 : Représentation d'un spectre fréquentiel d'un signal audio.....04

Figure1.3 . Structure de base d'un système RAP.....06

Figure 1.4. Schéma d'un système RAL. 06

Figure 1.5 schéma typique d'un système de vérification du locuteur.....08

Figure 2.1 Processus de l'apprentissage machine.....12

Figure 2.2 Processus de l'apprentissage supervisé.....13

Figure 2.3 Processus de l'apprentissage non-supervisé.....14

Figure 2.4. Interaction agent-environnement.....15

Figure 2.5. Le neurone biologique.....15

Figure 2.6. Réseau de neurone artificiel.....16

Figure 2.7. Réseaux de neurones à propagation avant.....17

Figure 2.8. Réseau neuronal convolutif.....18

Figure 2.9. Architecture des réseaux de neurones récurrents.....18

Figure 2.10 (à gauche) Un RNN (à droite) Sa version déroulé Source.....19

Figure 2.11 Les types de réseaux de neurones récurrents.....20

Figure 2.12 Récurrent Neural Network (RNN) Tutorial.....21

Figure 3.1: Les domaines d'applications de python.....26

Figure 3.2 : démarche méthodologique de notre travail.....27

Figure 3.3 : Modèle d'apprentissage.....28

Figure 3.4 : Taux de précision et de perte (entrainement et validation) du modèle RNN GTCC.....33

Figure 3.5 : Taux de précision et de perte (entrainement et validation) du modèle CNN GTCC.....34

Figure 3.6 : Taux de précision et de perte (entrainement et validation) du modèle RNN GTCC.....36

Figure 3.7 : Taux de précision et de pertes de modèle RNN GTCC pour l'ensemble d'enregistrements multilingue.....37
Figure 3.8 : Taux de précision et de perte de modèle CNN pour l'ensemble d'enregistrement monolingue.....37
Figure 3.9 : Taux de précision et de perte de modèle CNN pour l'ensemble d'enregistrement multilingue.....38
Figure 3.10 : Matrice de confusion RNN GTCC de la base de données monolingue.....39
Figure 3.11 : Matrice de confusion RNN de la base de données multilingue.....40
Figure 3.12 Matrice de modèle CNN pour l'ensemble d'enregistrement monolingue.....41
Figure 3.13 : Matrice de modèle CNN pour l'ensemble d'enregistrement multilingue.....42
Figure 3.14 : Taux de précision et de perte de modèle RNN pour l'ensemble d'enregistrement monolingue augmentée.....43
Figure 3.15 : Taux de précision et de perte de modèle RNN pour l'ensemble d'enregistrement multilingue augmentée.....44
Figure 3.16 : Taux de précision et de perte de modèle CNN pour l'ensemble d'enregistrement monolingue augmentée.....44
Figure 3.17 : Taux de précision et de perte de modèle CNN pour l'ensemble d'enregistrement multilingue augmentée. 45
Figure 3.18 : Matrice de confusion RNN GTCC de la base de données monolingue augmentée.....46
Figure 3.19 : Matrice de confusion RNN GTCC de la base de données multilingue augmentée.....47
Figure 3.20 : Matrice de confusion CNN de la base de données monolingue augmentée.....48
Figure 3.21 : Matrice de confusion CNN de la base de données multilingue augmentée.....49

Liste des tableaux

Tableau 1.1 : La différence entre MFCC et GTCC…….09
Tableau 2.1 : Comparaison entre CNN et RNN……..22
Tableau 3.1 : bibliothèque python utilisée dans le modèle d'identification……..29
Tableau 3.2 : Taux de : précision de validation, Perte et test pour RNN GTCC et CNN GTCC……34
Tableau 3.3 : Taux de précision et test des modèles RNN GTCC et CNN GTCC pour les ensembles d'enregistrement monolingue et multilingue après l'augmentation des données……..45

Introduction Générale

La reconnaissance du locuteur est une branche de la reconnaissance de formes qui vise à identifier automatiquement l'identité d'une personne à partir de sa voix. Ce domaine connaît un intérêt croissant en raison de ses multiples applications, notamment en sécurité, biométrie, assistance vocale et interaction homme-machine. Au fil du temps, de nombreuses approches ont été proposées afin d'améliorer la fiabilité et la précision de ces systèmes.

Toutefois, l'avènement de l'intelligence artificielle et, plus précisément, du deep learning a ouvert de nouvelles perspectives, notamment grâce à l'utilisation de modèles capables de traiter des données complexes de manière efficace et autonome.

Dans ce contexte, les réseaux de neurones récurrents (RNN) se sont révélés particulièrement adaptés au traitement des signaux vocaux, en raison de leur capacité à modéliser des séquences temporelles. Combinés aux coefficients cepstraux Gammatone (GTCC), ces réseaux offrent une meilleure représentation des caractéristiques acoustiques de la parole.

L'objectif principal de ce mémoire est donc de concevoir et d'évaluer un système de reconnaissance automatique du locuteur basé sur les RNN et les GTCC. Pour ce faire, nous mettrons en œuvre des techniques de traitement du signal, d'extraction de caractéristiques, et d'apprentissage supervisé, tout en évaluant les performances du modèle obtenu en termes de précision, de robustesse et de capacité de généralisation.

La structure de ce mémoire s'articule comme suit :

• Le premier chapitre présente une vue générale sur la reconnaissance du locuteur. Il aborde les mécanismes de production de la parole, les tâches liées à la reconnaissance vocale, les principes de fonctionnement des systèmes, ainsi qu'une comparaison entre les paramètres GTCC et MFCC, fréquemment utilisés dans ce domaine.

• Le deuxième chapitre est consacré aux fondements de l'intelligence artificielle, il s'attarde aussi sur les réseaux de neurones artificiels, avec un accent mis sur les RNN, en expliquant leur architecture, leur fonctionnement, ainsi qu'une comparaison avec les CNN afin de justifier les choix retenus.

• Le troisième chapitre est dédié à la mise en œuvre expérimentale de notre système de reconnaissance. Il présente les environnements de développement, les outils utilisés, les étapes de traitement des données, ainsi que la collecte d'une base de données vocale en français, anglais et dialecte algérien, réalisée selon plusieurs critères pertinents pour l'identification du locuteur. Ce chapitre décrit également la conception du modèle, l'évaluation de ses performances et l'analyse des résultats obtenus.

1.1. Introduction

La reconnaissance vocale est encore en chantier et attire beaucoup de développeurs amenés par la complexité de la technologie d'analyse de la voix (aussi appelée analyse du locuteur). Cette technologie s'applique avec succès là où les autres technologies sont difficiles à employer. Elle est utilisée dans des secteurs comme les centres d'appel, les opérations bancaires, l'accès à des comptes, sur PC domestiques, pour l'accès à un réseau ou encore pour des applications judiciaires. Le traitement vocal vise donc aussi un gain de productivité puisque c'est la machine qui s'adapte à l'homme pour communiquer, et non l'inverse et c'est pour ça que la reconnaissance vocale est quasiment imparfaite dans son domaine[1].

Pour cet objectif, ce chapitre présente essentiellement une introduction au domaine de la reconnaissance automatique de la parole RAP et ses principales composantes. Aussi l'utilité des réseaux de neurones RNN pour la reconnaissance vocale et explication des GTCC .

1.2. Parole humaine

1.2.1 Définition

La parole constitue le mode de communication le plus naturel dans toute société humaine du fait que son apprentissage s'effectue dès l'enfance. La parole se définit comme étant un signal réel, continu, d'énergie finie et non stationnaire, généré par l'appareil vocal humain [2].

1.2.2 Production de la parole

Les différences de voix entre les individus sont dues à la structure de leurs organes articulatoires, tels que la longueur du tractus vocal, les caractéristiques des cordes vocales et les variations dans leurs habitudes de parole. Chez un adulte, le tractus vocal mesure généralement environ 17 cm et fait partie des organes impliqués dans la production de la parole, situés au-dessus des plis vocaux (Anciennement appelés cordes vocales). Comme illustré dans la Figure 1.1, ces organes comprennent le pharynx laryngé (Situé sous l'épiglotte), le pharynx oral (derrière la langue, entre l'épiglotte et le voile du palais), la cavité buccale (en avant du voile du palais et délimitée par les lèvres, la langue et le palais), le pharynx nasal (Au-dessus du voile du palais, à l'extrémité arrière des cavités nasales) et la cavité nasale (Au-dessus du palais, s'étendant du pharynx aux narines). Le larynx est composé des plis vocaux, de la partie supérieure du cartilage cricoïde, des cartilages aryténoïdes et du cartilage thyroïde. La région située entre les plis vocaux est appelée la glotte[3].

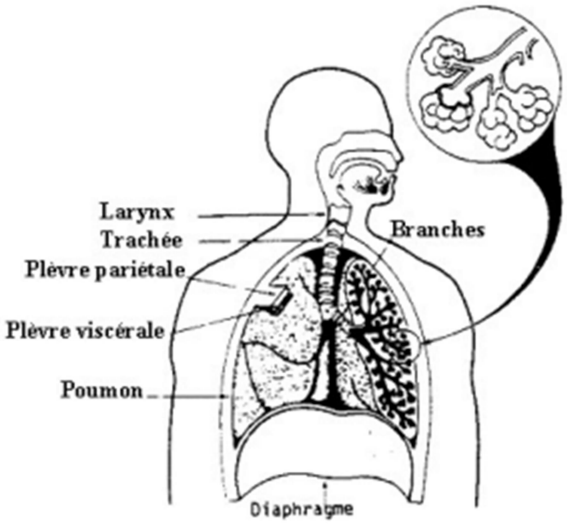

Figure 1.1. Organes de production de la parole humaine [4].

La source d'excitation de la voix humaine contient également des informations spécifiques à chaque locuteur. L'excitation est générée par le flux d'air provenant des poumons, qui passe ensuite par la trachée puis par les plis vocaux. L'excitation est classée en phonation, chuchotement, friction, compression, vibration ou une combinaison de ces éléments. L'excitation de la phonation se produit lorsque le flux d'air est modulé par les plis vocaux. Lorsque les plis vocaux se ferment, la pression s'accumule en dessous jusqu'à ce qu'ils se séparent. Les plis sont ensuite ramenés ensemble par leur tension, leur élasticité et l'effet Bernoulli. L'oscillation des plis vocaux provoque une excitation pulsée du tractus vocal. La fréquence d'oscillation est appelée fréquence fondamentale et elle dépend de la longueur, de la masse et de la tension des plis vocaux. La fréquence fondamentale est donc une autre caractéristique distinctive pour un locuteur donné[3].

1.2.3 Identification du locuteur

L'identification du locuteur consiste à déterminer l'identité d'un individu parmi une population de personnes connues. À partir d'un échantillon d'une voix enregistrée, on cherche à déterminer quel locuteur de la base de données a parlé. Pour ce faire, les données de la base sont comparées à une référence caractéristique de chaque utilisateur connu du système. Le résultat de chaque comparaison est un score, fonction de la similarité observée par le système entre les données du locuteur et la référence considérée. Le score le plus élevé correspond à la référence la plus proche des données de test et l'identité du locuteur correspondant à cette référence est renvoyée par le système[5].

1.2.4 Paramètres du signal de parole

Un processus naturel, variable dans le temps qui peut être directement représenté sous la forme de signal analogique. Ce dernier est un vecteur acoustique porteur d'informations d'une grande complexité, variabilité et redondance.

Analyser un tel signal est une tâche difficile vu le grand nombre de paramètres associés. Néanmoins, trois principaux paramètres s'imposent : la fréquence fondamentale, le spectre fréquentiel et l'énergie. Ces paramètres sont appelés traits acoustiques et sont énumérés ci-après[6,7].

1.2.4.1 Fréquence fondamentale (F_0)

C'est une caractéristique acoustique propre à chaque personne. Elle est fonction de plusieurs paramètres physiologiques tels que le volume de la glotte et la longueur de la trachée. Elle se définit par la cadence du cycle d'ouverture et de fermeture des cordes vocales pendant la phonation des sons voisés. La fréquence fondamentale varie d'un locuteur à un autre selon le genre et l'âge comme suit [8] :

• de 80 Hz à 200 Hz pour une voix d'homme.

• de 150 Hz à 450 Hz pour une voix de femme.

• de 200 Hz à 600 Hz pour une voix d'enfant

1.2.4.2 Spectre fréquentiel

C'est la représentation d'un signal dans le domaine fréquentiel (ensemble de fréquences en progression arithmétique). Une importante caractéristique permettant l'identification de tout locuteur par sa voix nommée timbre.

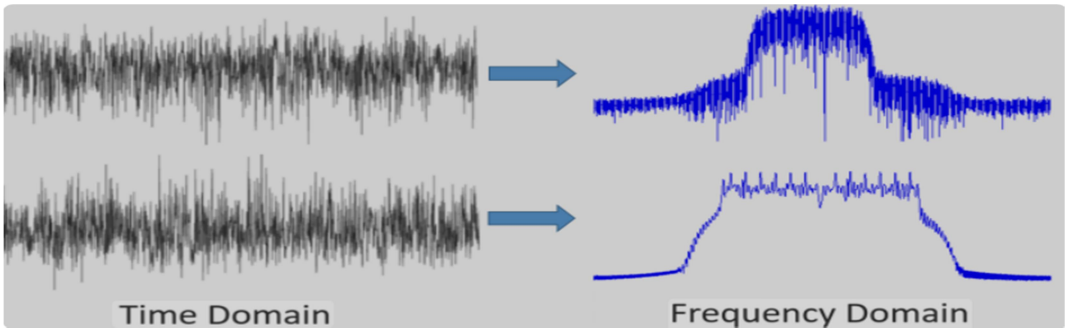

Figure1.2 : Représentation d'un spectre fréquentiel d'un signal audio

1.2.4.3 Energie

Elle est correspond à l'intensité sonore. Elle est généralement plus puissante pour les segments voisés de la parole que pour les segments non voisés.

1.2.4.4 Amplitude

L'amplitude est une mesure de son changement dans une seule période (telle que le temps ou la période spatial). L'amplitude d'un signal non périodique est son amplitude comparée à une valeur de référence. Il existe différentes définitions de L'amplitude qui sont toutes fonctions de L'amplitude des différences entre les valeurs extrêmes de la variable. Dans les textes plus enceins, la phase d'une fonction périodique est parfois appelée l'amplitude[9].

1.2.4.5 Spectrogramme

Le spectrogramme est un diagramme associant à chaque instant t d'un signal, son spectre de fréquence. Les spectrogrammes sont utilisés pour identifier des sons, comme des cris d'animaux et des sons d'instruments musicaux. Ils sont largement utilisés dans le domaine de la reconnaissance de la parole. On peut dire qu'Un spectrogramme affiche la force d'un signal dans le temps aux différentes fréquences d'une forme d'onde. Les spectrogrammes peuvent être des graphiques bidimensionnels avec une troisième variable représentée par des couleurs ou des graphiques tridimensionnels avec une quatrième variable de couleur.

1.3 La reconnaissance

En termes plus généraux, la reconnaissance fait référence à la capacité de reconnaître, de comprendre ou de distinguer quelque chose. Cependant, la reconnaissance peut être utilisée dans différents domaines et a des significations spécifiques.

1.4 La reconnaissance de parole

La reconnaissance automatique de la parole est une technologie informatique qui permet à un logiciel d'interpréter la parole humaine naturelle. Cela Permet à la machine d'extraire le message verbal contenu dans le signal vocal et de l'analyser ce signal en une chaîne de mots ou phonèmes représentant ce que la personne a prononcé. Cette technologie utilise des méthodes informatiques dans le domaine du traitement du signal et de l'intelligence artificielle [10].

Les systèmes de reconnaissance de parole se sont considérablement améliorés ces dernières années grâce aux progrès de l'apprentissage automatique, en particulier du deep learning, qui permettent une meilleure compréhension et interprétation de la parole humaine, même dans des environnements bruyants ou avec des accents différents. Cependant, ils ne sont pas encore parfaitement précis et peuvent avoir des difficultés avec certains mots ou accents particuliers, ce qui nécessite encore une relecture ou

une correction manuelle dans certains cas. Le schéma suivant résume les étapes de base d'un système RAP :

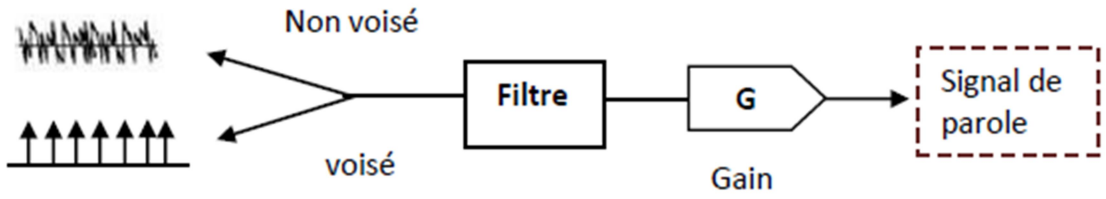

Figure1.3 . Structure de base d'un système RAP

1.5 La reconnaissance automatique de locuteur

La reconnaissance automatique du locuteur est interprétée comme une tâche particulière de reconnaissance de formes. Elle consiste à identifier la personne qui parle en se basant sur sa voix. La variabilité de la parole entre locuteurs est essentielle pour la RAL, car cela permet de distinguer une voix parmi plusieurs. Contrairement à la reconnaissance automatique de la parole, la RAL s'intéresse aux informations non linguistiques contenues dans le signal vocal. Cependant, la RAL bénéficie souvent des avancées de la reconnaissance automatique de la parole, avec de nombreuses techniques appliquées en RAP avant d'être adaptées à la RAL. Elle peut être réalisée à l'aide de différents algorithmes, tels que les réseaux de neurones, les machines à vecteurs de support et les modèles de Markov cachés. Les performances de ces algorithmes dépendent de nombreux facteurs, tels que la qualité de l'enregistrement audio, le bruit de fond et la variabilité interpersonnelle. Malgré ces défis, la reconnaissance automatique du locuteur continue de progresser et de trouver de nouvelles applications dans le monde de la technologie.

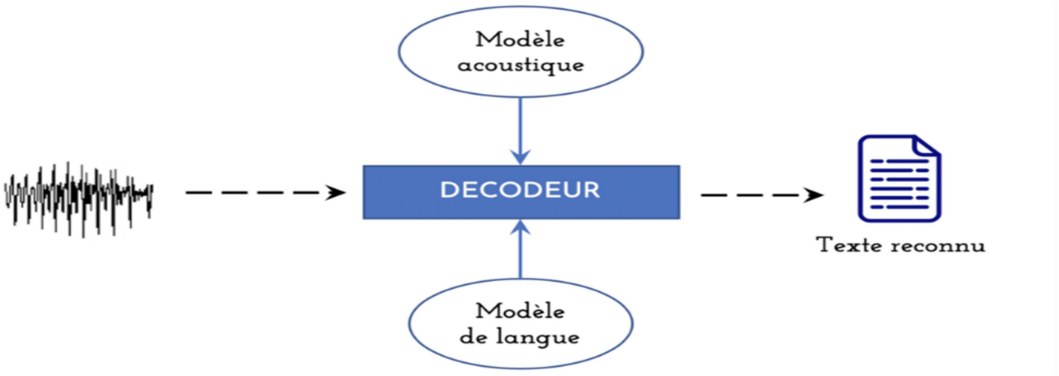

Figure 1.4. Schéma d'un système RAL.

1.6 Domaine d'application de la RAL

La reconnaissance automatique du locuteur est utilisée dans de nombreuses applications, Voici certains domaines d'application de cette technologie [11] :

a) Applications sur site

La personne doit faire l'objet d'une présentation physique à un endroit précis

. • Verrouillages vocaux (pour locaux, compte informatique, etc.)

• Interactivité matérielle (retrait d'argent à un guichet automatique, …)

b) Applications liées aux télécommunications

La vérification s'opère à distance :

• Accès à des services pour des abonnés, ou des données confidentielles.

• Transaction à distance.

c) Applications commerciales

• Associer un même mot de passe pour une petite population de locuteur (membre d'une famille, d'une société).

• Protection de matériel contre le vol.

d) Applications judiciaire

• Recherche de suspects et de preuves.

• Les juges, avocats, enquêteurs de police ou de gendarmerie souhaitent utiliser des procédés de reconnaissance vocale pour enquêter ou confirmer le coupable ou l'innocent.

1.7 Vérification Automatique du Locuteur

Consiste après que le locuteur a decliné son identité à vérifier l'adéquation du message vocal avec la référence acoustique du locuteur qu'il prétend être. C'est une décsion de tout ou rien. Les performances de vérification de locuteur sont données en termes de faux rejets f_r de fausses acceptations f_a [4].

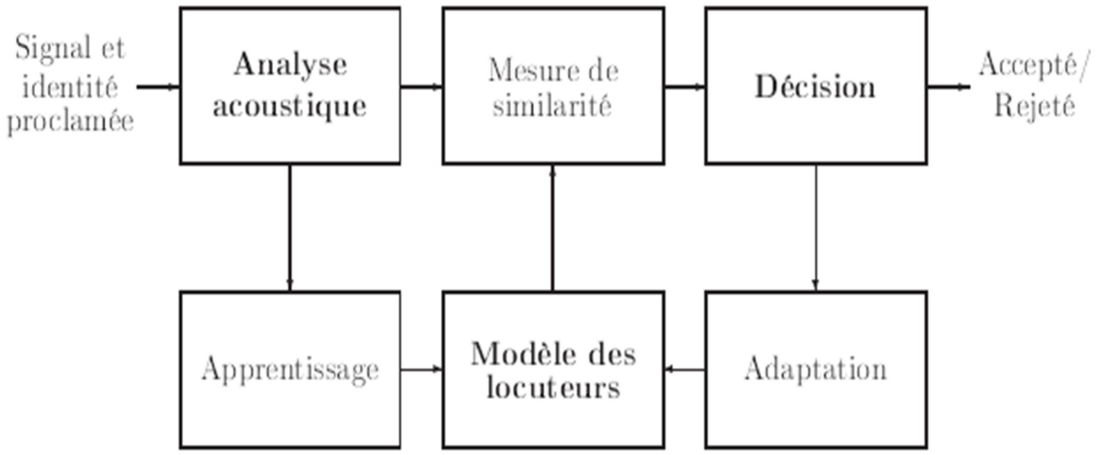

Figure 1.5 schéma typique d'un système de vérification du locuteur.[4]

1.8 Les paramètres MFCC

Le principe de calcul des MFCC (Mel Frequency Cepstral Coefficients) est issu des recherches psycho-acoustique sur la perception des différentes bandes de fréquences par l'oreille humaine. Le principal intérêt de ces coefficients est extraire des informations pertinentes en nombre limité en s'appuyant à la fois sur la production (théorie Cepstrale) et sur la perception de la parole (échelle des Mels).[12]

1.9 Les paramètres GTCC

Les Gammatone Frequency Cepstral Coefficients (GTCC) sont des descripteurs acoustiques utilisés pour l'analyse et la reconnaissance des signaux vocaux. Inspirés du fonctionnement de l'oreille humaine, les GTCC utilisent une banque de filtres Gammatone qui modélise plus précisément la perception auditive que les filtres triangulaires classiques utilisés dans les MFCC. Cette approche permet de capturer des caractéristiques plus robustes face au bruit et aux variations des conditions d'enregistrement. Grâce à leur efficacité dans la modélisation du spectre vocal, les GTCC sont de plus en plus utilisés dans les systèmes de reconnaissance automatique du locuteur, notamment dans les environnements acoustiques complexes.

1.10 Comparaison entre MFCC et GTCC

Critère	MFCC	GTCC
Inspiration biologique	Basé sur l'échelle de fréquence de Mel (perception auditive humaine)	Basé sur les filtres Gammatone (modèle plus proche de la cochlée)
Type de filtre utilisé	Filtres triangulaires sur l'échelle de Mel	Filtres Gammatone
Robustesse au bruit	Moyenne à faible dans les environnements bruyants	Meilleure robustesse, surtout en environnement bruité
Utilisation classique	Très répandu dans la reconnaissance vocale automatique	De plus en plus utilisé pour la reconnaissance du locuteur
Complexité computationnelle	Moins complexe, rapide à calculer	Légèrement plus complexe en raison de la banque de filtres Gammatone
Représentation perceptive	Approximation correcte de l'audition humaine	Représentation plus fidèle du traitement auditif naturel
Performance en reconnaissance	Bonnes performances en environnement contrôlé	Souvent meilleures performances en conditions réelles ou bruitées

Tableau 1.1 : Différence entre MFCC et GTCC [13,14]

Les MFCC ont longtemps été la référence dans l'analyse de la parole grâce à leur simplicité et leur efficacité (Davis & Mermelstein, 1980). Toutefois, pour les systèmes de reconnaissance dans des conditions bruitées ou non vocales, les GTCC ont été proposés comme une alternative plus robuste, car ils s'inspirent directement de la structure de la cochlée humaine (Valero & Alias, 2012). Plusieurs études récentes (Heittola et al., 2013 ; Schäfer et al., 2013) ont confirmé leur efficacité dans la classification de sons environnementaux complexes.

1.11 conclusion

Dans ce chapitre, nous avons présenté les bases théoriques de la reconnaissance automatique du locuteur, en détaillant les étapes clés du système : l'analyse acoustique du signal, la modélisation du locuteur et la décision. Nous avons exploré les paramètres acoustiques tels que les MFCC et les GTCC, en les comparant pour souligner leurs avantages respectifs.

Nous avons également abordé les techniques couramment utilisées, telles que le GMM pour la modélisation du locuteur, et les processus d'identification et de vérification. Ce premier chapitre constitue ainsi un socle théorique solide pour comprendre les développements plus avancés qui seront abordés dans le chapitre suivant, notamment l'application du deep learning et des Réseaux de Neurones Récurrents.

2.1 Introduction

L'intelligence artificielle permet aujourd'hui d'améliorer significativement la reconnaissance automatique du locuteur en modélisant efficacement les caractéristiques vocales. Ce chapitre se concentre sur les réseaux de neurones récurrents (RNN), adaptés au traitement des données séquentielles comme la parole, ainsi que sur les coefficients GTCC, une méthode avancée d'extraction des paramètres acoustiques. Ensemble, ces deux éléments forment la base de notre approche pour une reconnaissance plus robuste et précise.

2.2 Intelligence artificielle

2.2.1 Définition : L'intelligence artificielle est la simulation des processus de l'intelligence humaine par des machines, en particulier des systèmes informatiques. Les applications spécifiques de l'IA incluent les systèmes experts, le traitement du langage naturel, la reconnaissance vocale et la vision artificielle.

2.3 Deep learning

2.3.1 Définition : Le deeplearning ou apprentissage profond est un sous-domaine de l'intelligence artificielle (IA). Ce terme désigne l'ensemble des techniques d'apprentissage automatique (machine learning), autrement dit une forme d'apprentissage fondée sur des approches mathématiques, utilisées pour modéliser des données. Pour mieux comprendre ces techniques, il faut remonter aux origines de l'intelligence artificielle en 1950, année pendant laquelle Alan Turning s'intéresse aux machines capables de penser. Cette réflexion va donner naissance à la machine learning, une machine qui communique et se comporte en fonction des informations stockées. Ces neurones sont interconnectés pour traiter et mémoriser des informations, comparer des problèmes ou situations quelconques avec des situations similaires passées, analyser les solutions et résoudre le problème de la meilleure façon possible

2.3.2 Domaines d'application du deep learning

Le Deep Learning trouve aujourd'hui des applications dans un large éventail de domaines, notamment :

- la vision par ordinateur (ex. : reconnaissance faciale) ;
- le traitement automatique du langage naturel (ex. : traduction automatique) ;
- la reconnaissance vocale et du locuteur ;
- la santé (ex. : analyse d'images médicales) ;
- la finance (ex. : détection de fraudes) ;
- les jeux et l'intelligence artificielle interactive.

Dans cette étude, nous nous concentrerons particulièrement sur l'application du Deep Learning à la reconnaissance automatique du locuteur, en exploitant des architectures séquentielles telles que les réseaux de neurones récurrents (RNN), ainsi que des paramètres acoustiques comme les GTCC

2.4 Apprentissage automatique

L'apprentissage automatique, également connu sous le nom de machine learning, est un domaine de l'intelligence artificielle qui cherche à doter les machines de la capacité d'apprendre à partir de données en utilisant des modèles mathématiques. En substance, il s'agit du procédé par lequel des informations significatives sont extraites à partir d'un ensemble de données d'entraînement. L'objectif de cette phase est d'obtenir les paramètres d'un modèle qui atteindront les meilleures performances, notamment lors de l'exécution de la tâche assignée au modèle. Une fois l'apprentissage réalisé, le modèle peut ensuite être déployé en production.

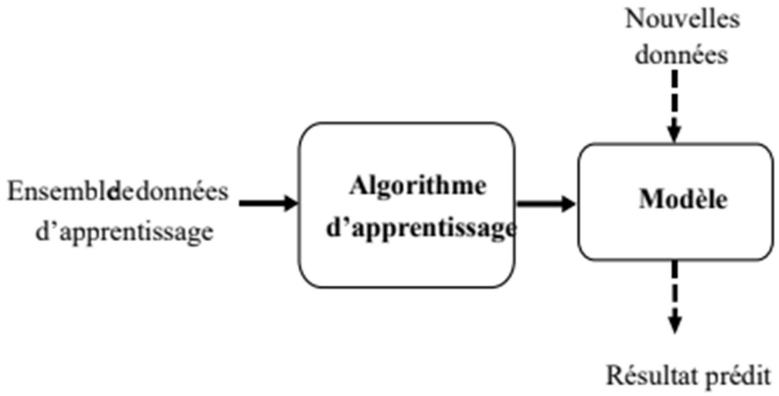

Figure 2.1: Processus de l'apprentissage machine [15]

2.5. Types d'apprentissages

2.5.1 Apprentissage supervisé

L'apprentissage supervisé est un processus d'apprentissage à partir d'un ensemble d'exemples d'apprentissage étiquetés fournis par un superviseur externe compétent. Cela signifie qu'une expertise humaine est nécessaire pour étiqueter les données. Chaque exemple consiste en une description d'une situation accompagnée d'une étiquette (une classe pouvant être représentée par des valeurs numériques ou nominales) indiquant l'action correcte que le système doit prendre dans cette situation. L'objectif de ce type d'apprentissage est que le système puisse généraliser ses réponses et agir correctement dans des situations non présentes dans l'ensemble d'apprentissage. Ainsi, l'utilisateur fournit à l'algorithme des paires d'entrées/sorties souhaitées (X, y), comme illustré dans la Figure 2.3, et l'algorithme trouve un moyen de produire la sortie souhaitée à partir des entrées. Plus précisément, l'algorithme est capable de générer une sortie pour une entrée qu'il n'a jamais rencontrée auparavant. [15]

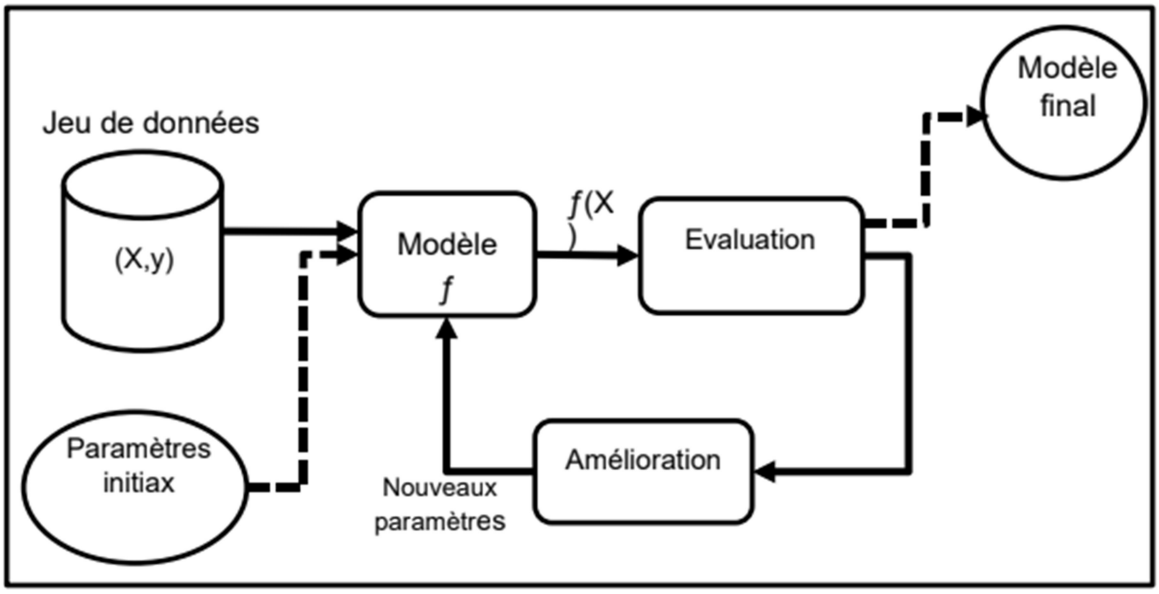

Figure 2.2 : Processus de l'apprentissage supervisé [15]

On distingue deux grands types de problèmes d'apprentissage supervisé : la régression et la classification.

a) La classification

Généralement, pour les problèmes de classification l'ensemble de données utilisé avec un nombre fini de classes, chaque exemple est associé à l'une d'entre elles. La cible pour chaque exemple a une valeur discrète représentant une classe particulière. Avec ces données, le modèle d'apprentissage automatique apprendra à attribuer des catégories aux entrées. Par exemple, dans le contexte de la classification de documents, imaginez un ensemble de données composé de trois catégories. Chaque classe correspond à une matière :"Economie", "Politique" et "Autre". Le modèle doit apprendre à déterminer si un document concerne l'économie, la politique ou d'autres sujets, puis doit associer le document à une valeur qui représente la catégorie correcte [16]

b) Régression

Dans le contexte d'un problème de régression, la variable cible est constituée d'un ou de plusieurs éléments ayant des valeurs continues. Un modèle d'apprentissage automatique est entraîné à prédire une ou plusieurs valeurs réelles. La météorologie offre un bon exemple de problème de régression, par exemple la prédiction de la température. En effet, la valeur à prédire dans ce cas est une quantité continue. On peut également inclure d'autres éléments dans la variable cible tels que la pression atmosphérique et le taux d'humidité, ce qui crée un vecteur de valeurs continues [16]

2.5.2 Apprentissage non-supervisé

Lorsque le système ou l'opérateur dispose uniquement d'exemples sans étiquettes, et que le nombre et la nature des classes n'ont pas été prédéterminés, on parle d'apprentissage non supervisé (ou regroupement). Aucune expertise n'est disponible ni requise. L'algorithme doit découvrir par lui-même la structure, plus ou moins cachée des données. Dans cet apprentissage, le système doit cibler les données dans l'espace de description en fonction de leurs attributs disponibles, afin de les regrouper en ensembles homogènes d'exemples. La similarité est généralement calculée à l'aide d'une fonction de distance entre les paires d'exemples. Ensuite, il revient à l'opérateur d'associer ou de déduire une signification pour chaque groupe [18]

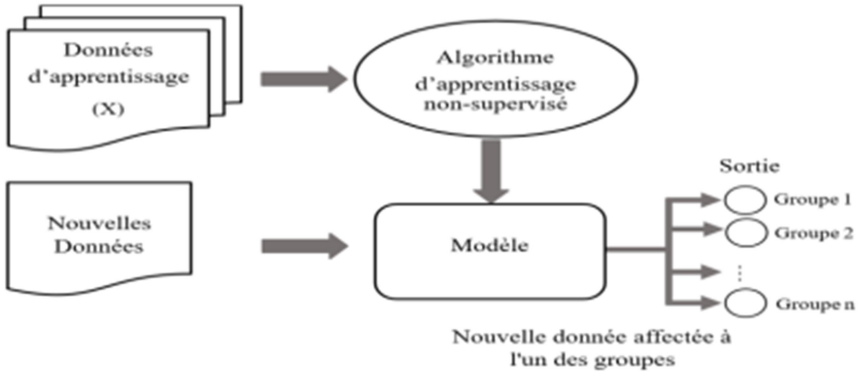

Figure 2.3 : Processus de l'apprentissage non-supervisé [16]

2.5.3 Apprentissage par renforcement

Le domaine de l'apprentissage par renforcement vise à enseigner à un agent comment se comporter de manière appropriée dans un environnement spécifique, c'est-à-dire atteindre un objectif préalablement choisi par l'utilisateur. Le problème à résoudre est divisé en une séquence d'étapes. À chaque étape, l'agent doit choisir parmi un ensemble d'actions, ce qui lui donne la possibilité d'interagir avec son environnement. Contrairement à l'apprentissage supervisé, il n'y a pas de cible permettant d'apprendre un comportement. À la place, l'agent reçoit un signal (déterminé par l'utilisateur) qui lui indique s'il a agi correctement. À chaque étape de la séquence, l'agent reçoit des informations sur son environnement qui l'aideront à choisir l'action appropriée. Pendant l'apprentissage, l'agent cherchera à maximiser le nombre de signaux positifs afin d'améliorer son comportement [18].

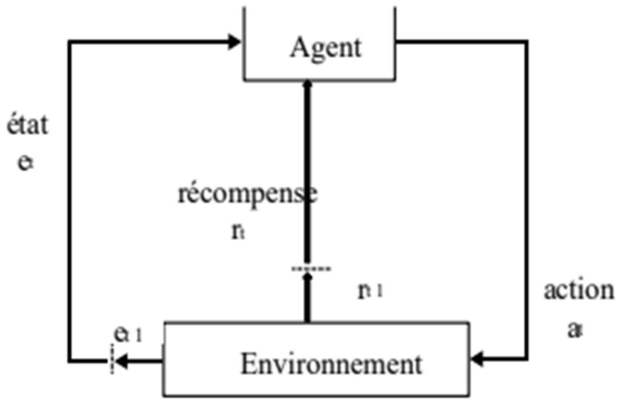

Figure 2.4 : Interaction agent-environnement[16]

2.6. Neurone biologique

Le cerveau humain est composé d'environ 1011 neurones, soit mille milliards, avec un nombre de connexions (synapses) allant de 1000 à 10000 par neurone. Le neurone est une cellule qui possède un corps cellulaire, qui agit comme un centre de contrôle et effectue la sommation des informations qui lui parviennent (voir Figure 2.5). Les dendrites, qui se ramifient à partir du corps cellulaire, permettent le transport des informations de l'extérieur vers le corps du neurone. Le neurone traite ensuite ces informations et les transmet le long de l'axone à d'autres neurones. La connexion entre deux neurones est appelée synapse. [16]

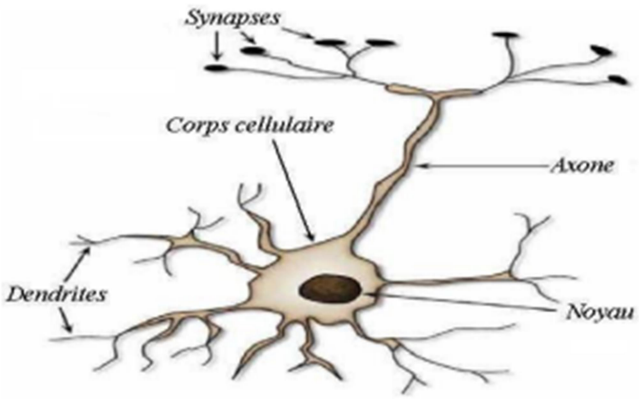

Figure 2.5 : Le neurone biologique

Les réseaux de neurones biologiques sont capables d'accomplir facilement certaines fonctions telles que la mémorisation, l'apprentissage par l'exemple, la généralisation, la reconnaissance des formes et le traitement du signal. À partir du principe que le comportement intelligent provient de la structure et du fonctionnement des neurones biologiques, des recherches ont conduit au développement des neurones formels, également appelés neurones artificiels [16].

2.7. Réseaux de neurones artificiels (RNA)

Un réseau de neurones artificiels est un système informatique inspiré du fonctionnement du cerveau humain, utilisé dans les ordinateurs dotés de capacités d'intelligence artificielle. Les réseaux de neurones artificiels sont conçus en se basant sur la structure des neurones biologiques du cerveau humain. Ils sont composés d'au moins deux couches de neurones - une couche d'entrée et une couche de sortie - et comprennent généralement des couches intermédiaires appelées "couches cachées ou hidden layer". La complexité du problème à résoudre détermine le nombre de couches nécessaires dans le réseau de neurones artificiels.

Chaque couche est composée d'un grand nombre de neurones artificiels spécialisés.

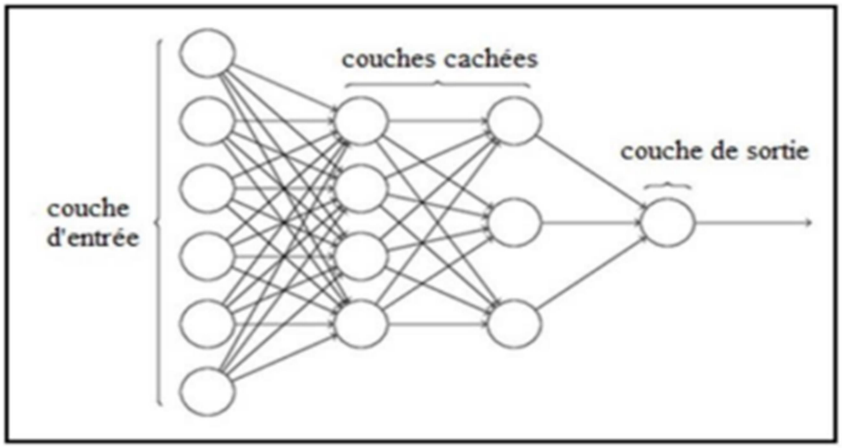

Figure 2.6 : Réseau de neurone artificiel

Les réseaux de neurones trouvent des applications dans divers domaines. Ils sont utilisés pour résoudre des problèmes de classification, de régression et même pour estimer la densité de probabilité. Ils sont employés à la fois dans l'apprentissage supervisé et non supervisé, ainsi que dans les modèles discriminatifs et génératifs. En résumé, ils constituent une famille de modèles extrêmement flexibles et puissants qui méritent d'être explorés davantage[18].

2.8 Fonctionnement des réseaux de neurones artificiels

Le réseau de neurones artificiels repose sur l'utilisation de plusieurs processeurs qui fonctionnent en parallèle. Ces processeurs sont organisés en couches. La première couche est responsable de la réception des entrées de données brutes. Chaque couche subséquente reçoit ensuite les sorties d'informations provenant de la couche précédente. La dernière couche génère les résultats du système. Pour traiter des problèmes plus complexes, il est souvent nécessaire d'avoir plusieurs couches. Chaque neurone possède

une valeur spécifique qui détermine quelle information peut être transmise dans le système. La fonction d'activation est utilisée pour calculer la valeur de sortie de chaque neurone. Ce calcul détermine combien de neurones doivent être activés pour résoudre le problème. Un algorithme est ensuite créé associant un résultat à chaque entrée. L'algorithme permet à l'ordinateur d'apprendre à partir de nouvelles informations qu'il reçoit.

Le réseau de neurones permet à l'ordinateur d'analyser des exemples et d'acquérir des capacités pour effectuer des tâches spécifiques. Ces exemples sont généralement étiquetés. Ce processus a permis aux ordinateurs de reconnaître des objets dans des images, parfois de manière plus performante que le cerveau humain lui-même.

Tout comme le cerveau humain, les réseaux de neurones artificiels ne peuvent pas être directement programmés, mais doivent apprendre en étudiant et en analysant des exemples [19].

2.9 Types de réseaux de neurones artificiels

Les types de réseaux de neurones sont généralement classifiés en fonction du nombre de couches nécessaires entre l'entrée des données et la sortie finale. De plus, le type de réseau est déterminé en fonction du nombre de nœuds cachés présents dans chaque modèle. On tient également compte du nombre d'entrées et de sorties de chaque nœud [20].

a) Réseaux de neurones à propagation avant

Le réseau de neurones de base est connu sous le nom de Feedforward. Les données de ce type de réseau se propagent directement depuis les entrées jusqu'aux nœuds de traitement. Ensuite, ils vont directement aux sorties [20].

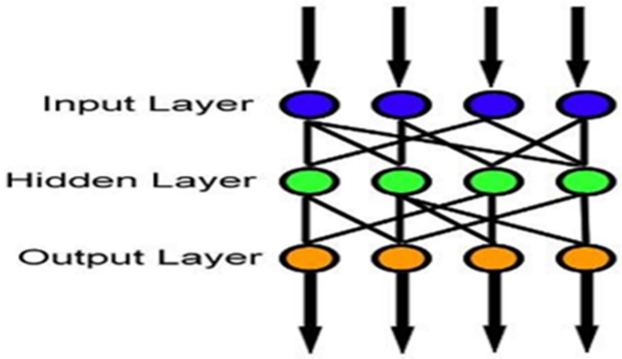

Figure 2.7 : Réseaux de neurones à propagation avant

b) Réseaux de neurones convolutives

Les réseaux de neurones convolutives également connus sous le nom de CNN, sont utilisés pour détecter des motifs simples à l'intérieur d'une image afin d'identifier son contenu en effectuant des recoupements. Leur usage est de plus en plus répandu dans une variété de domaines, comme la reconnaissance faciale et la numérisation de texte. Les CNN comprennent au moins cinq couches et le résultat s'étend d'une couche à l'autre [20].

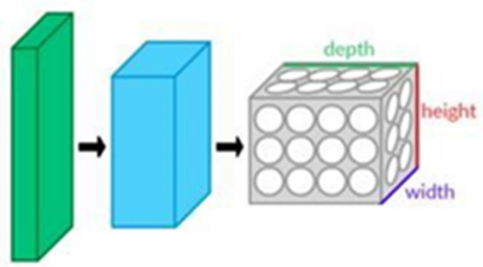

Figure 2.8 : Réseau neuronal convolutif

c) Réseaux de neurones récurrents

Les réseaux neuronaux récurrents (RNN) sont une variante très importante de réseaux neuronaux, largement utilisés dans le traitement du langage naturel. Ce qui distingue les RNN, c'est leur capacité à effectuer la même tâche pour chaque élément d'une séquence, où la sortie dépend des calculs précédents. On peut également dire que les RNN possèdent une "mémoire" qui capture des informations sur ce qui a été calculé jusqu'à présent. En théorie, les RNN peuvent utiliser des informations provenant de séquences de longueur arbitraire, mais en pratique ils se limitent souvent à examiner uniquement les étapes récentes. Les RNN sont une classe de réseaux neuronaux qui permettent aux prédictions antérieures d'être utilisées comme entrées grâce à l'utilisation d'états cachés. La structure typique d'un RNN est représentée dans la Figure 2.12 [20].

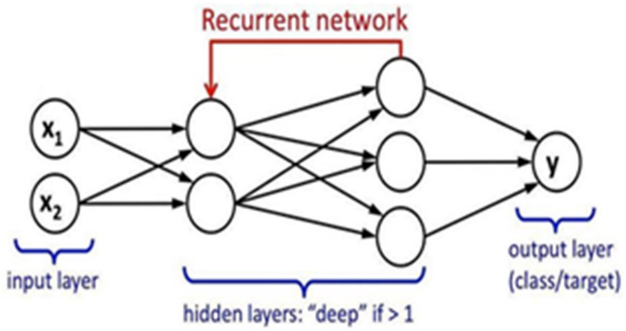

Figure 2.9 : Architecture des réseaux de neurones récurrents

2.10.Réseaux de neurones récurrents

Un réseau de neurones récurrent (RNN) est un type de réseau de neurones artificiel qui utilise des données séquentielles ou des données de séries temporelles. Ces algorithmes d'apprentissage en profondeur sont couramment utilisés pour des problèmes ordinaux ou temporels, tels que la traduction linguistique, le traitement du langage naturel, la reconnaissance vocale et le sous-titrage d'images ; ils sont incorporés dans des applications populaires telles que Siri, la recherche vocale et Google Translate. Comme les réseaux de neurones convolutifs (CNN) à propagation avant, les réseaux de neurones récurrents utilisent des données d'entraînement pour apprendre. Ils se distinguent par leur « mémoire » car ils prennent des informations d'entrées antérieures pour influencer l'entrée et la sortie en cours. Alors que les réseaux de neurones profonds traditionnels supposent que les entrées et les sorties sont indépendantes les unes des autres, la sortie des réseaux de neurones récurrents dépend des éléments antérieurs au sein de la séquence. Alors que les événements futurs seraient également utiles pour déterminer la sortie d'une séquence donnée, les réseaux de neurones récurrents unidirectionnels ne peuvent pas rendre compte de ces événements dans leurs prédictions [21].

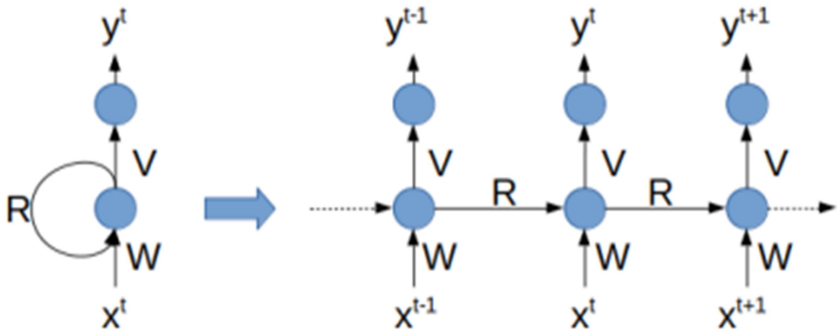

Figure 2.9 : (à gauche) Un RNN (à droite) Sa version déroulé Source

2.10.1 Types de réseaux de neurones récurrents

Les types de réseaux de neurones récurrents (RNN) comprennent plusieurs variantes qui ont été développées pour améliorer les capacités des RNN traditionnels. Pour traiter ce type de données, il existe trois grands types de réseaux de neurones récurrents :le RNN simple, le LSTM et le GRU.

• Le RNN simple est la forme la plus basique de RNN, ne possédant pas de portes pour contrôler le flux d'informations. Cependant, en pratique, les RNN simples ne sont généralement pas utilisés en raison de leurs limitations.

• Les LSTM sont une architecture populaire de RNN introduite pour résoudre le problème de la disparition du gradient. Les LSTM sont une variante avancée de RNN qui surmontent les limitations des RNN simples en introduisant des mécanismes de mémoire à court et long terme. Les LSTM utilisent des portes pour contrôler le flux d'informations et sont efficaces pour gérer des dépendances à long terme.

• Les GRU sont une autre variante de RNN qui simplifient l'architecture des LSTM en combinant les portes d'oubli et d'entrée en une seule porte d'update. Les GRU sont efficaces pour des tâches similaires aux LSTM mais avec une architecture plus simple.

Comme pour les réseaux de neurones traditionnels, les réseaux de neurones récurrents peuvent contenir plusieurs couches, ce qui leur permet de capturer davantage de non-linéarité parmi les données, mais augmente également le temps de calcul en phase d'apprentissage. On peut également combiner des couches récurrentes avec des couches classiques, telles que des couches denses (MLP) ou des couches de convolution (CNN) [22].

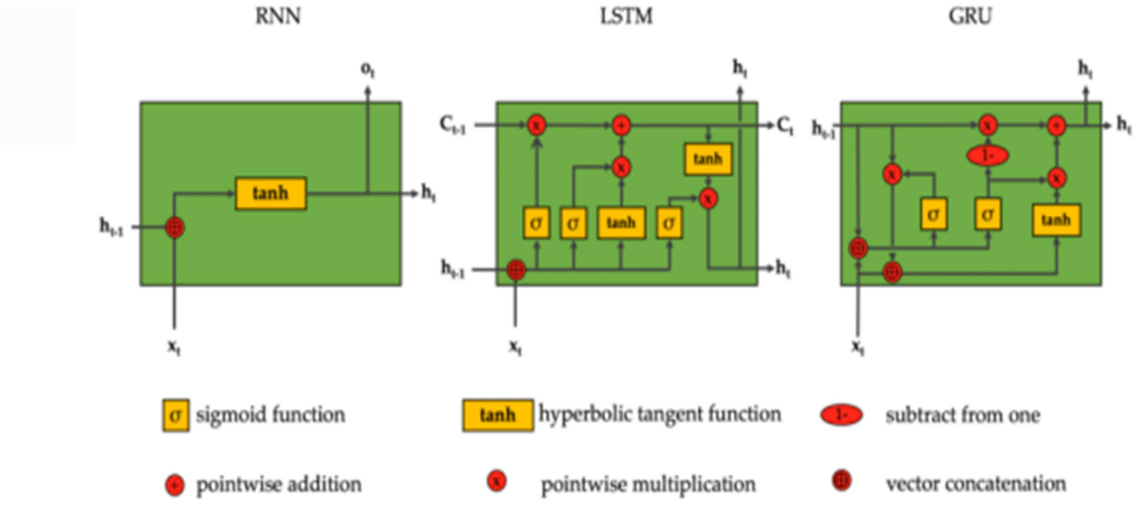

Figure 2.10 : Les types de réseaux de neurones récurrents

2.10.2 Architecture de RNN

L'architecture d'un réseau neuronal récurrent (RNN) se compose généralement de plusieurs couches répétitives qui sont empilées les unes sur les autres. Comme mentionné précédemment, chaque couche récurrente peut être un simple RNN, LSTM ou un GRU. L'architecture d'un RNN est constituée de :

- **Entrée (Input)**: Les données séquentielles sont introduites dans le réseau par l'intermédiaire de l'entrée. Chaque séquence de données est représentée par une série d'éléments (par exemple, des mots dans une phrase ou des instants temporels dans une série temporelle).
- **Couche récurrente (Récurrent Layer)**: Les informations séquentielles sont traitées dans la couche récurrente. Cette couche possède des connexions de rétroaction qui permettent aux informations de circuler d'une étape à l'autre dans la séquence. Cela permet au réseau de capturer les dépendances temporelles et de modéliser les relations complexes entre les éléments séquentiels.
- **Optionnel** : Stacking de couches récurrentes (Stacking Récurrent Layers) : Il est possible d'empiler plusieurs couches récurrentes les unes sur les autres pour former un réseau de neurones récurrents profond. Chaque couche récurrente traite les informations provenant de la couche précédente, permettant ainsi au réseau de capturer des niveaux de représentation plus abstraits et complexes.
- **Couche de sortie (Output Layer):** La couche de sortie génère les prédictions ou les sorties souhaitées en fonction des informations traitées par les couches récurrentes. La nature de la tâche détermine le type de couche de sortie utilisée. Par exemple, pour la classification, une couche de sortie dense avec une fonction d'activation appropriée peut être utilisée [23].

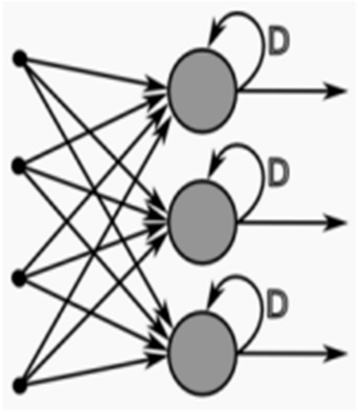

Figure 2.11: Recurrent Neural Network (RNN) Tutorial

2.10.3 Fonctionnement des RNN

Les réseaux de neurones récurrents (RNN) fonctionnent sur deux principes clés selon les sources fournies :

1. Astuce de la fenêtre glissante et Connexions Récurrentes

Les RNN reposent sur l'astuce de la fenêtre glissante pour traiter des signaux de taille variable. Ils utilisent des connexions récurrentes qui permettent d'analyser la partie passée du signal, offrant au réseau la capacité de "voir" la fenêtre correspondante à un instant donné et de se "souvenir" de sa décision à un instant précédent. Ces connexions récurrentes permettent au réseau de neurones de conserver une mémoire interne et de prendre en compte les dépendances temporelles dans les données.

2.Modélisation des Dépendances Temporelles

Les RNN sont conçus pour traiter des séquences de taille variable tout en modélisant les dépendances au sein de la séquence d'entrée. Ils peuvent être approximés par des réseaux non récurrents dépliés dans le temps, ce qui permet de visualiser leur fonctionnement de manière plus claire. Les RNN peuvent être utilisés pour des tâches telles que l'étiquetage de séquences, la classification de séquences et la génération de séquences, comme la prédiction de mots suivants dans un texte ou la classification de sentiments dans des avis en ligne [21].

2.10.4 Comparaison entre RNN et CNN

Critère	CNN (Convolutional Neural Network)	RNN (Recurrent Neural Network)
Structure	Composée de couches convolutives, souvent avec pooling	Inclut des boucles récurrentes reliant les étapes précédentes
Utilisation typique	Vision par ordinateur (image classification, détection d'objets)	Traitement de séquences (texte, audio, séries temporelles)
Mémoire	Pas de mémoire temporelle	Possède une mémoire courte grâce à la récursivité
Apprentissage de séquences	Moins adaptée	Bien adaptée, surtout avec LSTM GRU
Parallélisassions	Facile à paralléliser	Difficile à paralléliser à cause des dépendances temporelles
Performance sur données temporelles	Moins efficace	Plus performante, selon le contexte

Tableau 2.1 : Comparaison entre CNN et RNN[24,25].

2.10 Les réseaux Gammatone cepstral coefficients GTCC

Les Gammatone Cepstral Coefficients (GTCC) sont des coefficients cepstraux extraits à partir d'un signal audio en appliquant une banque de filtres Gammatone, suivie d'une opération de transformation logarithmique et de la transformée en cosinus discrète (DCT). Cette méthode vise à mieux imiter le comportement auditif humain par rapport aux MFCC [22].

2.11.1 Fonctionnement des GTCC

Le processus de calcul des GTCC se déroule en plusieurs étapes, inspirées de la modélisation du système auditif humain, notamment la cochlée. Voici les principales étapes :

Pré-accentuation et encadrement

Le signal audio est d'abord pré-accentué (pour amplifier les hautes fréquences), puis découpé en trames courtes (frames), généralement de 20 à 30 ms, avec un recouvrement entre trames (par exemple, 50%).

Fenêtrage

Chaque trame est multipliée par une fenêtre (souvent de type Hamming) pour réduire les effets de discontinuité aux bords de la trame.

Filtrage Gammatone

Une **banque de filtres Gammatone** est appliquée. Ces filtres sont conçus pour simuler le comportement de la cochlée humaine en réponse aux sons. Ils sont espacés selon une échelle fréquentielle proche de celle de la **Bark scale**, ce qui imite la sensibilité fréquentielle de l'oreille humaine.

Extraction de l'énergie par canal

L'énergie du signal filtré dans chaque bande Gammatone est calculée, souvent en utilisant la racine carrée ou le logarithme pour se rapprocher de la perception auditive.

Logarithme

Comme le système auditif humain perçoit l'intensité de manière logarithmique, on applique une transformation logarithmique à ces énergies.

- **Transformée en cosinus discrète (DCT)**

  Une DCT est appliquée aux énergies log-transférées pour obtenir les coefficients cepstraux. Cela permet de réduire la redondance et de ne garder que les composantes les plus significatives (généralement les premiers 12 à 20 coefficients).

- **(Optionnel)** Post-traitement

  Cela peut inclure la dérivation (coefficients delta et delta-delta) pour capturer les variations temporelles [16].

1.3. Conclusion

Dans ce chapitre, nous avons abordé les principales techniques d'extraction et de traitement de caractéristiques audio essentielles pour la reconnaissance des émotions à partir de la voix. Nous avons étudié les fondements des réseaux de neurones, notamment les architectures RNN et CNN, en soulignant leurs différences structurelles et leurs applications spécifiques. Les RNN, grâce aux cellules LSTM et GRU, sont particulièrement adaptés au traitement des données séquentielles telles que les signaux vocaux. En outre, nous avons exploré les coefficients cepstraux GTCC comme alternative prometteuse aux MFCC, grâce à une modélisation plus fidèle du système auditif humain. Le processus de génération des GTCC basé sur la banque de filtres Gammatone améliore la qualité des représentations spectrales pour les applications de reconnaissance émotionnelle. Ce chapitre pose les bases théoriques nécessaires pour l'analyse vocale émotionnelle et prépare le terrain pour les expérimentations pratiques à venir.

3.1 Introduction

Dans ce chapitre, nous menons une analyse comparative de deux systèmes d'apprentissage automatique dédiés à l'identification du locuteur, afin d'évaluer leur efficacité.

L'étude porte sur deux corpus distincts : l'un monolingue et l'autre multilingue fonctionnant avec une fréquence d'échantillonnage 16 kHz. Notre démarche consiste en une comparaison des performances de ces deux systèmes, en mettant l'accent sur leur capacité à s'adapter aux variations linguistiques.

La première phase de notre étude consiste en la constitution d'une base de données vocale, servant de fondement à l'entraînement et à l'évaluation de deux modèles d'identification de locuteurs. Ces modèles reposent sur des architectures de réseaux de neurones convolutifs (CNN) et récurrents (RNN), intégrant des coefficients cepstraux GTCC comme caractéristiques d'entrée. Chaque modèle est soumis à une série de tests en variant les paramètres afin d'optimiser leurs performances.

Dans la seconde phase la technique de l'augmentation de données est exploitée afin améliorer la robustesse des deux modèle

L'ensemble des expérimentations est mis en œuvre à l'aide de bibliothèques Python spécialisées dans le traitement du signal audio et l'apprentissage automatique, notamment Librosa pour l'extraction des caractéristiques audio et TensorFlow pour la construction et l'entraînement des modèles de réseaux de neurones.

3.2 Langage python

Conçu par Guido van Rossum et lancé en 1991, Python est un langage de programmation très prisé des professionnels de la donnée. Ses usages vont bien au-delà de la Data Science, englobant le développement de logiciels, la création d'algorithmes et même la gestion d'infrastructures web complexes pour les réseaux sociaux.

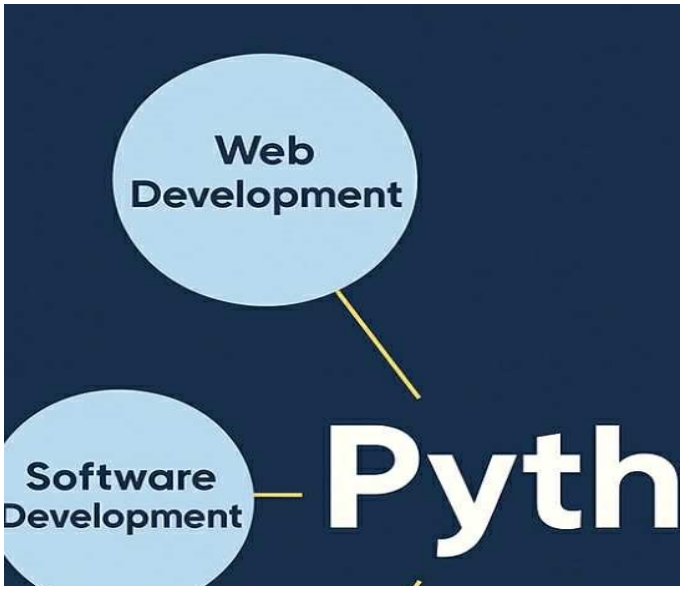

Figure 3.1: Les domaines d'applications de python

Python est un langage de programmation à la fois simple et puissant qui permet de créer des scripts faciles à écrire grâce à ses nombreuses bibliothèques. Il se distingue par son caractère interprété, ce qui signifie qu'il peut être exécuté directement sans compilation préalable. Sa polyvalence lui permet de fonctionner sur divers systèmes d'exploitation tels que Raspberry Pi, Mac OS X, Linux, Android et iOS.

Dans le cadre de nos travaux, nous avons utilisé Jupyter, implémenté via ANACONDA, comme environnement d'expérimentation.

3.3. DEMARCHE METHODOLOGIQUE

Dans le cadre de nos recherches expérimentales, nous avons mené une série d'études approfondies qui nous ont permis d'explorer en détail divers aspects de la reconnaissance automatique du locuteur. Le schéma illustré dans la figure 3.2 décrit de manière détaillée la démarche méthodologique que nous avons adoptée pour notre étude.

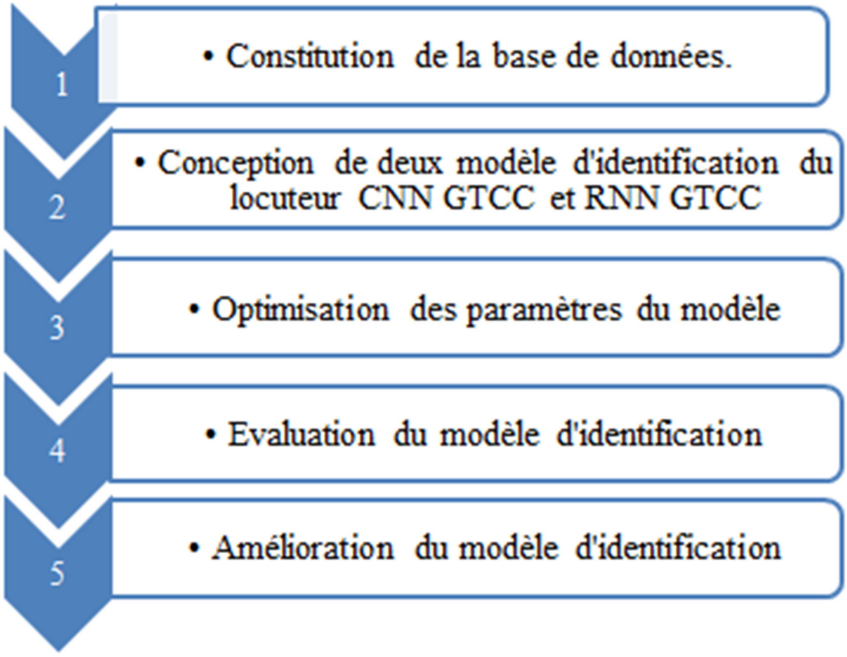

Figure 3.2 : démarche méthodologique de notre travail.

3.4 Collecte de la base de données

Pour la constitution de notre base de données multilingue, nous avons adopté une procédure similaire à celle utilisée pour la base de données TIMIT du dialecte américain, avec quelques adaptations nécessaires. Cependant, certaines conditions n'ont pu être strictement respectées pour certains enregistrements.

Les critères suivants ont guidé notre collecte :

1. Sélection de phrases phonétiquement équilibrées.
2. Enregistrements effectués dans un environnement acoustiquement contrôlé, au format WAV.
3. Utilisation d'un microphone haute performance placé à 20 cm du locuteur.
4. Uniformisation de la durée des échantillons d'apprentissage.
5. Protocole d'enregistrement : chaque locuteur a enregistré 23 phrases, réparties comme suit : 14 en dialecte algérien, 5 en français et 4 en anglais. Les enregistrements ont été réalisés à une

fréquence d'échantillonnage de 16 kHz, en incluant des participants de différentes tranches d'âge et des deux genres.

3.5.Conception des modèles d'identification du locuteur

Dans cette partie, nous décrivons en détail l'approche méthodologique déployée pour développer les deux modèles d'identification RNN GTCC et CNN GTCC.

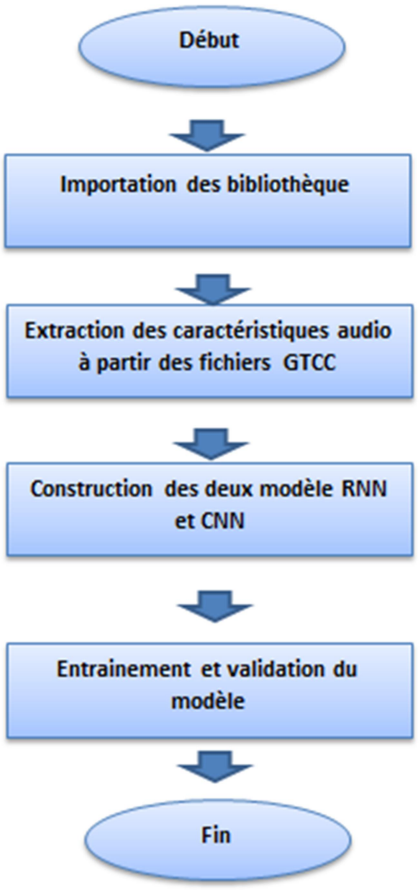

Figure 3.3 : Modèle d'apprentissage

Le processus méthodologique adopté dans cette étude se décompose en quatre étapes principales, qui sont cités dans l'organigramme. Chacune de ces étapes fera l'objet d'une description détaillée dans les sections suivantes

3.5.1. Importation de la bibliothèque

Les différentes bibliothèques utilisées et leurs fonctions sont présentées dans le tableau III.1 :

Bibliothèque	Fonction
Numpy	Opérations numériques
Tensorflow	Construction et entraînement du modèle de Deep Learning
Librosa	Traitement audio (lecture, analyse)
Os	Gestion des fichiers et des répertoires
Sklearn	Fournit des outils pour l'apprentissage automatique (modèles, évaluation, preprocessing).
Soundfile	Lecture et écriture de fichiers audio (formats WAV, FLAC, etc.).
matplotlib	Visualisation des données sous forme de graphiques et de diagrammes
Gammatone	Manipuler les filtres et représentation pour l'audio

Tableau 3.1 : Bibliothèque python utilisée dans le modèle d'identification

3.5.2. Extraction des caractéristiques audio

Cette étape comprend les opérations suivantes :

a. Chargement des fichiers audio à l'aide de librosa

b. Calcul des GTCC (Mel-Frequency Cepstral Coefficients) qui représentent les caractéristiques spectrales du signal audio

a. Chargement et préparation des données

Le Processus de Préparation des Données consiste à :

1. **Parcours des répertoires d'étiquettes**

   ✓ Exploration systématique de l'arborescence des dossiers

   ✓ Association automatique des noms de répertoire aux étiquettes de classification

   ✓ Détection des sous-dossiers contenant les échantillons audios

2. **Extraction des caractéristiques GTCC**

   - ✓ Charger les fichiers audios avec librosa .
   - ✓ Calculer le gammatonegram.
   - ✓ Appliquer la fonction inverse de la transformée de fourrier pour obtenir les coefficients cepstraux.
   - ✓ Filtrer et stocker les caractéristiques extraites dans des listes

3. **Encodage des étiquettes**

   - ✓ Transformation des libellés textuels en valeurs numériques via LabelEncoder
   - ✓ Création d'un mapping inversible (label → index numérique)
   - ✓ Vérification de l'équilibrage des classes

4. **Contrôle de qualité**

   - ✓ Exclusion des fichiers corrompus ou non lisibles
   - ✓ Validation de la cohérence entre le nombre d'échantillons et d'étiquettes
   - ✓ Vérification des dimensions des caractéristiques extraites

3.5.3. Construction du modèle RNN

La conception du modèle RNN repose sur les étapes suivantes :

- ✓ . Intégration de couches LSTM, adaptées au traitement des données séquentielles, afin de capturer les dépendances temporelles dans les caractéristiques audio.
- ✓ Application de la technique de régularisation Dropout, visant à réduire le risque de surapprentissage en désactivant aléatoirement certaines unités lors de l'entraînement.
- ✓ Compilation du modèle à l'aide de la fonction de perte categorical cross-entropy, appropriée pour les tâches de classification multi-classes, et de l'optimiseur Adam, reconnu pour sa performance et sa stabilité dans l'apprentissage des réseaux de neurones profonds.

3.5.4. Construction du modèle CNN

Notre approche pour construire le modèle CNN consiste en :

1. Création d'un modèle séquentiel pour empiler les couches linéairement.
2. Définition des couches nécessaires, notamment :
   - ✓ Convolution (Conv1D) pour extraire les caractéristiques.
   - ✓ Pooling pour réduire la dimensionnalité.
   - ✓ Flattening pour aplatir les données.
   - ✓ Fully connected pour prendre des décisions.
3. Utilisation de Conv1D pour détecter des modèles dans les données séquentielles.

3.5.5. Entraînement et évaluation du modèle

Dans cette section, nous décrivons le processus d'entraînement et d'évaluation du modèle.

1. Entraînement du modèle

Le modèle est entraîné à l'aide des données d'apprentissage. Plusieurs paramètres d'entraînement sont spécifiés, notamment :

- ✓ **Valid split :** Cette variable représente la proportion de données qui sont utilisées comme ensemble de validation lors de l'entraînement de notre modèle. Dans notre cas, 15% des données sont utilisées pour la validation et 15%. Pour le test
- ✓ **Sample rate :** La variable ‘sample rate’ spécifie le taux d'échantillonnage des enregistrements audio de notre base de données. Cela indique la fréquence à laquelle le signal audio a été enregistré.
- ✓ **Batch size :** Cette variable spécifie la taille des lots (Batches) utilisés lors de l'entraînement du modèle. Nous avons utilisé un lot de **:** 32

- ✓ **Epochs :** La variable `epochs` définit le nombre d'epochs (itérations complètes) pendant les quelle nous avons entraîné le modèle. Un epoch correspond à une passe complète sur l'ensemble des données d'entraînement. Le nombre d'époques utilisé dans notre code : 100.

- ✓ **Fonction d'activation :** elle se comporte comme un interrupteur dans un neurone artificiel : elle décide si le neurone doit "s'activer" ou pas, en fonction de ce qu'il reçoit. Elle aide le réseau à apprendre des choses complexes en ajoutant de la non-linéarité, un peu comme un cerveau qui ne réagit pas toujours de manière prévisible**.** Dans notre cas, Les couches caché utilisent la fonction RELU et la couche : SOFTMAX.

- ✓ **Fonction de perte** : La fonction de perte est un outil mathématique qui quantifie l'écart entre les prédictions du modèle et les valeurs réelles (étiquettes cibles). Elle sert de critère d'optimisation pendant l'apprentissage en guidant l'ajustement des poids du réseau. Une valeur de perte faible indique que le modèle fait de bonnes prédictions.

- ✓ **Optimiseur** : Un optimiseur est un algorithme qui ajuste les poids du modèle afin de minimiser la fonction de perte. Des exemples courants incluent. Son objectif est de minimiser l'erreur de prédiction sur les données d'entraînement tout en évitant le surapprentissage. L'optimiseur déployé dans notre modèle : ADAM.

2. Évaluation et prédiction

Une fois l'entraînement terminé, le modèle est évalué sur les données de test afin de mesurer sa performance sur des exemples non traités auparavant.

- ✓ **Évaluation** : les métriques de performance (telles que la précision, la perte, la matrice de confusion, etc.) sont calculées à partir des prédictions du modèle sur l'ensemble de test.

- ✓ **Affichage des résultats de précision** : un résumé des résultats est présenté, mettant en évidence la précision globale du modèle. La précision est définie comme le rapport entre les échantillons correctement classés et le nombre total d'échantillons dans l'ensemble de validation, constituant ainsi une mesure clé pour évaluer les performances d'un modèle de classification.

- ✓ **Prédictions** : le modèle effectue des prédictions sur les données de test. Les classes prédites sont comparées aux classes réelles pour illustrer la qualité de la classification.

3.6. Optimisation des paramètres du modèle

Pour optimiser les performances des deux modèles, nous avons mené plusieurs expérimentations étudiant les phénomènes de sous-apprentissage et de surapprentissage sur le corpus de dialecte algérien (monolingue). Les figures 3.4 et 3.5 Illustrent respectivement les taux de précision et de perte (Entrainement et Validation) des modèles RNN GTCC et CNN GTCC.

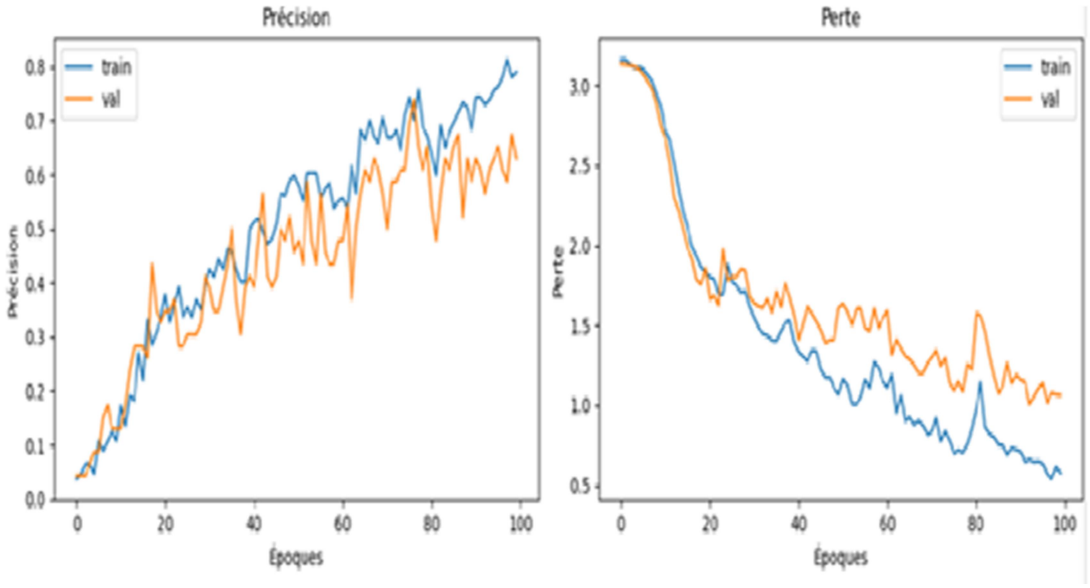

Figure 3.4 : Taux de précision et de perte (entrainement et validation) du modèle RNN GTCC.

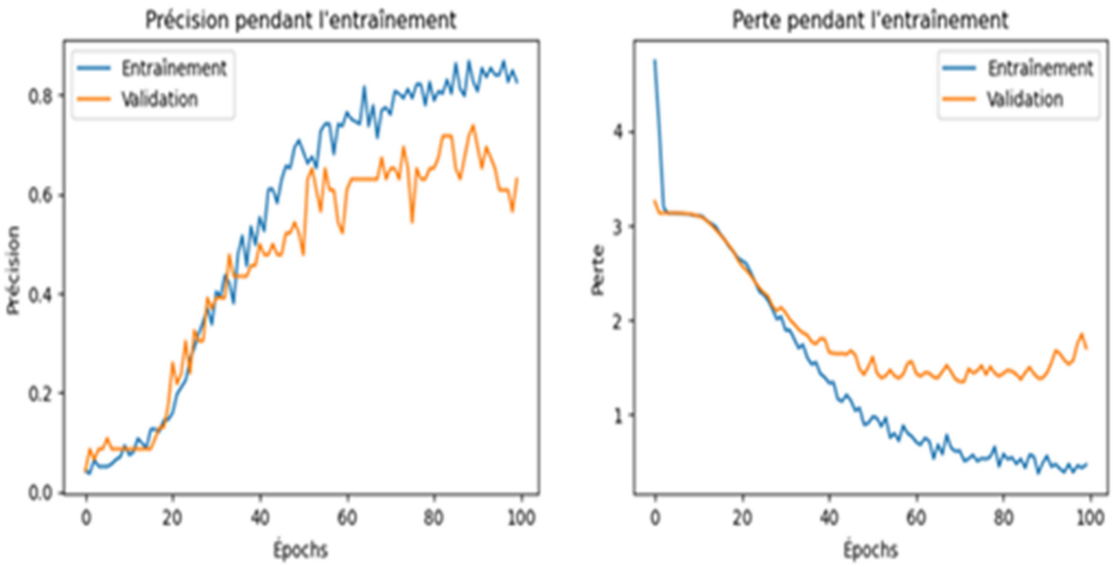

Figure 3.5 : Taux de précision et de perte (entrainement et validation) du modèle CNN GTCC.

Les résultats obtenus présentent notamment un sur-apprentissage important pour les deux modèles. Les taux de précision de validation, de perte et de test sont donnés par le tableau III.1 :

Modèle / Taux	RNN GTCC	CNN GTCC
Précision de validation maximale	71,74%	73,91%
Test	52,39%	60,87%

Tableau 3.2 : Taux de : précision de validation, Perte et test pour RNN GTCC et CNN GTCC.

D'après le tableau, On note :

- ✓ Hormis le phénomène de sur-apprentissage remarqué dans les résultats des deux modèles, les taux de CNN GTCC sont nettement meilleur.

- ✓ les Taux de test des deux modèles sont insuffisants pour un système de classification qui doit atteindre ,au moins, un taux de test de 90 % .

Dans le cadre de l'amélioration les performances des deux modèles, nous avons employé une stratégie d'optimisation basée sur l'algorithme RANDOM SEARCH. Cette approche nous a permis de déterminer les paramètres les plus appropriés pour maximiser la précision et minimiser la perte de validation, ainsi que pour améliorer la classification.

En plus de cette optimisation, nous avons intégré plusieurs paramètres supplémentaires pour renforcer la robustesse des deux modèles :

- ✓ **Régularisation L1 et L2** : Pour prévenir le surapprentissage, nous avons appliqué les techniques de régularisation L1 et L2, qui ajoutent une pénalité aux poids importants dans la fonction de perte.
- ✓ **Fonction Shuffle** : Nous avons utilisé la fonction Shuffle pour mélanger aléatoirement les données d'entraînement. Cela empêche le modèle d'apprendre des ordres artificiels ou des motifs non représentatifs dans les données.
- ✓ **Taux d'apprentissage adaptatif** : L'utilisation d'un taux d'apprentissage adaptatif nous permet d'ajuster dynamiquement le taux d'apprentissage pendant l'entraînement. Cela aide à éviter le surapprentissage et le sous-apprentissage en ajustant la vitesse d'apprentissage en fonction des besoins du modèle.
- ✓ **Early Stopping** : Enfin, nous avons mis en œuvre la fonction Early Stopping pour arrêter l'entraînement lorsque les performances du modèle sur les données de validation commencent à se dégrader. Cela permet d'éviter le surapprentissage en stoppant l'entraînement au moment opportun.

3.7. Evaluation du modèle d'identification

Dans cette expérimentation, nous avons appliqué les deux modèles à deux ensembles d'enregistrements multilingues et monolingues (dialecte algérien), les performances est évaluée par les : taux de perte(learning et validation) , taux de précision (learning et validation) et la matrice de confusion avec le taux de test.

3.7.1.Analyse des taux de perte et précision

Les figures 3.6 3.7 3.8 et 3.9 illustrent respectivement les courbes de perte et de précision(entrainement et validation des modèle RNN GTCC monolingue et multilingue, et CNN GTCC monolingue et multilingue:

En général, la précision du modèle augmente à mesure que le nombre d'époques d'entraînement augmente, ce qui indique un meilleur ajustement du modèle aux données pendant l'apprentissage. La différence dans Le nombre d'époques observé est induite par l'algorithme Early Stopping.

Les résultats obtenus montrent que l'ensemble de données multilingues donne les meilleurs résultats, démontrant ainsi sa robustesse en termes d'extraction de caractéristiques vocales. De plus, le modèle CNN GTCC affiche les taux de performance les plus élevés pour cette même base de données, ce qui indique son efficacité pour l'identification du locuteur sur notre jeu de données spécifique. En revanche, il présente un taux de classification atteignant 82%.

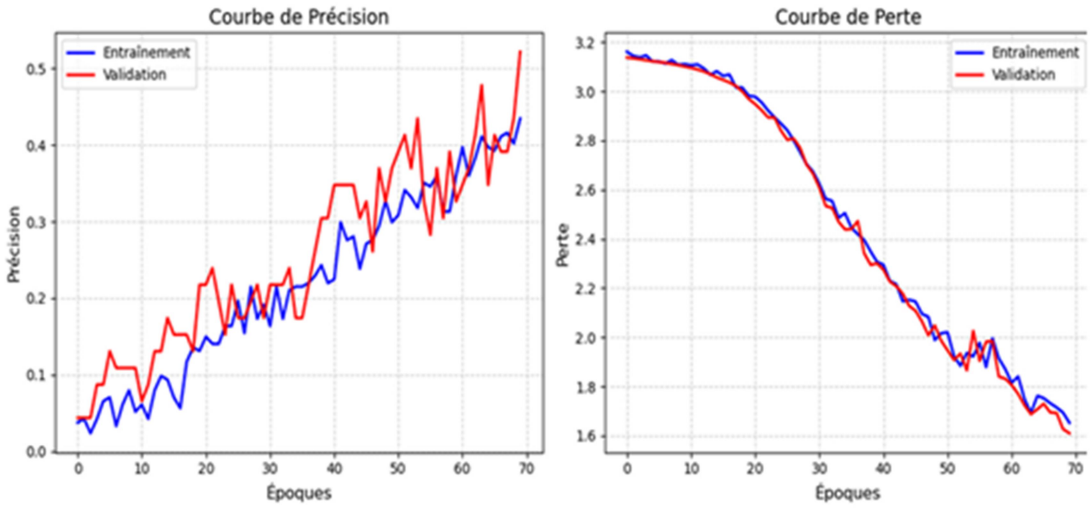

Figure 3.6 : Taux de précision et de perte (entrainement et validation) du modèle RNN GTCC.

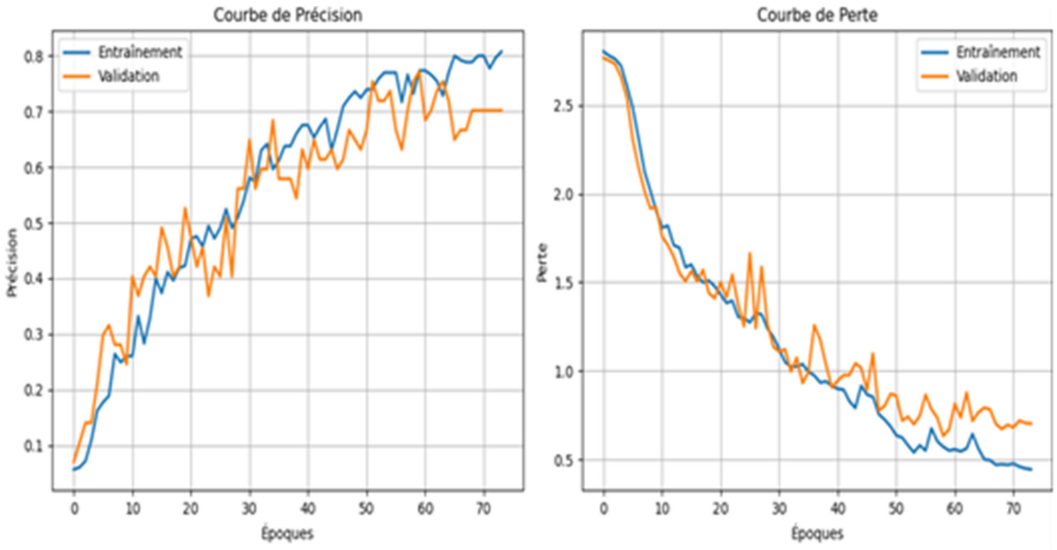

Figure 3.7 : Taux de précision et de pertes de modèle RNN GTCC pour l'ensemble d'enregistrements multilingue.

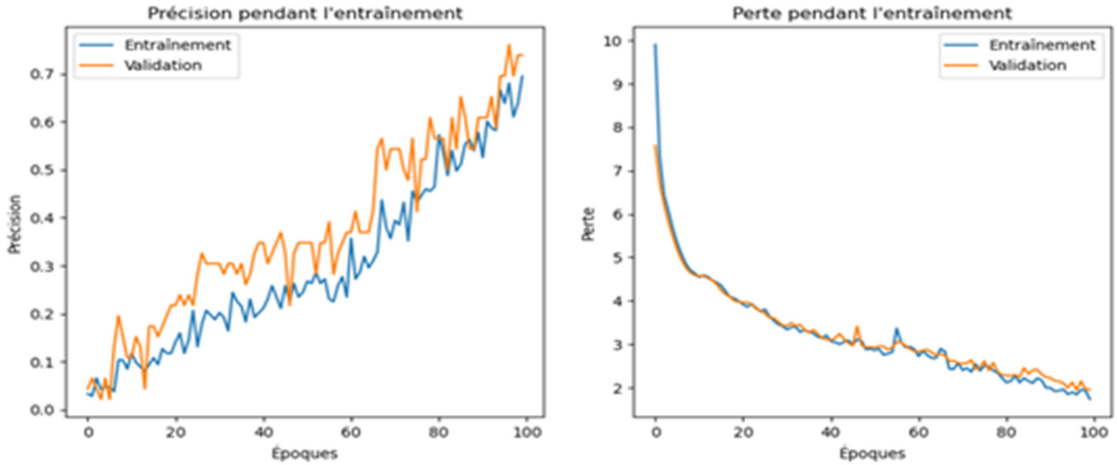

Figure 3.8 : Taux de précision et de perte de modèle CNN pour l'ensemble d'enregistrement monolingue.

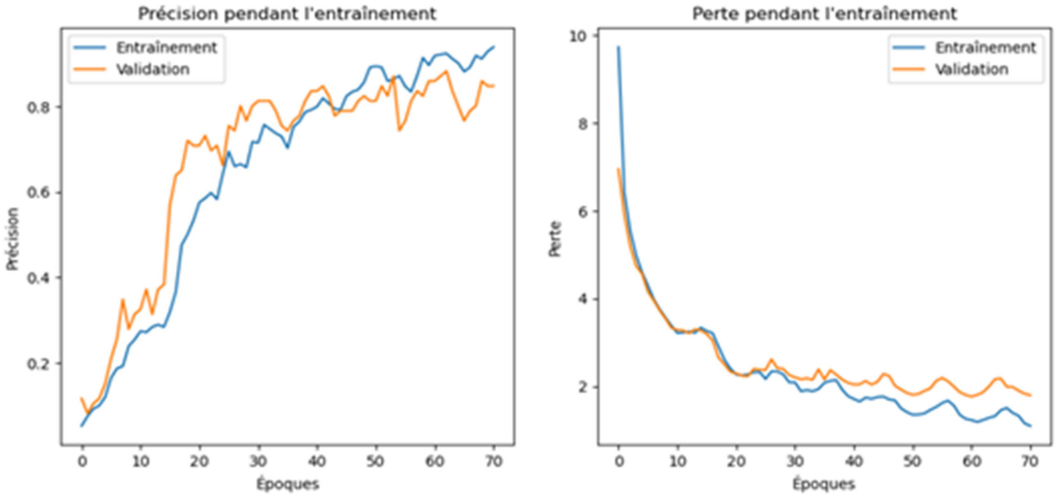

Figure 3.9 : Taux de précision et de perte de modèle CNN pour l'ensemble d'enregistrement multilingue.

3.7.2.Analyse de la matrice de confusion et taux de test

- ✓ La matrice de confusion est un outil clé pour évaluer les performances d'un modèle de classification. Elle compare les prédictions du modèle aux valeurs réelles en affichant les résultats sous forme de tableau à double entrée.
- ✓ Dans ce tableau :
- ✓ Les lignes représentent les classes réelles.
- ✓ Les colonnes représentent les classes prédites.
- ✓ La diagonale montre les prédictions correctes.
- ✓ Les valeurs hors diagonale indiquent les erreurs de classification, révélant les confusions entre les différentes classes.

Les figures 3.10 3.11 3.12 3.13 illustrent respectivement les matrices de confusion des modèles RNN GTCC monolingue et multilingues et CNN GTCC monolingue et multilingue:

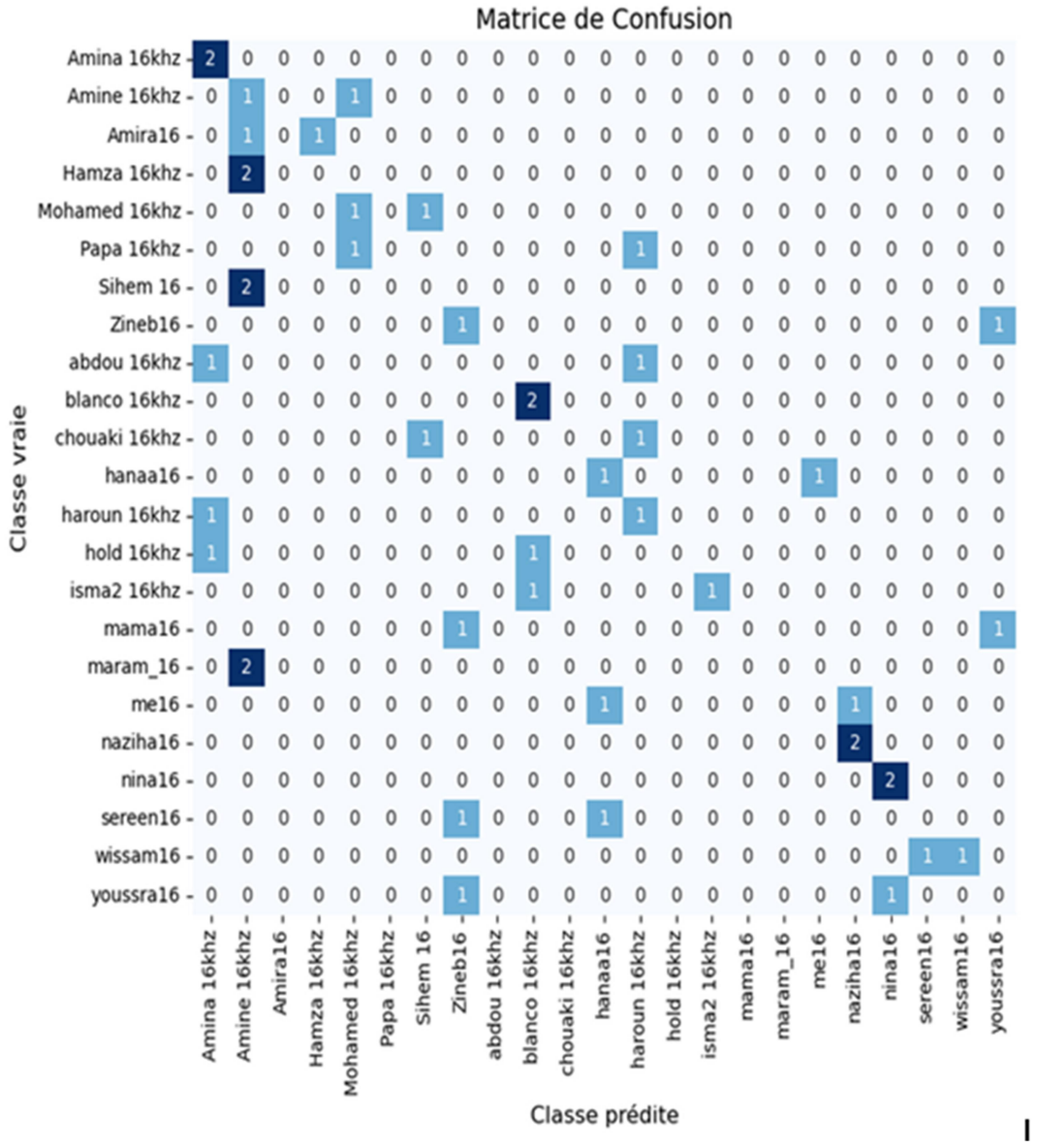

Figure 3.10 : Matrice de confusion RNN GTCC de la base de données monolingue.

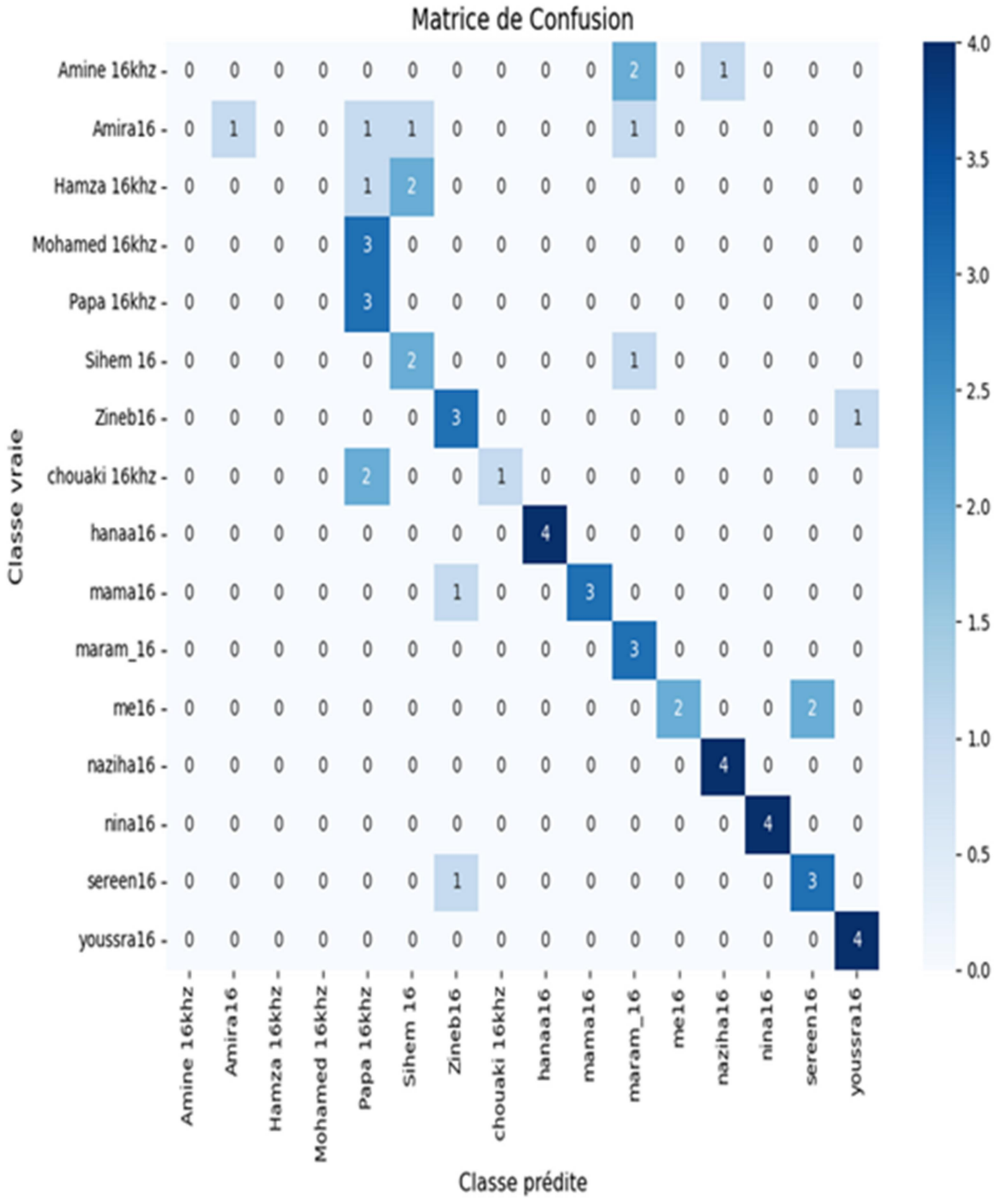

Figure 3.11 : Matrice de confusion RNN de la base de données multilingue.

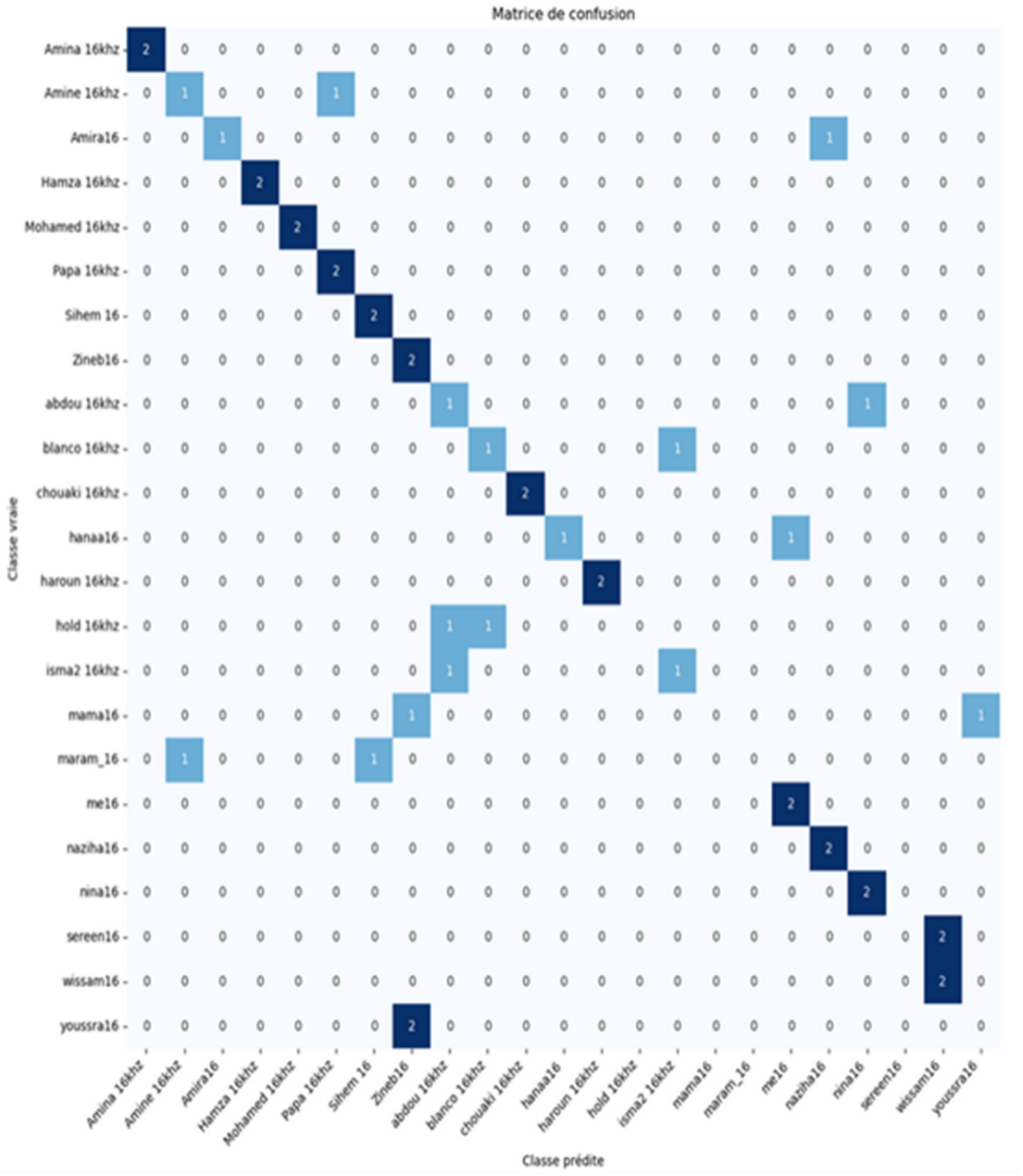

Figure 3.12 Matrice de modèle CNN pour l'ensemble d'enregistrement monolingue.

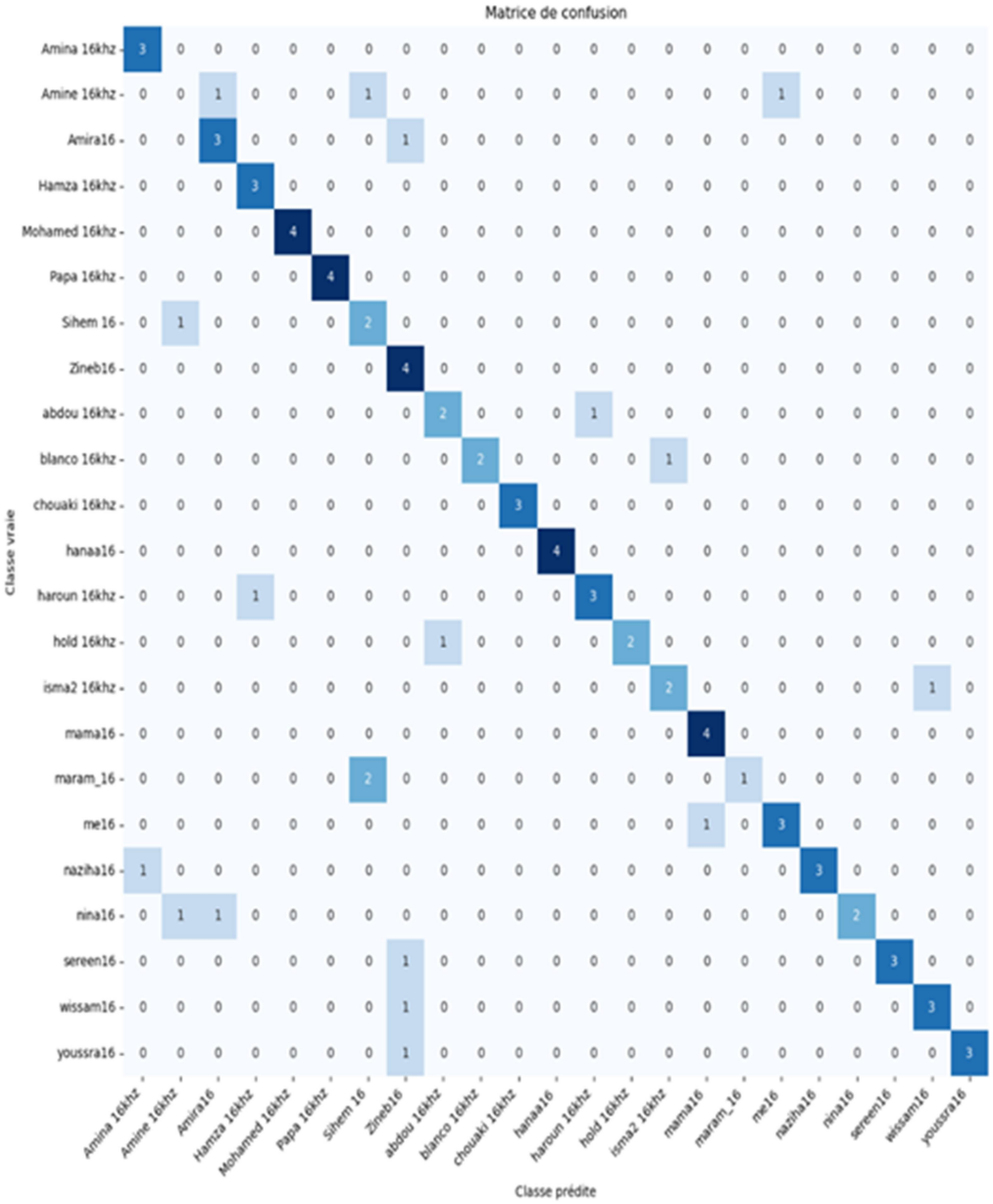

Figure 3.13 : Matrice de modèle CNN pour l'ensemble d'enregistrement multilingue.

Comparativement au modèle RNN GTCC, les matrices de confusion du modèle CNN GTCC montrent moins d'erreurs, indiquant de meilleures performances. De plus, La base de données multilingues appliquée à ce même modèle se distingue particulièrement par ses résultats supérieurs, probablement grâce à la richesse et à la diversité des caractéristiques vocales qu'elle contient, ce qui permet une reconnaissance plus précise des locuteur

II.8. Amélioration du modèle d'identification

Pour améliorer les performances de notre modèle d'identification, nous avons utilisé l'augmentation de données (Data Augmentation), une technique éprouvée.

Cette approche consiste à enrichir notre base de données en appliquant diverses transformations aux enregistrements existants, notamment :

- ✓ La modification de la vitesse de lecture
- ✓ L'ajout de bruit de fond
- ✓ La modification de la hauteur tonale
- ✓ L'ajustement du volume avec différentes amplitudes

Ces transformations ont permis de multiplier par cinq le nombre d'enregistrements, ce qui a considérablement enrichi notre base de données. Nous avons ensuite évalué les performances de notre modèle sur ces ensembles de données augmentés.

III.8.1. Analyse des taux de perte et précision

Après augmentation, **les figures 3.14 3.15 3.16 et 3.17** montrent respectivement les courbes de perte et de précision(entrainement et validation des modèle RNN GTCC monolingue et multilingue , et CNN GTCC monolingue et multilingue:

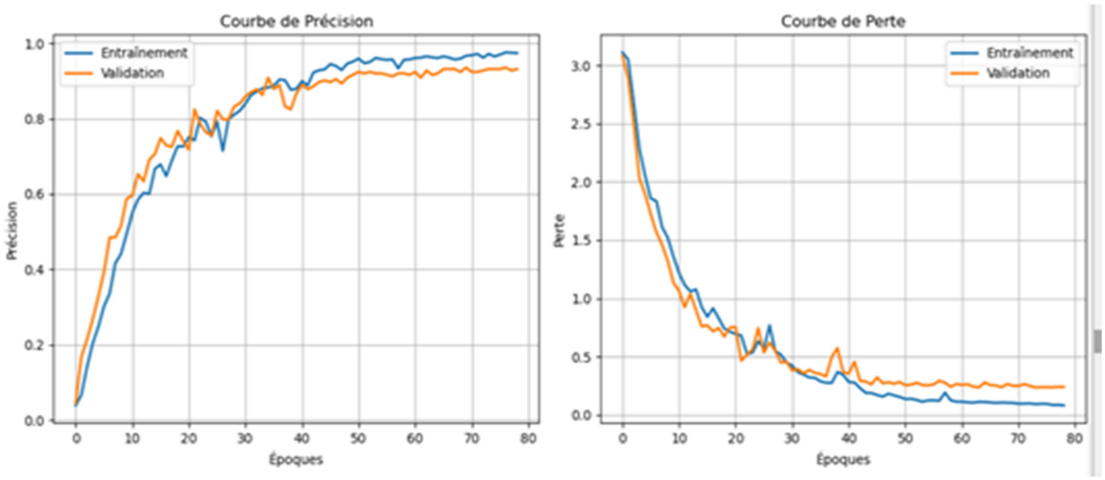

Figure 3.14 : Taux de précision et de perte de modèle RNN pour l'ensemble d'enregistrement monolingue augmentée.

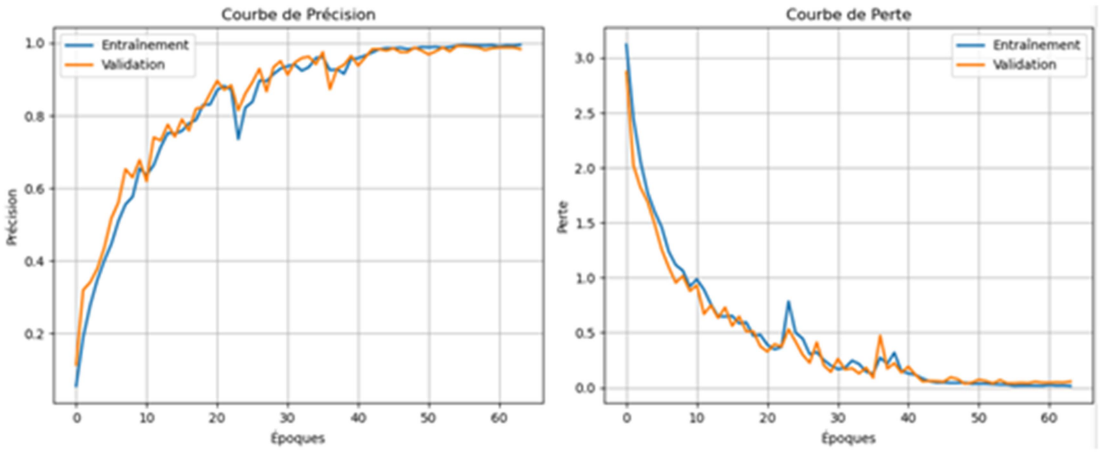

Figure 3.15 : Taux de précision et de perte de modèle RNN pour l'ensemble d'enregistrement multilingue augmentée.

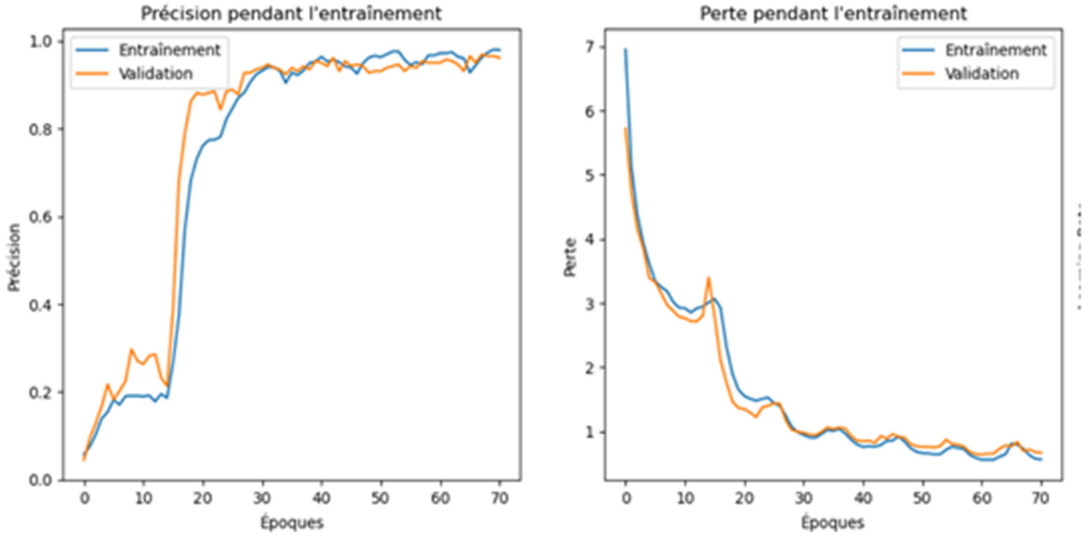

Figure 3.16 : Taux de précision et de perte de modèle CNN pour l'ensemble d'enregistrement monolingue augmentée.

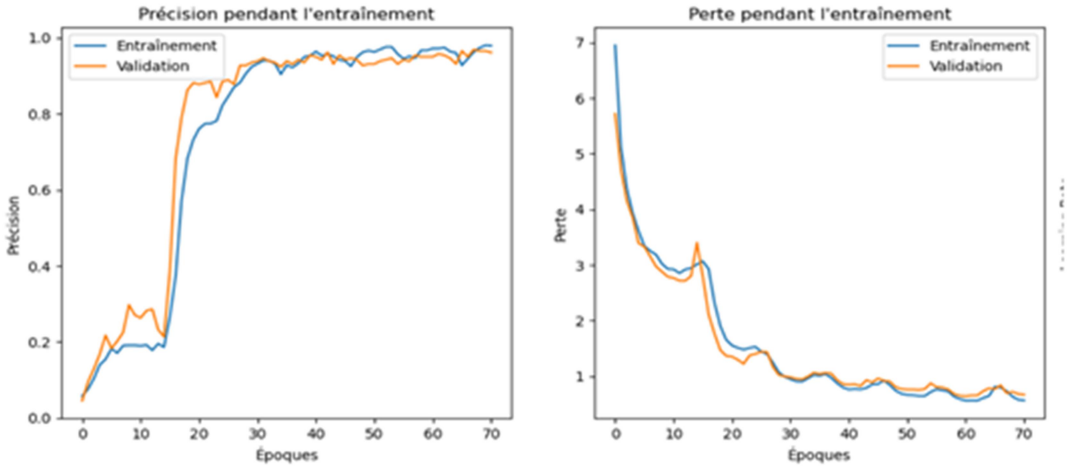

Figure 3.17 : Taux de précision et de perte de modèle CNN pour l'ensemble d'enregistrement multilingue augmentée.

Après avoir augmenté les données, nous observons une amélioration significative des taux de précision, de perte et de classification. Le RNN GTCC montre une amélioration notable des taux de précision de validation et de test.

Le tableau 3.3. montrent les différents taux de précision de validation et de test pour les différents modèles (CNN GTCC et RNN GTCC) et ensembles d'enregistrement (monolingue et multilingue) après augmentation des données :

	RNN GTCC		CNN GTCC	
	Validation	**Test**	**Validation**	**Test**
Monolingue	94,27%	90,46%	94,83%	95,24%
Multilingue	98,75%	97,72%	97,09%	98,13 %

Tableau 3.3 : Taux de précision et test des modèles RNN GTCC et CNN GTCC pour les ensembles d'enregistrement monolingue et multilingue après l'augmentation des données.

D'après le tableau, nous remarquons que : pour la base de données monolingue, le CNN GTCC affiche les meilleures performances. En revanche, pour la base de données multilingue, les deux approches obtiennent des taux de classification similaires, avec un léger avantage pour le CNN GTCC.

III.8.2. Analyse de la matrice de confusion

Les figures 3.18 3.19 3.20 et 3.21 illustrent respectivement les matrices de confusion des modèles RNN GTCC monolingue et multilingues augmentées et CNN GTCC monolingue et multilingue augmentées. Les deux modèles montrent des performances de classification améliorées, avec un léger avantage pour le CNN GTCC par rapport au RNN GTCC.

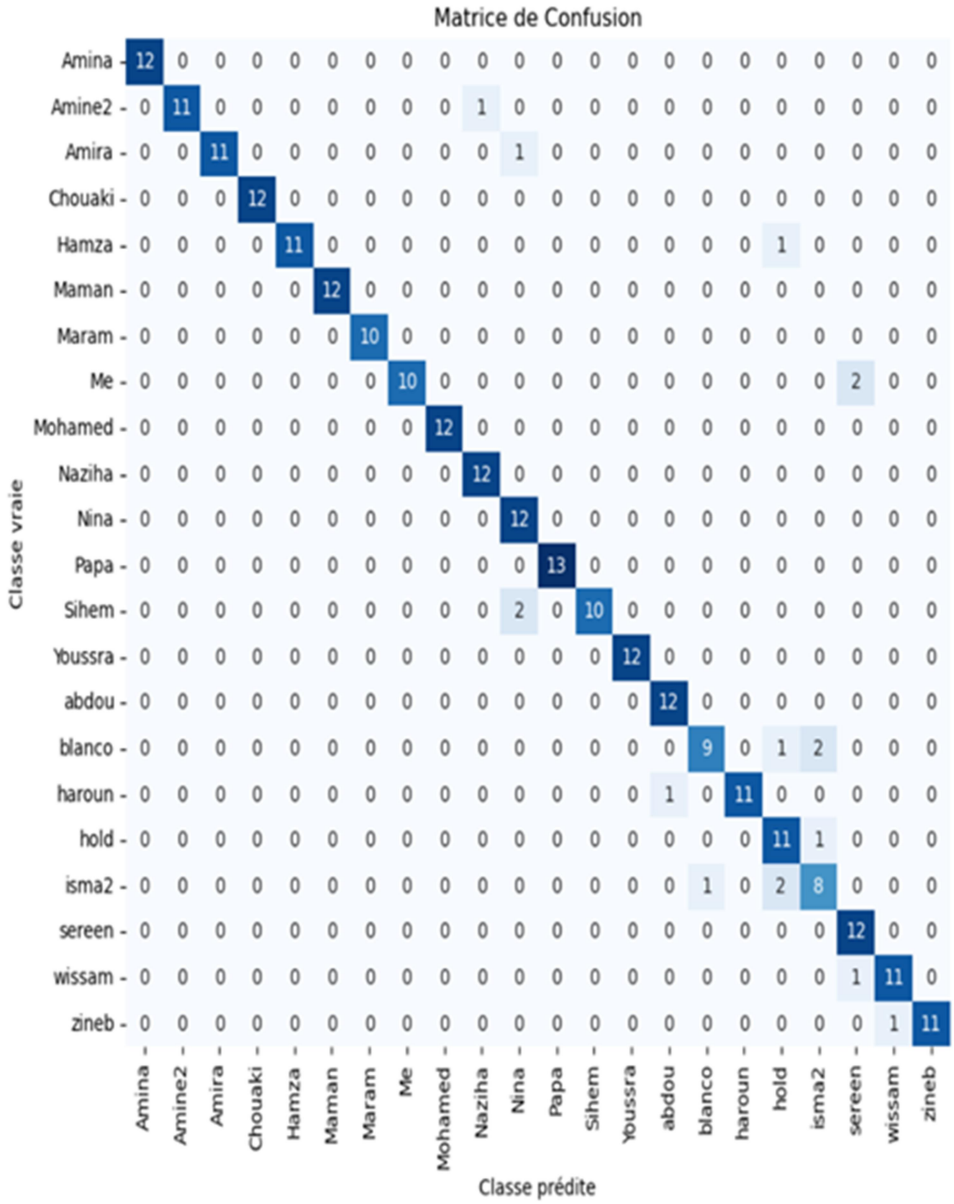

Figure 3.18 : Matrice de confusion RNN GTCC de la base de données monolingue augmentée

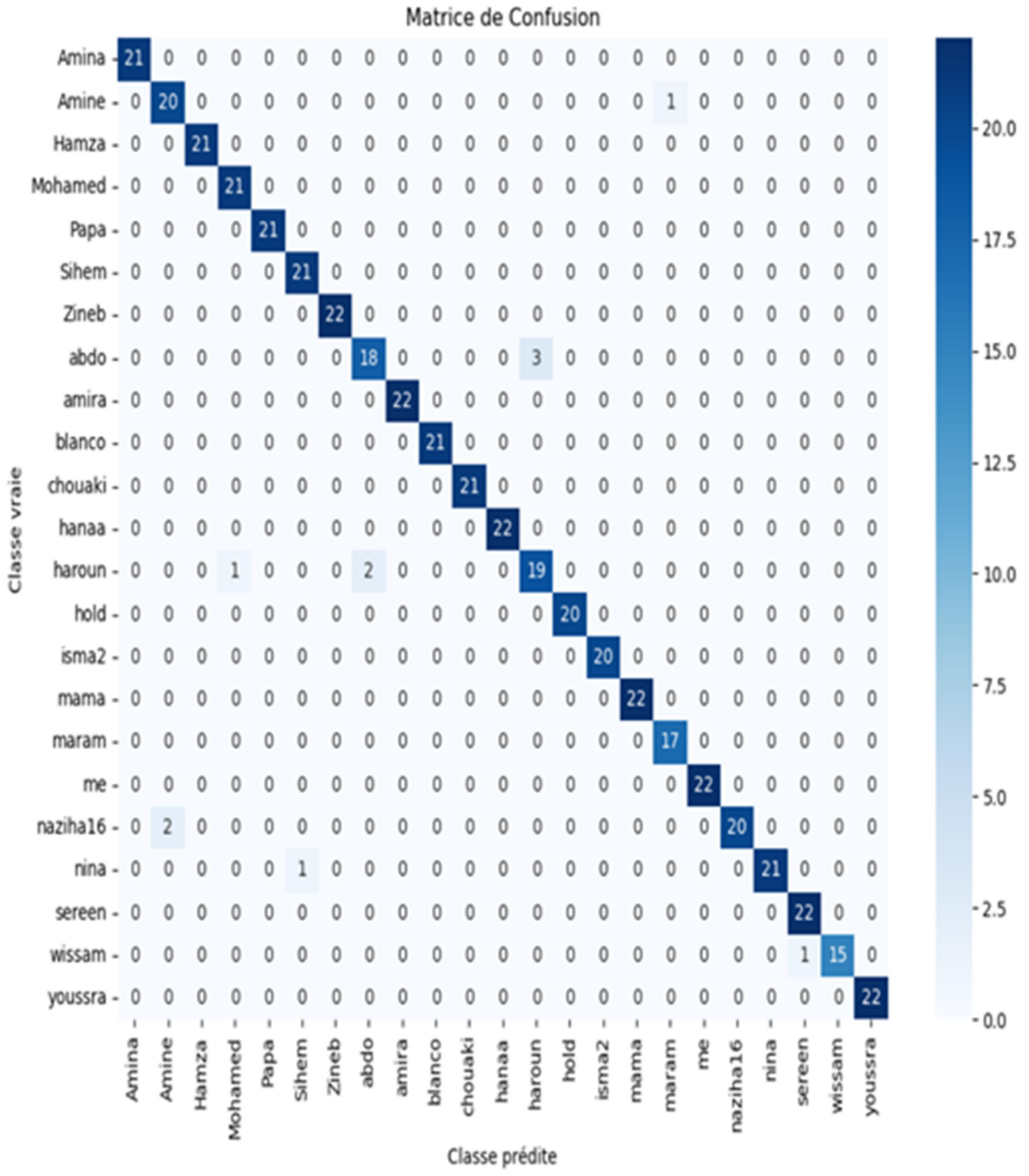

Figure 3.19 : Matrice de confusion RNN GTCC de la base de données multilingue augmentée

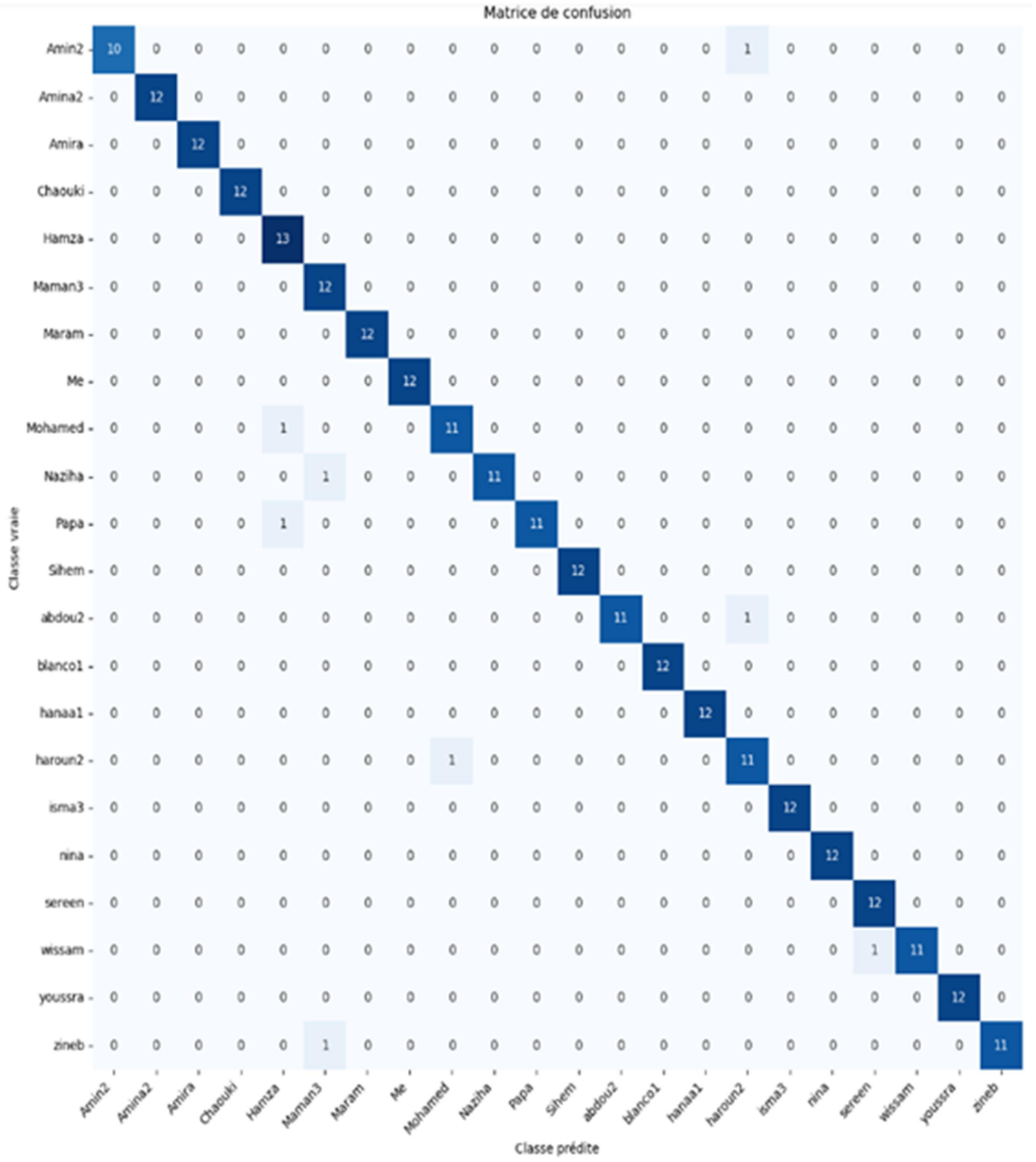

Figure 3.20 : Matrice de confusion CNN de la base de données monolingue augmentée

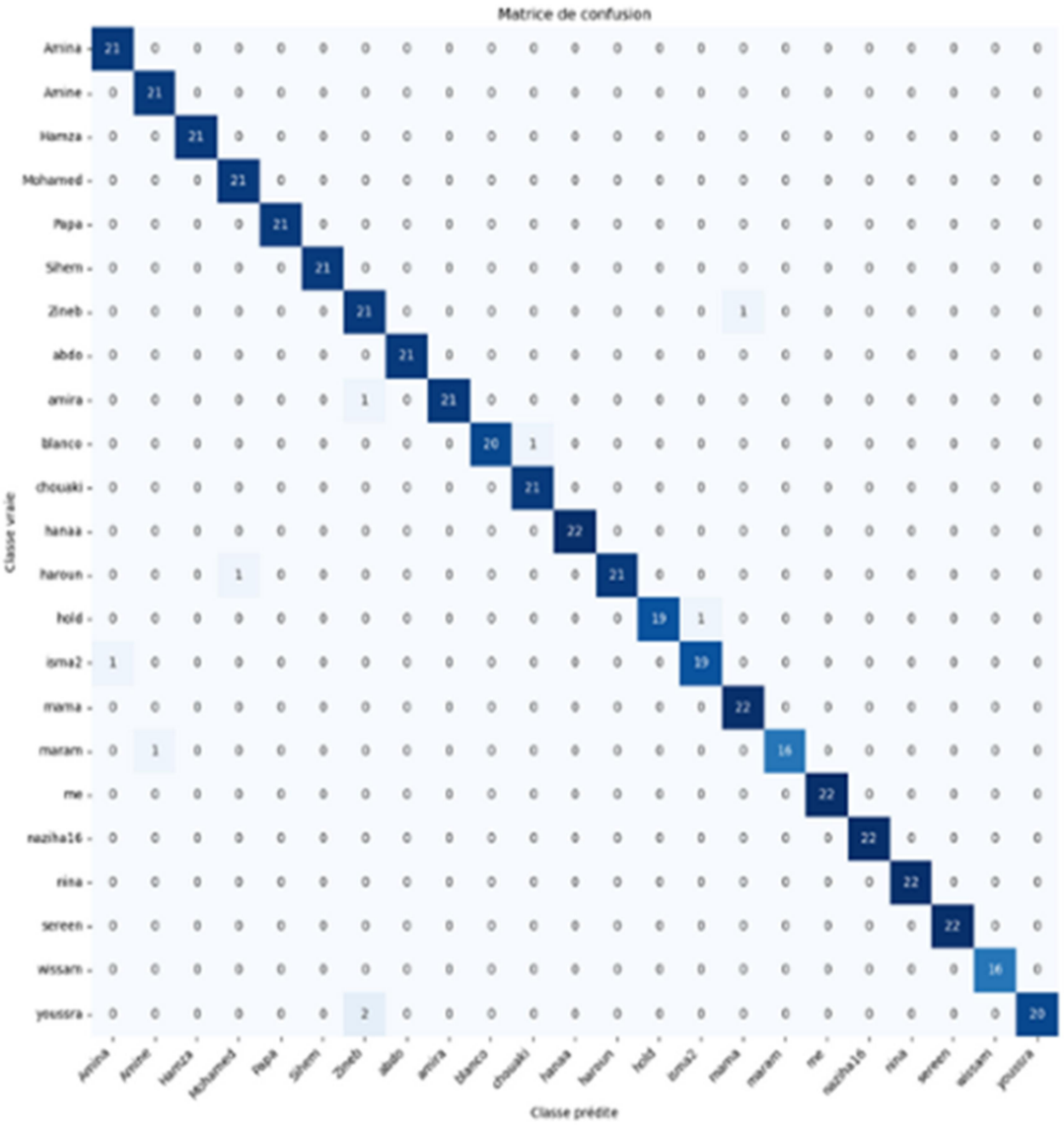

Figure 3.21 : Matrice de confusion CNN de la base de données multilingue augmentée

Conclusion:

Ce chapitre présente une étude comparative entre deux modèles d'identification du locuteur : CNN GTCC et RNN GTCC, appliqués à des corpus vocaux monolingue et multilingue. L'ensemble du processus expérimental, depuis la collecte des données jusqu'à l'évaluation des performances, a été décrit en détail.

Les expériences ont montré que le modèle CNN GTCC offre de meilleures performances que le RNN, particulièrement sur les données multilingues.

CONCLUSION GÉNÉRALE ET PERSPECTIVES

Dans ce travail, nous avons mené une étude comparative approfondie entre deux architectures de réseaux de neurones – le **CNN (Convolutional Neural Network)** et le **RNN (Recurrent Neural Network)** – appliquées à la tâche d'identification du locuteur à partir de caractéristiques audio GTCC. Deux corpus ont été exploités : l'un **monolingue** (dialecte algérien) et l'autre **multilingue** (algérien, français, anglais), afin d'évaluer la robustesse des modèles dans des contextes linguistiques variés.
Les résultats expérimentaux ont mis en lumière plusieurs constats clés :

- **Avant augmentation des données**, les performances globales étaient limitées, en particulier en phase de test, avec un phénomène notable de **surapprentissage** affectant les deux modèles. Le **CNN GTCC** s'est toutefois montré **légèrement supérieur** au RNN en termes de précision sur l'ensemble de validation et de test.
- L'intégration de techniques d'**optimisation** telles que la régularisation L1/L2, le *shuffle*, l'adaptation dynamique du taux d'apprentissage et le *early stopping* a permis d'améliorer la stabilité de l'entraînement et de limiter l'overfitting.
- L'**augmentation de données** (modification de la vitesse, ajout de bruit, changement de pitch, variation de volume) a eu un **impact décisif** sur les performances, permettant d'atteindre des taux de classification très élevés. Après augmentation :
    - Le **CNN GTCC** a atteint jusqu'à **98,13 % de précision** sur les données multilingues.
    - Le **RNN GTCC** a également progressé de manière significative, culminant à **97,72 %** dans le même contexte.
    - Sur les données monolingues, les deux modèles ont dépassé les **90 %** en test, avec un avantage net pour le CNN.

En conclusion, cette étude démontre que l'architecture **CNN GTCC**, enrichie par des techniques de traitement et d'optimisation appropriées, s'avère **la plus efficace** pour l'identification automatique du locuteur, particulièrement sur des jeux de données riches et diversifiés. Ces résultats ouvrent des perspectives prometteuses pour des applications en environnements multilingues ou en conditions variées de parole.

En perspectives, plusieurs pistes d'amélioration et d'exploration futures peuvent être envisagées pour approfondir et renforcer notre système d'identification du locuteur :

1. **Extension à d'autres architectures de Deep Learning** :
   Explorer des modèles plus avancés tels que les **Transformers**, **ResNet** ou **attention-based RNNs** pourrait améliorer la capacité du système à capturer des informations temporelles et spectrales complexes.
2. **Utilisation de représentations audio complémentaires** :
   Intégrer d'autres types de caractéristiques acoustiques, comme les **spectrogrammes**, **PLP (Perceptual Linear Prediction)** ou les **embeddings pré-entraînés** (par ex. x-vectors ou Wav2Vec), pourrait enrichir la représentation des signaux vocaux.
3. **Tests en conditions réelles** :
   Valider la robustesse du système dans des **environnements réels et bruités**, avec des microphones de qualité variable, permettrait d'évaluer la généralisabilité du modèle hors laboratoire.
4. **Identification ouverte** :
   Étendre le système à un **contexte open-set**, où des locuteurs inconnus peuvent apparaître lors du test, représenterait une évolution vers une solution plus proche des besoins du monde réel.

Bibliographie

[1] M. F. Clemente Giorio, Kinect in Motion - Audio and Visual Tracking by Example, Packt Publishing, 2013.

[2] Lawrence Rabiner. Fundamentals of speech recognition.PTR Prentice Hall,1993.7,8,16

[3] S. K. Singh, P.C.Pandey, « Featues and Technique For Speaker Recognition », Seminar Report, page 5,6, Novembre 03.

[4] Yassine Mami. Reconnaissance de locuteurs par localisation dans un espace de locuteurs de référence. Interface homme-machine [cs.HC]. Télécom ParisTech, 2003. Français. ffNNT : tel-00005757

[5] J.P. Haton, C. cerisara, D. Fohr, Y. Laprie, and K. Smaili, "Reconnaissance automatique de la parole: du signal à son interprétation". Paris: Dunod, 2006.

[6] Asmaa Amehraye. Débruitage perceptuel de la parole. PhD thesis, Télécom Bretagne, 2009. 11, 28, 29

[7] Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon. Spoken language processing: a guide to algorithms and system development. Prentice-Hall, 2001. 11, 25

[8] René Boite. Traitement de la parole. PPUR presses polytechniques, 2000. 11

[9] Othman Lachhab. Reconnaissance Statistique de la Parole Continue pour Voix Laryngée et Alaryngée. PhD thesis, Université Mohamed V-Agdal, Ecole Nationale Supérieure d'Informatique et d'Analyse des Systèmes, 2017. 13

[10] H. Satori, M. Harti and N. Chenfour, « Système de Reconnaissance Automatique de l'arabe basé sur CMUSphinx », mémoire master, Dhar Mehraz Fès Morocco.

[11] Mr. Haddab, « reconnaissance automatique du locuteur par la méthode du taux passage par zéro », mémoire master, université Mouloud mamri de Tizi-Ouzou, 2007/2008

[12] Rachedi Julien Mémoire Master 2005 « Reconnaissance et classification de phonèmes ».

[13] Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing, 28(4), 357–366.
https://doi.org/10.1109/TASSP.1980.1163420

[14] Valero, X., & Alias, F. (2012). Gammatone Cepstral Coefficients: Biologically Inspired Features for Non-Speech Audio Classification. IEEE Transactions on Multimedia, 14(6), 1684–1689.
https://doi.org/10.1109/TMM.2012.2213763

[15] M. MOUSS Mohamed Djamel, « Intégration D'un Module De Reconnaissance De La Parole Au Niveau D'un système Audiovisuel – Application Téléviseur », thèse de doctorat, Université Batna 2, AVRIL 2021.

[16] (consulté le 12/06/2023), disponible sur : https://www.editionseni.fr/open/mediabook.aspx?idR=f6e7a7353a3574180124387fa03fdcl,

[17] La Ryax Team, « Deep learning : comprendre les réseaux de neurones artificiels (artificial neural networks) », article, page 3, 2020.

[18] Pr. BILAMI Azeddine, « Apprentissage Incrémental & Machines à Vecteurs Supports », Université HADJ LAKHDAR – BATNA, 18 /12 /2013

[19] Guillaume Saint-Cirgue, « Apprendre la machine learning en une semaine », 2019.

[20] Houcine Noura & Khelifa Nadia, « classification des textures par les réseaux de neurones convolutifs », mémoire master, université mouloud Mammri tizi-ouzou, 2018/2019.

[21] "Réseaux de neurones récurrents." Data Analytics Post, [date non spécifiée]. Disponible sur: dataanalyticspost.com/Lexique/reseaux-de-neurones-recurrents/

[22] TSCHIRHART, Fabien. "Réseau de Neurones Formels Appliqués à l'Intelligence et au Jeu." Mémoire de recherche, sous la direction de M. Alain Lioret, École Supérieure de Génie Informatique, Paris, 2009.

[23] CHRAIBI KAADOUD, Ikram. "Apprentissage de séquences et extraction de règles de réseaux récurrents : application au traçage de schémas techniques." Thèse de doctorat en informatique, sous la direction de Frédéric Alexandre, Université de Bordeaux, 2018.

[24] LeCun et al., 1998 ; krizhevsky et al., 2012

[25] Hochreiter & Schmidhuber,1997 et Graves et al., 2013.