الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne démocratique et populaire

وزارة التعليم العالي والبحث العلمي

Ministère de l'enseignement supérieur et de la recherche scientifique

سعد دحلب جامعة البليدة 1
Université SAAD DAHLAB BLIDA1

كلية التكنولوجيا
Faculté de Technologie

قسم الالكترونيات
Département d'Électronique

MEMOIRE

En vue d'obtention du diplôme de master en

Présenté par :

DOUIDENE SIHEM

&

CHAMBI NOUR EL HOUDA

Filière : Télécommunication

Spécialité : ST

CLASSIFICATION DU LOCUTEUR A L'AIDE DE RESEAUX DE NEURONES CONVOLUTIONNELS ET DES GTCC

Proposé par : Mme N. BOUTALEB

Année Universitaire : 2024-2025

REMERCIEMENTS

Nous tenons tout d'abord à adresser mes remerciements les plus chaleureux à ma promotrice, Madame ***N. Boutaleb****, pour la confiance qu'elle m'a accordée, sa disponibilité constante, la qualité de son encadrement, et la rigueur scientifique dont elle a fait preuve tout au long de ce projet. Son accompagnement m'a été d'un soutien inestimable, tant sur le plan méthodologique qu'humain. Grâce à ses conseils avisés et à son regard critique, elle a su orienter mes réflexions, approfondir ma démarche et m'encourager à viser l'excellence.*

Je souhaite également exprimer toute ma reconnaissance aux membres du jury, ***Monsieur H. Ait Saadi*** *et* ***Monsieur S. Dahmani*** *pour l'honneur qu'ils m'ont fait en acceptant d'évaluer mon travail. Leur expertise, leurs remarques pertinentes et leurs suggestions constructives ont grandement enrichi la qualité de ce mémoire. Je les remercie pour le temps qu'ils ont consacré à la lecture et à l'analyse de ce travail,ainsi que pour l'intérêt qu'ils lui ont porté.*

Mes remerciements vont aussi à l'ensemble des enseignants et intervenants de la formation, qui ont su transmettre leur savoir avec passion, exigence et bienveillance, contribuant ainsi à la richesse de mon apprentissage durant ces années d'études.

DÉDICACES

Je tiens à dédier ce travail :

À toutes les personnes qui ont étémes côtés tout au long de ce parcours, avec amour, patience et bienveillance.

À mon père, ce modèle de sagesse, de courage et de détermination, qui m'a toujours soutenu(e) avec discrétion mais puissance. Ton exemple m'a appris à ne jamais baisser les bras, à aller au bout de mes objectifs, même dans les moments les plus difficiles.

À ma chère mère, source inépuisable de tendresse, de patience et de prières. Merci pour ton amour infini, pour tes encouragements quotidiens, et pour cette force silencieuse que tu m'as transmise à chaque étape. Ton soutien moral et affectif m'a porté(e) plus que tu ne l'imagines.

À mes frères, mes alliés de toujours, pour leur présence rassurante et leurs encouragements qui m'ont donné du courage à chaque étape surtout mon petit ange AMIR.

À mes chères cousines Amira et Marame et chers cousins Hocine et Chaouki, pour leur douceur, leurs encouragements et leur affection sincère. Merci d'avoir été là dans les moments de doute comme dans les moments de joie.

À mon chouchou, K.I, une personne très spéciale à mes yeux. Tu as été là dans les moments durs, toujours à l'écoute, toujours à me motiver, à me pousser à donner le meilleur de moi-même. Merci pour ta patience, ton amour et ta lumière.

À toute ma famille, élargie, présente dans ma vie de près ou de loin, je vous dédie aussi ce travail. Vos encouragements, vos prières, vos mots d'encouragement ont formé un socle solide sur lequel j'ai pu m'appuyer.

Et enfin, une pensée sincère et remplie de gratitude pour mon binôme, Noor, avec qui j'ai partagé cette belle aventure académique. Ton sérieux, ton esprit d'équipe et ta bonne humeur ont rendu ce parcours plus agréable et plus humain. Merci d'avoir été là à chaque étape.

À vous tous, merci du fond du cœur. Ce travail vous appartient autant qu'à moi.

Sihem

Dédicaces

Je tiens à exprimer toute ma gratitude à Allah, le Tout-Puissant, pour m'avoir accordé la force, la patience et la persévérance nécessaires tout au long de ce parcours.

À ma chère mère, source d'amour inconditionnel, de prières silencieuses et de soutien indéfectible : merci pour ta tendresse et ta foi en moi.

À mon père, pilier de sagesse et d'encouragement, merci pour ton exemple, tes sacrifices et tes mots qui m'ont souvent remis sur le bon chemin.

À mes sœurs, pour leur affection, leurs encouragements constants et les moments de complicité qui m'ont tant réconfortée.

À ma binôme, pour sa précieuse collaboration, son esprit d'équipe et les efforts partagés qui ont enrichi ce travail.

Et à ma petite compagne à quatre pattes, ma chatte, dont la présence douce et apaisante a souvent su calmer mes moments de stress.

NOUR

ملخص

تستكشف هذه المذكرة استخدام الشبكات (CNN) ومعاملات سيبسترال غاماماتون (GTCC) لتصنيف المتحدثين تلقائيا. تظهر نتائج التجارب أن الجمع بين CNN GTCC.يقدم نتائج مرضية، حيث يصل إلى معدل 97% على قواعد بيانات متعددة اللغات.

الكلمات المفتاحية CNN GTCC

Résumé

Ce mémoire de master explore l'utilisation des réseaux de neurones convolutionnels (CNN) et des coefficients cepstrauxgammatone (GTCC) pour classifier automatiquement les locuteurs. Les résultats des expérimentations montrent que la combinaison CNN+GTCC présente des résultats satisfaisants, atteignant un taux de 97% sur des bases de données multilingues.

Mots clés : CNN, GTCC

Abstract

This master'sthesis explores the use of Convolutional Neural Networks (CNN) and GammatoneCepstral Coefficients (GTCC) for automatic speaker classification. The experimentalresults show that the combination of CNN and GTCC yieldssatisfactoryresults, achieving a rate of 97% on multilingualdatabases.
Keywords: CNN, GTCC

LISTE DES ACRONYMES

CNN:Convolutional Neural Network

CPU:CentralProcessing Unit

FFT : Fast Fourier Transform

GTCC: GammatoneCepstralCofficients

GPU:GraphicalProcessingUnit

IA:IntelligenceArtificielle

ID :Identification

MFCC:Mel_FrequencyCepstral Coefficients

RAP: Reconnaissance Automatique De la Parole

RAL : Reconnaissance Automatique De Locuteur

ReLu : RectifiedLinear Unit

TD : Texte Dépendent

TI : Texte Indépendent

VAL:Vérification automatique du locuteur

TABLE DES MATIERES

TABLE DES MATIERES

TABLE DES MATIERES

Liste des Figures

Chapitre1 :

Chapitre 2 :

Chapitre 3 :

Liste des Tableaux

INTRODUCTION GENERALE

INTRODUCTION GENERALE

La voix est devenue un outil d'identification aussi essentiel que l'empreinte digitale, et son analyse s'avère particulièrement efficace, notamment dans les applications forensiques. Depuis des décennies, les sciences criminelles reposent sur des méthodes traditionnelles telles que l'analyse auditive et auditive-instrumentale, où certains paramètres acoustiques sont mesurés pour comparer les voix. Cependant, ces approches présentent des limites en raison de leur subjectivité et de leur sensibilité aux variations naturelles de la voix .

L'essor des technologies a permis d'introduire de nouvelles méthodes afin de renforcer ces techniques anciennes, notamment grâce à l'intelligence artificielle et au Machine Learning. Ces systèmes exploitent les caractéristiques uniques de chaque voix pour construire des modèles capables de reconnaître un individu avec une précision accrue. Toutefois, leur performance peut être affectée dans des environnements bruités.

Inspirés des principes fondamentaux des réseaux de neurones (précédemment utilisés en Machine Learning), les réseaux de neurones profonds, en particulier les réseaux convolutifs (CNN : Convolutional Neural Network), ont profondément transformé le domaine de l'intelligence artificielle. Ces modèles offrent non seulement des performances améliorées, mais ils présentent également une meilleure robustesse face aux bruits, une adaptabilité aux différences linguistiques et une précision accrue .

L'objectif de notre projet de fin d'études est de concevoir un système d'identification automatique des locuteurs efficace, intégrant les réseaux de neurones et les GTCC (Gammatone-FrequencyCepstral Coefficients), afin qu'il puisse être utilisé dans des contextes forensiques exigeants, tout en gérant la diversité des enregistrements et des langues. Ce mémoire s'articule autour de trois chapitres :

Dans le premier chapitre. Nous introduisons les bases scientifiques de l'identification vocale. Aprèsavoir expliqué le fonctionnement de la parole humaine, nous passons en revue les méthodes traditionnelles en sciences forensiques, notamment les approches auditives, auditive-instrumentales et automatiques. Ce chapitre établit ainsi le cadre de la problématique en exposant les défis et les limites des approches classiques dans un contexte judiciaire

Ce deuxième chapitre met en avant la révolution apportée par l'intelligence artificielle dans la reconnaissance du locuteur. Nous y présentons les principes fondamentaux du Machine Learning et du Deep Learning, avec un focus sur les réseaux de neurones, plus spécifiquement les CNN. Leur capacité à extraire automatiquement des caractéristiques complexes des signaux audio est mise en lumière, démontrant leur potentiel pour l'identification vocale.

Ce dernier chapitre est consacré à la partie expérimentale de notre travail. Nous détaillons la collecte de notre propre base de données en dialecte algérien,francais et anglais, la conception de notre modèle CNN basé sur les GTCC, ainsi que les différentes étapes d'entraînement et d'évaluation. Nous analysons l'impact de divers paramètres (fréquence d'échantillonnage, nombre d'époques, fonctions d'activation, optimiseurs) sur les performances du système. Grâce à la technique telle que l'augmentation de données , nous avons atteint des niveaux de précision élevés, renforçant ainsi la robustesse de notre modèle.

CHAPITRE 1

RECONNAISSANCE DE LOCUTEUR

1.1 Introduction

L'identification d'une personne par sa voix devient un enjeu majeur pour l'authentification et la sécurité. Dans ce chapitre, nous allons explorer la reconnaissance du locuteur en adoptant une approche globale. Nous commencerons par examiner le processus de production de la parole afin de mieux comprendre comment les caractéristiques vocales permettent d'identifier un individu. Ensuite, nous passerons en revue l'état actuel de la reconnaissance du locuteur, en analysant son architecture, ses différentes applications et ses principales branches. Enfin, nous aborderons la reconnaissance de la parole et du langage, qui complètent ces avancées technologiques.

1.2 Parole

La parole constitue le mode de communication le plus naturel au sein de toute société humaine. Elle se définit comme un signal réel, continu, d'énergie finie et non stationnaire, produit par l'appareil vocal humain. Grâce à elle, les individus peuvent établir une communication fluide et compréhensible [1].

Dans le domaine du traitement du signal, la parole occupe une place centrale en raison de ses caractéristiques acoustiques uniques, directement liées aux mécanismes spécifiques de sa production .

1.2.1 Production de la parole

La production de la parole repose sur un processus complexe impliquant l'activation coordonnée de nombreux muscles et articulateurs, dont le fonctionnement est précisément synchronisé et contrôlé par le système nerveux.

Dans ce chapitre, nous nous intéressons à ce système afin de mieux comprendre le rôle des principaux articulateurs intervenant dans la production de la parole. Nous présenterons une vue d'ensemble des éléments constituant le système vocal, suivie d'une brève description du système de phonation. Cette dernière nous permettra de caractériser et d'identifier les grandes classes de sons élémentaires, ainsi que d'expliquer leurs variations. Enfin, nous conclurons par une explication du principe d'acquisition du signal de parole.

1.2.2 Mécanisme de la production de la parole :

1. L'appareil vocal : une machine à produire des sons

L'appareil vocal humain est capable de générer une grande diversité de sons grâce à unensemble d'organes qui travaillent en synergie. Il comprend les poumons, le larynx et le conduit vocal, qui se divise en plusieurs parties : le pharynx, la cavité buccale et lecavités nasales.

2. Les poumons : le moteur de la parole

Les poumons jouent un rôle essentiel dans la production des sons en fournissant l'énergie nécessaire sous forme d'air. Ils agissent comme un générateur qui alimente le larynx via la trachée. Bien que cette dernière ne participe pas activement à la production sonore, elle assure la liaison entre les poumons et le larynx.

3. Le larynx : le vibrateur naturel

Situé au milieu du cou, le larynx est composé de plusieurs muscles mobiles entourant une cavité placée à l'extrémité supérieure de la trachée.

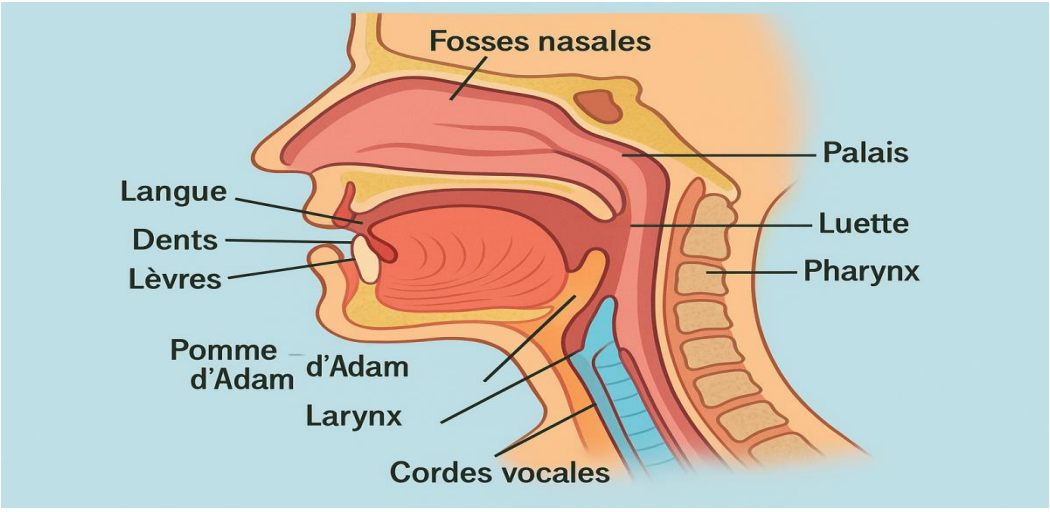

Figure1.1 : Coupe de l'appareil phonatoire humaine [1]

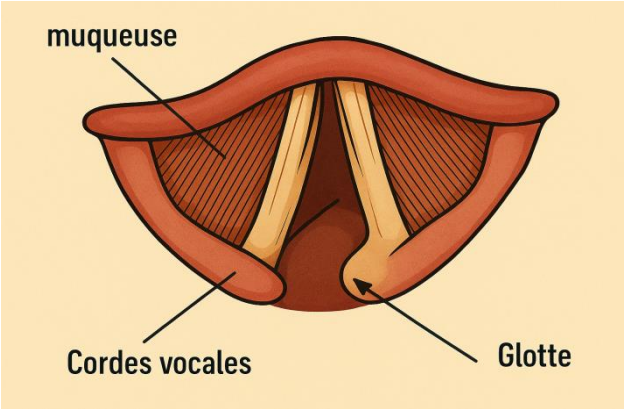

Figure 1.2 : Vue de haut du larnyx[1]

À l'intérieur du larynx se trouvent les cordes vocales, qui sont en réalité des muscles et des ligaments capables de s'ouvrir et de se fermer. Lorsqu'elles s'écartent, elles forment une ouverture triangulaire appelée glotte. La taille de cette ouverture influence le flux d'air qui traverse le conduit vocal et détermine la nature du son produit. En vibrant sous l'effet de l'air, les cordes vocales jouent le rôle d'un véritable générateur de son.

4. Le conduit vocal : le sculpteur des sons

Le conduit vocal agit comme un résonateur, modulant les sons produits par le larynx. Il est constitué de plusieurs cavités, notamment la cavité pharyngo-buccale et la cavité nasale. Les mouvements des muscles articulatoires – mâchoires, langue, voile du palais et lèvres modifient la forme du conduit vocal et influencent des sons. Ce système permet non seulement de produire des sons distincts, mais aussi des nuances grâce à deux fonctions principales : la génération de bruit et la résonance.

Ainsi, la parole résulte d'une coordination complexe entre la respiration, la vibration des cordes vocales et la modulation du conduit vocal, donnant naissance à une infinité de sons et de nuances expressives[2].

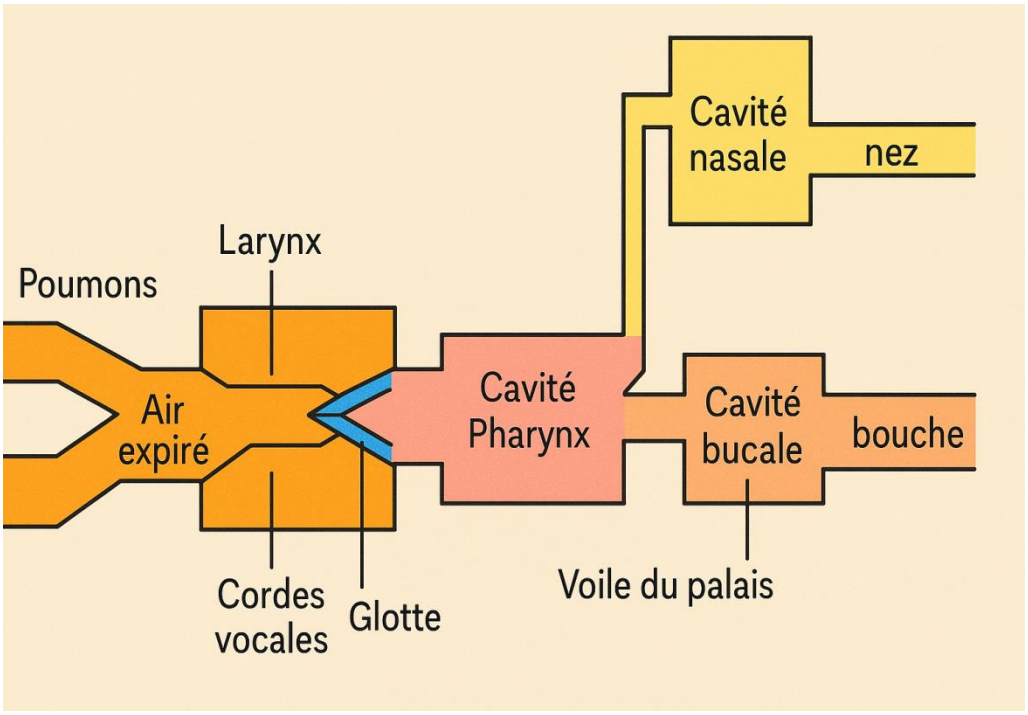

Figure 1.3: Schéma simplifié de l'appareil phonatoire[2]

La production de la parole est un processus qui débute par la formulation d'un message linguistiqueetsetransformeenuneséried'actionsmotricesimpliquantdiversespartiesducorps humain. Ce processus aboutit à la création d'un signal vocal intelligible. On peut le diviser en trois étapes distinctes selon les travaux de Brown et Haooort (2000) ainsi que Blanketa .(2002) [4].

a) La conceptualisation

Cette première étape transforme l'intention de communiquer en une élaboration d'idées et de concepts clairs, correspondant au message que l'on souhaite transmettre.

b) La formulation

Ensuite, le message prend une forme linguistique. Cela implique de sélectionner les mots adaptés, de construire des phrases grammaticalement correctes, de diviser les mots en syllabes et de leur attribuer une structure phonétique.

c) L'articulation et l'exécution motrice de la parole

Ensuite, le message prend une forme linguistique. Cela implique de sélectionner les mots adaptés, de construire des phrases grammaticalement correctes, de diviser les mots en syllabes et de leur attribuer une structure phonétique.

La production de la parole est une symphonie complexe et parfaitement synchronisée**, où diverses parties du corps collaborent sous l'impulsion de mécanismes de contrôle sophistiqués, aboutissant à une communication fluide et compréhensible.

1.2.3 Paramètres du signal de parole

Le signal de parole est un phénomène dynamique, continu, et limité en énergienantes. Sa structure évolue dans le temps et peut être représentée sous forme de signal analogique.

Analyser ce type de signal s'avère complexe, car de nombreux paramètres sont impliqués. Cependant, trois caractéristiques acoustiques essentielles se distinguent : [3]

a) La fréquence fondamentale

La fréquence fondamentale désigne le rythme d'ouverture et de fermeture des cordes vocales. On l'appelle également la hauteur de la voix. Elle est influencée par la taille du larynx : un enfant, possédant un larynx plus petit, a une voix aiguë. En comparaison, les femmes et les hommes ont des larynx de tailles différentes, ce qui explique leurs tonalités variées. Cette fréquence ne concerne que les sons sonores et peut fluctuer en fonction des individus et des situations.

-

Plages de fréquences vocales :

Les voix humaines se répartissent dans des plages de fréquences bien distinctes [5] :

Catégorie	Plage fondamentale	Plage harmonique
Homme	85-180hz	Jusqu'à 8khz
Femme	165-255hz	Jusqu'à 10khz
Enfant	250-400hz	Jusqu'à 12khz

Tableau 1.1 : Plagedefréquence de voix humaine [5]

b) Le spectre spatial

Le spectre 4 représente la répartition des fréquences d'un signal sonore. En termes simples, il s'agit d'une sorte de carte qui montre quelles fréquences sont présentes dans un signal. Cette carte permet notamment de distinguer chaque personne grâce au timbre unique de leur voix [3].

c) L'énergie

L'énergie sonore reflète l'intensité du signal vocal. Elle est généralement plus forte lorsque les sons sont "voisés" (comme les voyelles), par opposition aux sons "non voisés" (comme certaines consonnes). Par exemple, le mot "Tashghil", qui signifie "Allumer", illustre cette différence dans le signal sonore [3].

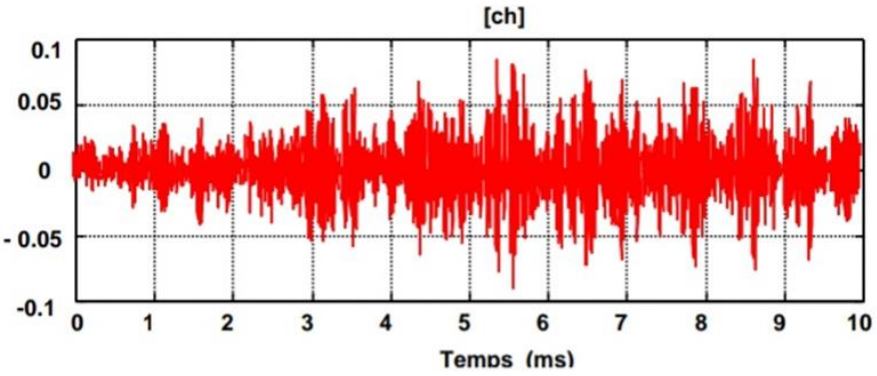

Figure1.4 Signal voisé [3]

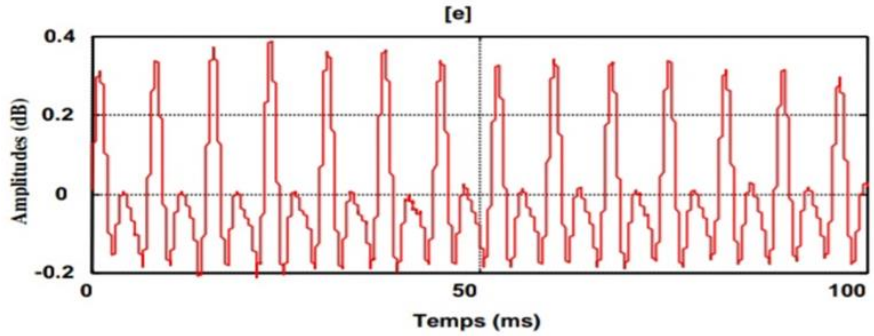

Figure1.5 Signalnon voisé [3]

1.3 Reconnaissance

Dans une perspective générale, la reconnaissance désigne la faculté d'identifier, de comprendre ou de différencier un élément, une idée ou une expérience. Toutefois, ce concept prend des significations variées en fonction du contexte ou du domaine d'application. Par exemple, en sciences cognitives, elle peut se rapporter au processus par lequel le cerveau distingue des formes ou des sons. En linguistique, elle peut évoquer la capacité de comprendre ou de déchiffrer des structures langagières. Ainsi, la reconnaissance est une notion qui s'inscrit dans une multitude de disciplines, enrichissant chaque domaine par ses spécificités et applications [6].

1.3.1 Reconnaissance de parole RAP

La reconnaissance automatique de la parole est une technologie informatique conçue pourpermettre à un logiciel d'interpréter la parole humaine de manière naturelle. Elle vise à extraire le contenu verbal encapsulé dans un signal vocal et à le convertir en une séquence de mots ou de phonèmes, reflétant ainsi les propos de l'utilisateur. Cette technologie s'appuie sur

des approches avancées issues du traitement du signal et de l'intelligence artificielle [6].

Ces dernières années, les systèmes de reconnaissance vocale ont connu des progrès significatifs grâce aux avancées en apprentissage automatique, en particulier dans le domaine du deeplearning. Ces innovations ont permis d'améliorer considérablement la précision et la capacité des systèmes à interpréter la parole, y compris dans des environnements bruyants ou face à divers accents. Cependant, malgré ces progrès, des limitations subsistent : certains mots ou accents demeurent difficiles à interpréter, rendant parfois nécessaire une révision ou une correction manuelle.

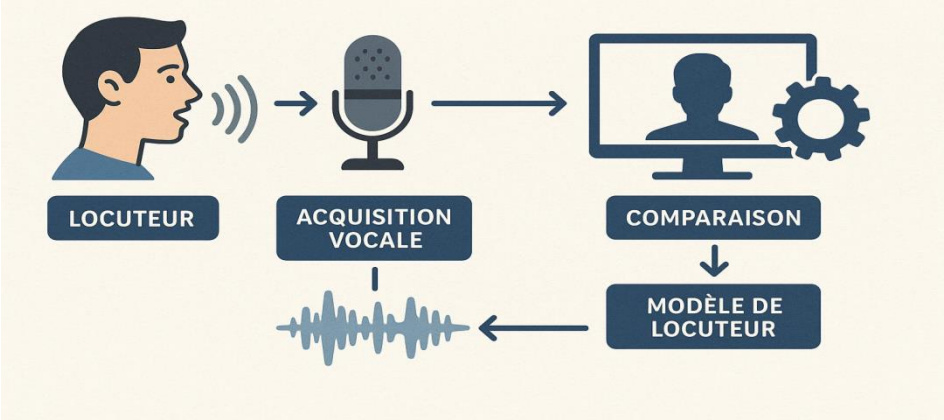

Figure1.6 Structure de base d'un système RAP

1.3.2 Applications d'un système RAP

Les applications de la reconnaissance automatique de la parole (RAP) sont variées et dépendent de leur catégorie. Grâce aux avancées des techniques de reconnaissance automatique du langage (RAL), ces systèmes ont considérablement gagné en efficacité. Par ailleurs, la plupart des systèmes RAL se divisent en deux catégories : ceux dépendants du texte et ceux indépendants du texte. Les systèmes dépendants du texte exigent une phase d'apprentissage, nécessitant souvent

de nombreuses heures de données vocales. Cette section présente de manière succincte les quatre principaux types de systèmes de reconnaissance
vocale[8].

Cette partie décrit brièvement les quatre principaux types de systèmes de reconnaissance vocale.

A) Commandes vocales

Les systèmes de reconnaissance de la parole trouvent leur application dans divers assistants vocaux comme Alexa ou Google Assistant. Ils permettent aux utilisateurs de piloter des appareils électroniques ou d'accéder rapidement à des informations grâce à des commandes vocales intuitives.

B) Applications de sécurité

Dans le domaine de la sécurité, ces systèmes sont employés pour reconnaître et authentifier les voix autorisées, offrant ainsi une méthode efficace pour sécuriser des espaces sensibles contrôlés.

C) Systèmes de dictée automatique

La dictée de texte figure parmi les applications les plus répandues de la reconnaissance vocale. Elle permet de convertir des paroles en texte écrit avec précision, facilitant ainsi la transcription dans divers contextes professionnels et personnels.

D) Systèmes de compréhension

Essentiellement, ils permettent de dialoguer avec une machine. Par conséquent l'utilisateur prononce une série de mots-clés que le système est peu de reconnaitre [9].

1.3.3 Reconnaissance automatique du locuteur RAL

La reconnaissance automatique du locuteur se distingue comme une sous-discipline de la reconnaissance de formes, dédiée à l'identification d'une personne à partir de son empreinte vocale unique. Cette spécialisation s'appuie sur les variations vocales propres à chaque individu, essentielles pour différencier une voix parmi plusieurs. Contrairement à la reconnaissance automatique de la parole qui s'intéresse aux aspects linguistiques, cette approche se concentre

sur les données non verbales contenues dans les signaux vocaux.

Profitant des avancées en reconnaissance de la parole, la reconnaissance automatique du locuteur intègre des algorithmes tels que les réseaux de neurones, les machines à vecteurs de support et les modèles de Markov cachés. La performance de ces techniques dépend cependant de facteurs comme la qualité des enregistrements audio, le bruit ambiant et les différences interpersonnelles. Malgré ces obstacles, cette technologie continue d'évoluer, ouvrant la voie à de nouvelles applications dans le monde technologique.

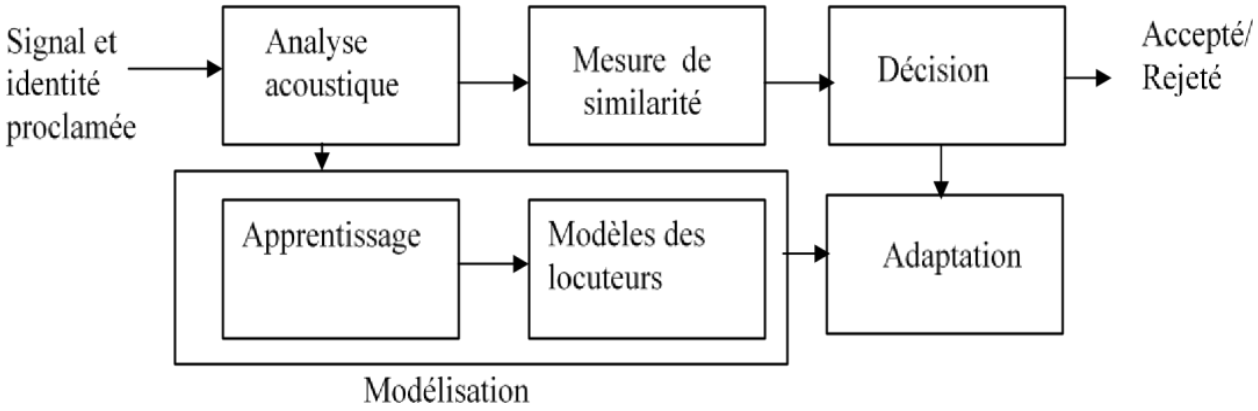

Figure1.7 Schéma d'un système RAL [11]

1.3.4 Différentes tâches en reconnaissance du locuteur :

La reconnaissance automatique du locuteur repose sur deux grandes missions : l'identification et la vérification [5].

A) Identification du locuteur

L'identification vise à déterminer qui parle parmi un ensemble d'individus. Elle peut se faire selon deux approches :

- **Identificationenensemblefermé** : Le système cherche à reconnaître le locuteur parmi un groupe prédéfini de personnes.

- **Identificationenensembleouvert** : Ici, le locuteur peut être une personne inconnue du système, ce qui permet d'identifier un nouvel individu n'ayant jamais été enregistré auparavant.

B) Vérification du locuteur

La vérification consiste à confirmer ou infirmer l'identité annoncée par une personne. Le système compare la voix du locuteur à une référence préenregistrée et prend une décision binaire : soit il valide l'identité revendiquée, soit il la rejette [5].

1.3.5 Modes d'élocution d'un système RAL

Tout comme en reconnaissance automatique de la parole, la manière dont un utilisateur s'exprime influence le fonctionnement du système. En reconnaissance du locuteur, plusieurs modes d'élocution sont possibles [12] :

A) Mode indépendant du texte :

Dans ce mode, l'identification ou la vérification ne dépend pas du contenu des phrases prononcées. Le système analyse uniquement les caractéristiques vocales du locuteur, sans tenir compte des mots utilisés [12].

B) Mode dépendant du texte :

Ce mode impose au locuteur de prononcer une phrase ou un mot clé défini à l'avance par le système. Selo Système le niveau de contrainte, plusieurs variantes existent [12] :

- **Systèmeàtextelibre** : L'utilisateur parle librement, sans contrainte sur le contenu de ses phrases. Les phrases utilisées pour l'apprentissage et les tests sont différentes.

- **Systèmeàtextesuggéré** : Le système propose une phrase différente à chaque session et pour chaque utilisateur. Les phrases d'apprentissage et de test varient.

- **Systèmedépendantduvocabulaire** : Le locuteur doit s'exprimer en utilisant un vocabulaire restreint. L'apprentissage et les tests se basent sur des phrases construites à partir de ce vocabulaire.

- **Systèmepersonnaliséavecmotdepassefixe** : Chaque utilisateur possède un mot de passe vocal unique. Le système apprend et vérifie l'identité en se basant sur ce même mot de passe.

1.3.6 Domaine d'applications d'un système RAL :

La reconnaissance automatique vocal du locuteur trouve des applications variées selon le contexte et

les besoins [11] :

A) Applications sur site

Ces usages nécessitent la présence physique de l'utilisateur dans un lieu spécifique :

- **Serruresvocales** : Contrôle d'accès à des locaux sécurisés ou à des comptes informatiques via l'authentification vocale.

- **Interactionavecdesdispositifsmatériels** : Par exemple, validation d'identité pour le retrait d'argent à un guichet automatique.

B) Applications dans les télécommunications

La vérification de l'identité s'effectue à distance :

Accèssécuriséauxservicesabonnés : Permet aux utilisateurs d'accéder à des plateformes réservées ou à des données confidentielles.

- **Transactionsàdistance** : Authentification vocale pour sécuriser des opérations financières ou administratives.

C) Applications commerciales

L'identification vocale peut être utilisée pour renforcer la sécurité et simplifier l'accès à certains services :

- **Motdepassevocalpartagé** : Un groupe restreint d'utilisateurs (famille, entreprise) peut utiliser une authentification vocale commune.

- **Protectioncontrelevol** : Sécurisation des dispositifs électroniques ou des objets de valeur grâce à la reconnaissance du locuteur.

D) Applications judiciaires

La reconnaissance vocale joue un rôle clé dans les enquêtes et le système judiciaire :

- **Identification** de **suspects** : Analyse vocale pour aider à l'identification et à la collecte de

preuves.

- **Outilsd'aideauxprofessionnelsdudroit** : Juges, avocats et enquêteurs peuvent utiliser ces technologies pour renforcer des preuves ou confirmer un verdict.

E) Applications stratégiques

Certains usages relèvent de la surveillance et de la sécurité nationale :

-Surveillancetéléphonique : Détection et identification de voix dans des contextes sensibles.

- **Protectiondesprincipesdémocratiques** : Prévention des menaces et des actes criminels.

- **Respectdelavieprivée** : Nécessité d'un encadrement strict pour éviter les abus et les intrusions injustifiées.

1.3.7 Structure de base de la RAL :

Un système de reconnaissance des locuteurs repose généralement sur trois composantes essentielles .Comme illustré dans la figure 1.8, la première étape concerne le traitement frontal,où le signal vocal capté est numérisé. À ce stade, des caractéristiques sont extraites discours. Il convient de souligner qu'aucune caractéristique unique ne permet d'identifier exclusivement un locuteur à partir d'un signal vocal. Toutefois, selon la théorie du filtre source en production de la parole, la forme du spectre vocal contient des informations cruciales. Ces

informations incluent la configuration du conduit vocal, déduite des formants, ainsi que la source glottale, identifiable grâce aux harmoniques de la hauteur [2].Ainsi, les caractéristiques spectrales sont largement exploitées dans les systèmes de reconnaissance des locuteurs.

La dernière phase du traitement frontal consiste à compenser les effets du canal. En effet, les variations des dispositifs d'entrée (tels que les combinés téléphoniques) peuvent altérer les caractéristiques spectrales du signal, notamment par des effets comme la limitation de bande ou la modification de la forme spectrale. Pour corriger ces perturbations, des techniques de compensation

sont appliquées, telles que la soustraction moyenne cepstrale à court et long terme, permettant d'atténuer ces distorsions et d'améliorer la robustesse du système.

Figure1.8 Schéma typique d'un système de RAL [2].

La reconnaissance du locuteur se décompose en deux étapes : l'apprentissage et la reconnaissance.

Les caractéristiques d'un signal de parole de haut-parleur sont conserver comme caractéristiques de référence. Ces vecteurs caractéristiques servent à construire un modèle vocal spécifique à chaque locuteur. Le nombre de modèles requis pour assurer une reconnaissance efficace dépend des caractéristiques ou des techniques utilisées par le système. Dans la phase de reconnaissance, Dans la phase de reconnaissance, des caractéristiques similaires sont extraites d'un nouvel énoncé afin de confirmer l'identité du locuteur. Le processus de décision s'appuie sur la mesure de la distance entre le modèle de référence et celui généré à partir de l'énoncé en question.

En cas d'identification, l'énoncé d'entrée est comparé avec tous les modèles disponibles, et l'utilisateur correspondant au modèle présentant la plus petite distance est sélectionné. Pour la vérification, seule la distance entre l'énoncé et le modèle du locuteur revendiqué est évaluée. Si cette distance est inférieure à un seuil défini à l'avance, le locuteur est validé ; sinon, il est considéré comme un imposteur [2].

1.3.8 Évolution de la reconnaissance dulocuteur

La reconnaissance du locuteur a parcouru un long chemin grâce aux avancées technologiques et scientifiques. À ses débuts, elle s'appuyait principalement sur des propriétés acoustiques, mais ces systèmes montraient leurs limites face aux variations dans l'environnement ou dans la façon de parler. L'intégration des aspects linguistiques a ensuite permis de mieux différencier les locuteurs.

Avec l'arrivée de l'apprentissage automatique véritable révolution s'est opérée. Des techniques comme les modèles de mélange gaussuneien (GMM), les réseaux neuronaux et l'apprentissage profond ont permis d'améliorer significativement la précision. Les caractéristiques comportementales, comme le rythme de la parole, viennent enrichir cette analyse pour une identification encore plus fiable.

Par ailleurs, l'apprentissage par transfert a ouvert la porte à des progrès même en présence de données limitées. Plus récemment, des innovations majeures ont intégré des aspects visuels à l'analyse, en exploitant les mouvements des lèvres et les expressions faciales [15].

1.4 Conclusion

La technologie de reconnaissance automatique du locuteur (RAL) a connu des avancées notables, atteignant un degré de maturité qui lui confère le statut de solution fiable pour des applications concrètes. L'authentification vocale se généralise de plus en plus dans les systèmes de sécurité contemporains.

Dans ce chapitre, nous avons d'abord analysé la production de la parole et ses différents paramètres. Ensuite, nous avons exploré l'état actuel des systèmes de reconnaissance automatique du locuteur, en présentant leur structure fondamentale ainsi que leurs deux principales fonctions, tout en abordant leurs applications, leur fonctionnement et l'évolution de la RAL.

CHAPITRE 2

RESEAUX DE NEURONES CONVOLUTIONNELS CNN

2.1 Introduction

La croissance du Deep Learning améliore notre capacité à comprendre et à analyser des données complexes, ouvrant ainsi de nouvelles perspectives dans divers domaines tels que la médecine, la fabrication, le commerce, le marketing et d'autres domaines .Avec sa capacité à extraire des connaissances et à atteindre une grande précision dans les tâches de calcul le Deep Learning est devenu un outil essentiel pour améliorer les systèmes intelligents et stimuler le progrès technologique .Au cours de ce chapitre nous allons examiner plusieurs aspects clés liés à l'utilisation des réseaux de neurones convolutifs (CNN) dans le domaine de la reconnaissance du locuteur.

2.2 Intelligence artificielle

L'Intelligence Artificielle (IA) est une branche de la science dédiée à la conception de programmes dits intelligents. On entend généralement par programmes intelligents des logiciels capables de résoudre des problématiques qui ont habituellement été associés aux compétences humaines [16].

Il est à noter que l'intelligence artificielle comprend le champ de l'apprentissage automatique, aussi appelé « machine learning », qui lui-même englobe l'apprentissage profond, également connu sous le terme « deeplearning ». Ces trois notions sont étroitement connectées et dépendantes les unes des autres [17].

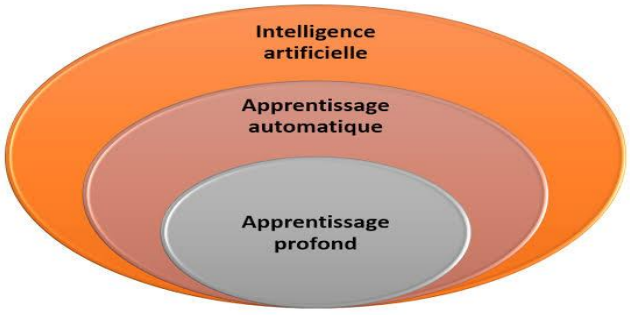

Figure2.1 Les différents domaines de l'intelligence artificielle [11]

2.3 L'Apprentissage

La caractéristique la plus fascinante des réseaux de neurones est leur capacité d'apprentissage, qui vise à extraire les informations pertinentes pour l'identification. Ainsi, c'est la phase de développement du réseau où le comportement de ce dernier est ajusté jusqu'à parvenir au comportement souhaité [22]. Au cours de l'entraînement d'un réseau, des ajustements de poids sont effectués pour améliorer la performance du réseau. Cette amélioration est considérée comme réussie lorsque le réseau parvient à se stabiliser. Il est généralement difficile de déterminer à l'avance les valeurs appropriées des poids des connexions pour une application spécifique. Une fois l'apprentissage terminé, les poids sont figés et on entre alors dans la phase de généralisation. Le réseau peut alors, dans une certaine mesure, être en mesure de généraliser. C'est-à-dire de produire des résultats corrects sur de nouveaux cas qui ne lui avaient pas été présentes au cours de l'apprentissage.

2.3.1 Apprentissage automatique

L'apprentissage automatique, également connu sous le nom machine learning, est un domaine de l'intelligence artificielle vise à donner aux machines la capacité d'apprendre à partir de données utilisant des modèles mathématiques. Essentiellement, c'est le processus Extraire des informations significatives à partir d'un ensemble de données former.

L'objectif de cette phase est d'obtenir des paramètres de modèle qui permettent d'atteindre les objectifs suivants : Meilleures performances, notamment lors de l'exécution des tâches assignées au modèle. Une fois formé, le modèle peut être déployé en production.

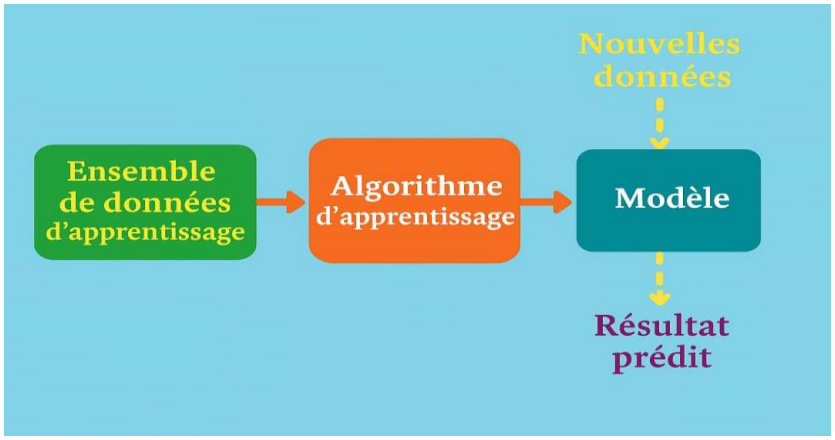

Figure2.2 Processus de l'apprentissage machine [3]

L'apprentissage automatique se subdivise en différents types, chacun étant défini par la nature des tâches à accomplir. Dans les prochaines sections, nous examinerons les principaux types d'apprentissage [16].

2.3.2 Apprentissage profond

L'apprentissage profond est une spécialisation de l'apprentissage automatique servant à former des systèmes informatiques connus sous le nom de réseaux de neurones artificiels (RNA). Ces méthodes s'appuient sur des algorithmes conçus pour imiter les actions du cerveau humain. Les réseaux neuronaux artificiels sont capables de traiter des problèmes sophistiqués comme la reconnaissance d'images, la reconnaissance de la parole et le traitement du langage naturel. Les algorithmes d'apprentissage profond utilisent plusieurs niveaux de neurones ou de nœuds synthétiques, interconnectés de diverses manières. Les différentes connexions lient les couches de nœuds entre elles. Ces réseaux ont été formés pour identifier et comprendre les caractéristiques d'un ensemble de données spécifique. Cette configuration offre aux algorithmes l'opportunité d'apprendre de leurs expériences et de perfectionner leur façon d'exécuter leurs missions.

2.3.3 Types d'apprentissage

A) Apprentissage supervisé

L'apprentissage supervisé consiste à apprendre à partir d'un ensemble d'exemples annotés fournis par un superviseur externe qualifié. Cela indique qu'une intervention humaine est requise pour l'annotation des données. Chaque instance comprend une description d'une circonstance ainsi qu'un label (une catégorie qui peut être exprimée par des valeurs numériques ou nominales) indiquant l'intervention appropriée que le système doit effectuer dans ce contexte. Le but de ce genre d'apprentissage est que le système soit capable de généraliser ses réponses et d'agir de manière appropriée dans des situations qui ne figurent pas dans l'ensemble d'apprentissage. Donc, l'utilisateur propose à l'algorithme des couples de données d'entrée et de sortie désirées (X ,Y) comme illustré dans la Figure2.3,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 rencontrépar avant [3].

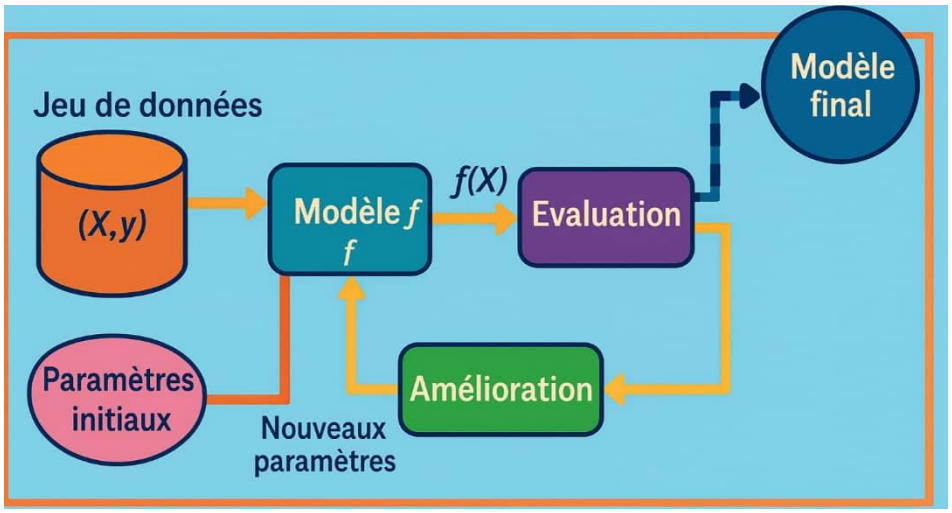

Figure 2.3 Processus de l'apprentissage supervisé [3]

Il existe deux principales catégories de problèmes d'apprentissage supervisé : la régression et La classification.

B) Apprentissage non-supervisé

Quand le système ou l'opérateur n'a que des exemples non étiquetés, et que la quantité et le type de classes ne sont pas définis d'avance, on évoque alors l'apprentissage non supervisé (ou regroupement). Il n'est pas nécessaire ni disponible d'expertise. L'algorithme est censé déterminer de manière autonome la structure, plus ou moins dissimulée, des données. Dans ce processus d'apprentissage, le système est censé orienter ses efforts vers les données dans l'espace descriptif en fonction des attributs à sa disposition, pour les classer en ensembles homogènes

d'exemples. On calcule généralement la similarité en utilisant une fonction de distance entre les couples d'exemples. Par la suite, c'est à l'opérateur qu'il incombe d'attribuer ou de tirer une signification pour chaque groupe [18].

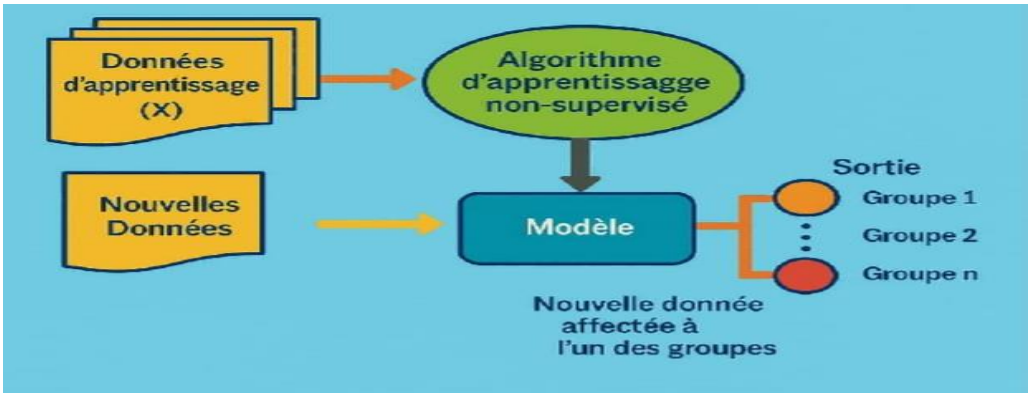

Figure2.4 Processus de l'apprentissage non-supervisé [3]

Il existe plusieurs problèmes parmi ceux-ci : le regroupement et la réduction de dimensionnalité.

Type de donnée	Apprentissage supervisé	Apprentissage non-supervisé
Discrète	Classification	Regroupement
Continue	Régression	Réduction de dimensionnalité

Tableau2.1 Type de données vs type d'apprentissage [3]

C) Apprentissage par renforcement

L'objectif de l'apprentissage par renforcement est d'apprendre à un agent à adopter un comportement adéquat dans un environnement donné, c'est-à-dire atteindre une cible définie par l'utilisateur au préalable. On décompose le problème à résoudre en une série d'étapes successives. À chaque instant, l'agent est confronté à un choix d'actions, lui permettant ainsi de communiquer avec son environnement. À l'opposé de l'apprentissage supervisé, il n'existe pas d'objectif précis

pour l'acquisition d'un comportement. Au lieu de cela, l'agent reçoit un retour (défini par l'utilisateur) lui signalant s'il a agi de manière appropriée. À chaque moment de la séquence, l'agent reçoit des données relatives à son environnement qui le guideront dans le choix de l'action adéquate. Au cours de la formation, l'agent s'efforcera d'augmenter le nombre de signaux positifs pour optimiser [16]

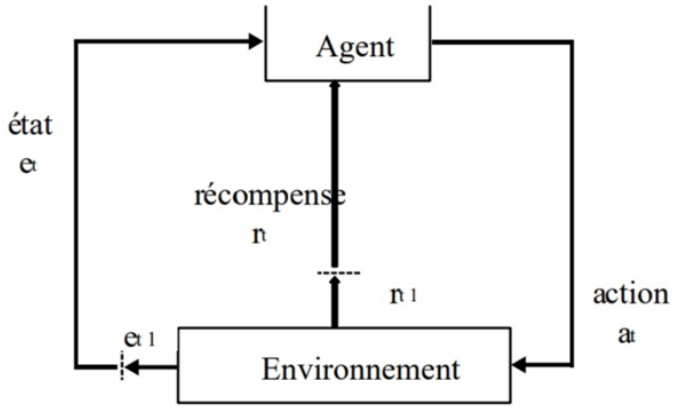

Figure 2.5 Interaction agent-environnement [3]

Ce genre d'apprentissage se prête tout particulièrement à un large éventail d'applications en robotique. Il se différencie de l'apprentissage supervisé et non-supervisé par l'emploi d'un signal de récompense qui signale simplement si l'action entreprise par l'agent est appropriée ou inappropriée, sans donner d'informations sur la meilleure option à choisir. En outre, il ne fait pas appel aux données d'entraînement ni aux étiquettes [3].

2.4 Neurone biologique

Le cerveau humain est une véritable merveille biologique, abritant environ 100 milliards de neurones. Chacun de ces neurones est connecté à 1 000 à 10 000 autres par le biais de synapses, formant un réseau incroyablement complexe.

Un neurone fonctionne comme une unité de traitement de l'information. Son corps cellulaire joue le rôle de centre de contrôle, intégrant les signaux qu'il reçoit. Les dendrites, sortes de ramifications qui s'étendent depuis le corps cellulaire, captent les informations provenant de l'extérieur et les acheminent vers le neurone. Une fois ces données traitées, elles sont transmises à d'autres neurones via l'axone, un prolongement qui agit comme un canal de communication. La

jonction entre deux neurones, où l'information est échangée, est appelée synapse [3].

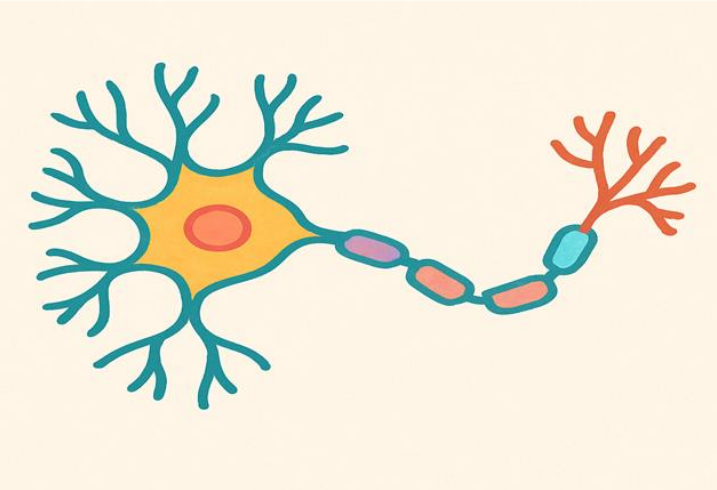

Figure2.6 : Schéma représente le neurone biologique[3]

Grâce à cette architecture sophistiquée, les réseaux de neurones biologiques excellent dans des tâches comme l'apprentissage, la mémorisation, la reconnaissance des formes et le traitement du signal. C'est en s'inspirant de leur fonctionnement que les chercheurs ont conçu les neurones artificiels, qui imitent ces mécanismes pour reproduire des formes d'intelligence [3].

2.4.1 Du neurone biologique au neurone artificiel

Le neurone biologique est constitué d'un corps cellulaire, responsable du traitement des informations, ainsi que de dendrites, qui agissent comme des capteurs de signaux. Les axones, quant à eux, assurent la transmission des signaux entre les neurones, tandis que les synapses jouent un rôle clé en régulant la force de connexion entre une dendrite et l'axone de deux neurones.

2.5 Perceptron

C'est **en 1957 que le Perceptron fut inventé par Frank Rosenblatt** au laboratoire aéronautique de Cornell. En se basant sur les premiers concepts de neurones artificiels, il proposa la « règle d'apprentissage du Perceptron ».

Un **Perceptron est un neurone artificiel**, et donc une unité de réseau de neurones. Il effectue des calculs pour détecter des caractéristiques ou des tendances dans les données d'entrée [18].

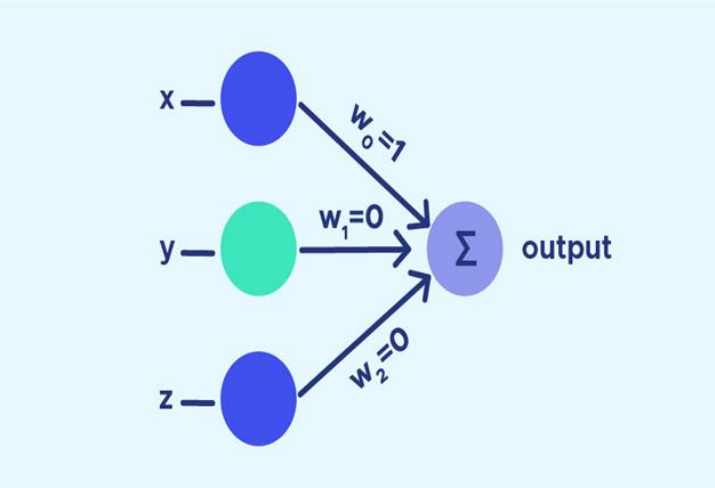

Figure2.7 Réseaumonocouche[18]

2.5.1 Perceptron multicouche

Le perceptron multicouche, ou multi layer perceptron (MLP) en anglais, est l'un des premiers réseaux de neurones à avoir été largement utilisé dans des applications concrètes, comme la reconnaissance de fleurs ou la détection de fraudes.

Ce type de réseau est structuré en plusieurs couches de neurones :

- Une couche d'entrée, qui reçoit les données initiales.

- Une ou plusieurs couches cachées, qui traitent et transforment l'information.

- Une couche de sortie, qui fournit le résultat final.

Chaque neurone d'une couche est connecté à tous les neurones de la couche précédente et de la couche suivante. Ces connexions sont associées à des poids, qui déterminent l'impact de l'activation d'un neurone sur ceux de la couche suivante. Ce mécanisme permet au réseau d'apprendre et d'affiner ses prédictions au fil du temps[18].

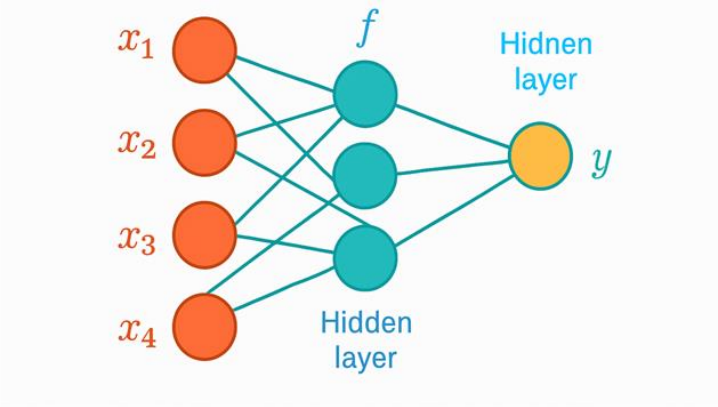

Figure2.8 Perceptron multicouche [18]

2.6 Réseaux de neuronesconvolutionnels CNN

2.6.1 Définition

Un réseau neuronal convolutif (CNN) est un type particulier de réseau neuronal,utilisé mathématique convolutionelle CNN considéréé comme le meilleur algorithme d'apprentissage pour effectuer une convolution peut extraire des fonctionnalités utiles à partir de données corrélées localement. Contrairement à la multiplication Une matrice courante utilisée dans d'autres types de réseaux neuronaux, la convolution est Utilisé dans au moins une couche d'un CNN.

Les données sont traitées via des couches convolutives, où les unités sont traitement non linéaire (fonctions d'activation) et couches de sous-échantillonnage.

Les noyaux de convolution effectuent une convolution pour extraire des caractéristiques des données L'entrée est ensuite envoyée à l'unité de traitement non linéaire. Ces unités ne sont pas fonctionnelles. Cette non-linéarité permet différentes réponses [23]

2.6.2 Architecture des CNN

Les réseaux de neurones convolutifs (CNN) sont constitués de plusieurs couches, chacune remplissant une fonction essentielle dans le traitement des données. En général, la première couche est une couche de convolution qui capte les caractéristiques élémentaires présentes dans l'entrée. Elle est suivie d'une couche de sous-échantillonnage, qui diminue la taille spatiale des cartes d'entités afin de réduire la complexité computationnelle tout en conservantl'essence de l'information.

Pour approfondir l'analyse, plusieurs couches de convolution et de sous-échantillonnage sont empilées, permettant ainsi l'extraction de caractéristiques de plus en plus abstraites. Une fois ces informations cruciales obtenues, elles sont transmises à une couche totalement connectée, qui joue un rôle fondamental dans la classification finale. Afin d'optimiser les performances du modèle et de prévenir le sur-apprentissage, les CNN intègrent également des stratégies tellesque la normalisation par lot, la régularisation et le dropout[18].

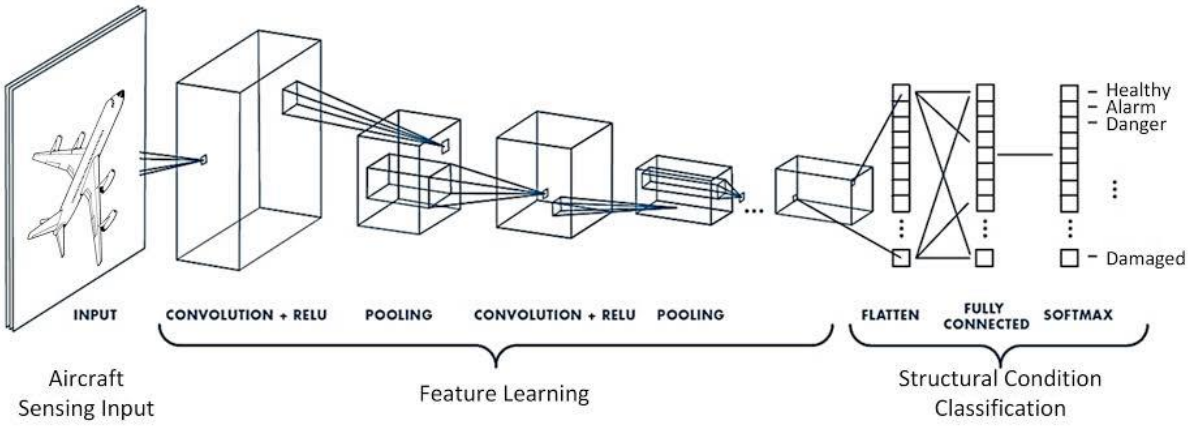

Figure 2.9 Réseau de neurones avec de nombreuses couches convolutives [18]

2.6.3 Différentes couches des CNN

Les réseaux de neurones convolutifs (ou CNN pour (Convolutional Neural Networks) sont constitués de multiples strates qui servent à dégager des attributs pertinents à partir des informations d'entrée. La configuration précise des strates peut fluctuer selon le problème et les buts particuliers.

a) Couche de convolution

La couche la plus importante dans le réseau CNN, elle joue un role essentiel dans l'extraction des caractéristiques. La figure 2.14 illustre l'opération de la convolution qui prend une image qui représente une matrice compose de pixels de 0 et 1, elle est une dimension de 7×7 pour appliquer un calcul `a l'aide d'un noyau ou d'un filtre (3x3), afin de produire une matrice Moins de dimensions (5x5) ou ce qu'on appelle une carte des caract´eristiques (FeaturesMap).

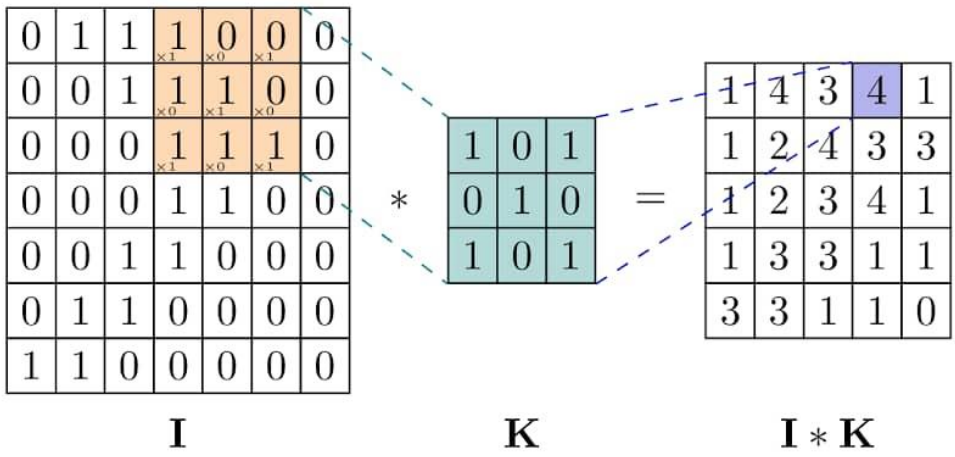

Figure2.10 Couche de convolution [22]

b) Couche d'activation

Après avoir appliqué les couches convolutives, une fonction d'activation est appliquée sur chaque carte de caractéristiques afin d'introduire des non-linéarités au sein du modèle. La fonction d'activation la plus fréquemment employée est la fonction ReLU (RectifiedLinear Unit), qui substitue les valeurs négatives par des zéros tout en préservant les valeurs positives.

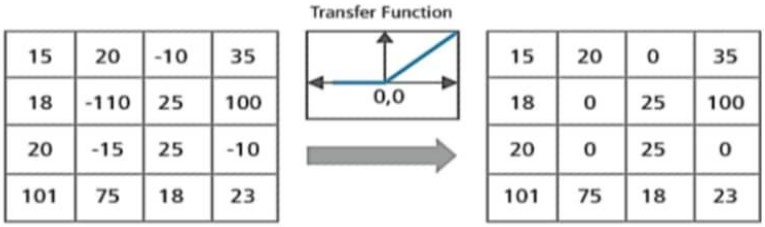

Figure2.11 Couche d'activation[22]

c) Couche de pooling

C'est une couche qui réduit les dimensions de chaque carte tout en préservant les informations importantes., il existe différents types d'opérations de pooling [18].

Maxpooling: Il choisit les aspects les plus significatifs de la fiche technique .Les aspects essentiels de la carte des caractéristiques sont conservés dans la couche maxpooling résultante. C'est la technique la plus couramment utilisée car elle donne les résultats les plus performants. [18].

Averagepooling : L'objectif est de calculer la moyenne pour chaque région spécifique de

la carte de caractéristiques. Pour ce faire, la somme des éléments dans les zones prédéfinies est d'abord obtenue à l'aide de la méthode de sum pooling [18].

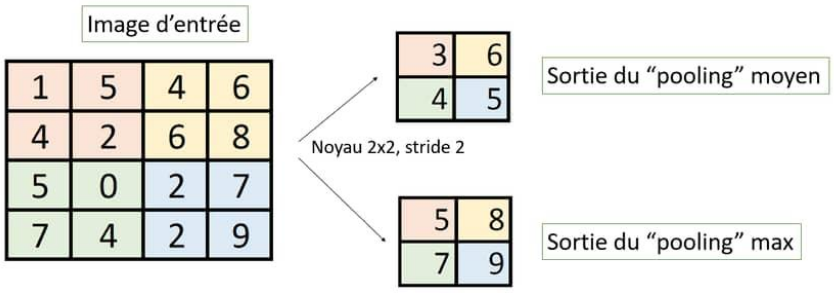

Figure2.12 Pooling moyen & poolingmaximal

d) Couche entièrement connecter (Fullyconnected layer) :

Dans une architecture de réseau de neurones, la couche entièrement connectée joue un rôle comparable à celui des réseaux classiques. Elle reçoit les informations issues de la première phase de traitement, qui combine des opérations de convolution et de mise en commun successives. Ces données sont ensuite transformées via un calcul de produit scalaire entre le vecteur de poids et le vecteur d'entrée, conduisant ainsi à la génération de la sortie finale [29].

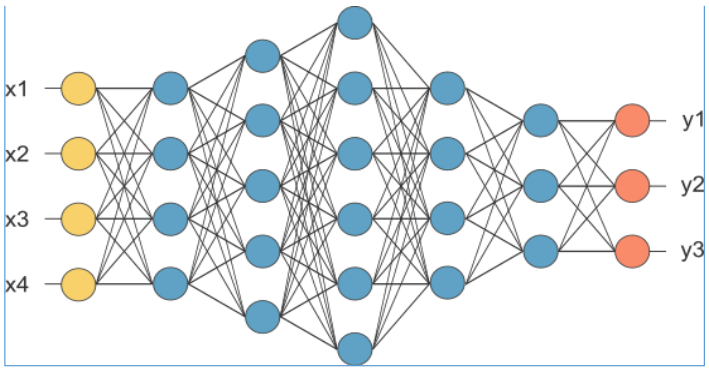

Figure2.13 Couchefully-connected [22].

e) Couche de sortie :

La dernière couche d'un CNN est la couche de sortie, qui génère des prédictions ou des probabilités associées à différentes classes ou catégories. Selon le problème, cette couche peut

utiliser une fonction d'activation appropriée telle qu'une fonction softmax pour la classification multi-classes [18].

2.6.4 Fonctions d'activation :

Les fonctions d'activation d'un réseau neuronal convolutif (CNN) jouent un rôle important dans capacité du réseau à introduire la non-linéarité et à capturer des relations complexes entre les données Le choix de la fonction de déclenchement dépend du problème spécifique et peut être testé pour de meilleurs résultats ..Les fonctions d'activation couramment utilisées dans les CNN sont [25] :

a) Fonction d'activation ReLU (RectifiedLinearUnit)

Cette fonction est définie comme f(x) = max (0, x), où x est l'entrée du neurone. Il conserve les valeurs positives et rejette les valeurs négatives, introduisant ainsi une non-linéarité dans le réseau.

b) Fonction d'activation Sigmoïde

La fonction sigmoïde est une fonction non linéaire qui transforme les valeurs d'entrée dans une plage de 0 à 1. Elle est définie comme f(x) = 1 / (1 + exp(-x)). Il est couramment utilisé dans les classes de sortie CNN pour la classification binaire ou probabiliste.

c) Fonction d'activation Softmax

La fonction softmax est utilisée pour classer plusieurs classes. Il convertit les valeurs d'entrée en une distribution de probabilité où la somme de toutes les sorties est égale à 1. Il est souvent utilisé dans la couche finale des CNN pour générer les probabilités de classe.

2.6.5 Paramètres des CNN

Pour construire un modèle efficace, plusieurs choix sont nécessaires à chaque couche :

- **Couchesprincipales** : Un réseau CNN se compose généralement de couches de convolution, de couches de correction ReLU, de couches de pooling et de couches entièrement connectées.

- Paramètres des couches de convolution :

- Le nombre de noyaux de convolution, qui définit le nombre de filtres appliqués aux données.

- Le pas de chevauchement, qui détermine l'écart entre chaque application de filtre.

- La marge à zéro (zéro padding), qui ajoute des zéros autour des bords des données pour maintenir la taille des cartes de caractéristiques. Dans certains cas, il peut être préférable de ne pas ajouter de zéros, ce qui réduit la taille de la carte des caractéristiques.

- Paramètres des couches de pooling :

- La taille de la fenêtre de traitement, qui définit la zone d'agrégation des données.

- Le pas de chevauchement, qui détermine le décalage de la fenêtre à chaque étape. Une fenêtre de **2x2 avec un pas de 1** est souvent utilisée.

2.6.6 Optimisation et évaluation du CNN :

L'entraînement d'un réseau de neurones convolutifs (CNN) vise à ajuster ses poids afin d'améliorer la précision de ses prédictions. Au départ, ces poids sont attribués aléatoirement. Le CNN est ensuite entraîné à l'aide d'un ensemble de données étiquetées, où chaque entrée correspond à une classe spécifique. Lors du traitement des données, le réseau compare ses prédictions aux étiquettes réelles et ajuste ses poids en conséquence grâce à la rétropropagation .

Une fois l'entraînement terminé, le modèle est évalué sur un ensemble de test composé de données inédites. Cette phase permet de mesurer la précision du CNN et sa capacité à généraliser ses prédictions. Un écart significatif entre la performance sur les données d'entraînement et celles de test peut révéler un sur-apprentissage, indiquant que le réseau s'est trop adapté aux données d'entraînement et peine à traiter de nouvelles informations. Ce phénomène est souvent lié à une taille insuffisante de l'ensemble de données.

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 [3].

2.6.7 Pourquoi les CNN sont-ils adaptés à la reconnaissance du locuteur ?

Les CNN offrent plusieurs avantages majeurs pour cette tâche

- **Extractionintelligentedescaractéristiques** : Grâce à leurs couches de convolution, ils identifient efficacement des motifs spécifiques dans les signaux vocaux.

- **Robustessefaceauxvariations** : Les CNN restent performants même lorsque l'enregistrement du locuteur varie (intonation, vitesse de parole, stress).

- **Réductiondeladimensionnalité** : Les couches de pooling compressent les données tout en conservant les informations essentielles, facilitant le traitement et améliorant la classification.

- **Apprentissageautomatisé** : Contrairement aux méthodes classiques nécessitant une extraction manuelle, les CNN apprennent directement à partir des données, ce qui optimise l'identification du locuteur sans dépendre d'expertise humaine.

En résumé, les CNN sont de précieux alliés en reconnaissance du locuteur, combinant flexibilité, performance et capacité d'apprentissage automatisé pour capturer l'essence d'une voix.

2.7 GammatoneCepstral Coefficient (GTCC)

2.7.1 Définition

Le GammatoneCepstral Coefficient (GTCC) est une méthode d'extraction de caractéristiques audio inspirée du fonctionnement de l'oreille humaine, et plus précisément de la cochléa[27].
Contrairement au MFCC (Mel-FrequencyCepstral Coefficient) qui repose sur l'échelle de Mel,le GTCC utilise une banque de filtres Gammatone, qui modélise plus fidèlement lecomportement

des cellules ciliées dans la cochlée.

Les filtres Gammatone simulent la réponse fréquentielle des fibres nerveuses auditives, et permettent de mieux capturer les informations perceptuelles, en particulier dans lesenvironnements bruités. Le processus d'extraction GTCC comprend généralement :

1. Le passage du signal audio dans une banque de filtres Gammatone.
2. Le calcul de l'énergie dans chaque bande fréquentielle.
3. L'application d'une transformation cepstrale (typiquement la DCT) pour obtenir un vecteur de coefficients représentant les caractéristiques du signal.

Les GTCC sont de plus en plus utilisés dans des tâches telles que la reconnaissance vocale, l'identification de locuteur, ou encore la classification sonore, où une robustesse perceptuelle est essentielle[27]

2.7.2-Coefficients CepstrauxGammatone (GTCC)

Avec la banque de filtres gammatone, il est possible de calculer les GTCC de manière analogue au schéma d'extraction des MFCC. Tout d'abord, les signaux audio d'origine sont découpés en courtes trames. La durée de chaque trame est fixée par défaut .Ensuite, pour chaque trame, on applique la transformée de Fourier rapide (FFT) afin d'analyser le spectre de la trame. Par la suite, la banque de filtres gammatone est appliquée à la FFT du signal, ce qui permet dobtenir un spectre en sous-bandes. L'énergie de chaque sous-bande est alors calculée et notée (Xn).

Dans la dernière étape, la fonction logarithmique et la transformée en cosinus discrète sont appliquées pour modéliser la perception humaine de l'intensité sonore et décorréler les sorties logarithmiquement compressées des filtres. Les GTCC peuvent être calculés comme suit :

$$GTCC_m = \sqrt{\frac{2}{N}} \sum_{n=1}^{N} log_{10}(X_n) \cos\left[\frac{\pi n}{N}(m - \frac{1}{2})\right], 1 \leq m \leq M \quad \text{.....2.1}$$

Où (Xn) représente l'énergie de la nth sous-bande, (N) le nombre de filtres gammatone et (M) le nombre de coefficients GTCC.

Ce processus permet une meilleure concentration de l'énergie. En parallèle, un groupe de caractéristiques plus représentatif est extrait avec une complexité de calcul similaire à celle des MFCC. Les résultats expérimentaux montrent la supériorité des caractéristiques GTCC dans la

classification du bruit ambiant.[28]

2.7.3 Comparaisson entre MFCC et GTCC

Le tableau 2.2 présente une comparaison entre les 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 cochléé)
Type de filtre utilisé	Filtres triangulaires	Filtres Gammatone
Robustesse au bruit	Moyenne à faible dans les environnements bruyants	Meilleure robustessse surtout en environnements bruité
Utilisation classique	Trés répandu dans la reconnaissance vocale automatique	De plus en plus utilisé pour la reconnaissance du locuteur
Complexité computationelle	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	Bonne performances en environnement controlé	Souvent meilleures performances en conditions réelles et bruitées

Tableau2.2 : Comparaison entre MFCC Et GTCC[28]

2.8 Conclusion

Dans ce chapitre Nous avons présenté le domaine de l'apprentissage profond, commencerons par une introduction générale à l'intelligence artificielle et à l'apprentissage automatique, soulignant l'importance de l'apprentissage profond et des CNN. Nous aborderons couches qui composent les CNN et leur fonction respective dans le traitement des données d'entrée. Ensuite, nous mettrons en évidence les avantages distincts des CNN dans le domaine de la reconnaissance du locuteur et leur capacité à extraire automatiquement des caractéristique pertinentes à partir de donnes audios.Ainsi que les architectureset les méthodes d'entraînement spécifiques qui contribuent à l'efficacité et à la précision de ces modèles dans le domaine de la télécommunication moderne de plus on a parler de GTCC et son fonctionnement et ses avantages par rapport au MFCC.

Chapitre 3

TRAVAUX D'ÉXPERIMENTATION ET RÉSULTATS

3.1 Introduction

Dans ce chapitre, nous présentons et analysons nos travaux expérimentaux visant à évaluer l'efficacité de l'apprentissage automatique pour l'identification du locuteur en multi langage multifréquences.

Notre travail consiste à élaborer un modèle d'identification capable de reconnaître un individu dans plusieurs scénarios : enregistrements de très haute qualité (44 KHz), de bonne qualité (16 KHz), ainsi que des enregistrements de qualité téléphonique fixe (8Khz). La variation linguistique est également prise en considération.

La première étape consiste en la collecte d'une base de données vocale, suivie du développement et de l'évaluation d'un modèle d'identification CNN combiné au GTCC testant diverses paramètres.la data augmentation et validation sont également exploité afin d'optimiser notre modèle, ensuite, nous améliorons les performances de notre système par l'utilisation de l'augmentation de données. L'ensemble de ces expérimentations seront implémenté via une bibliothèque Python.

3.2 Environnements expérimentaux utilisés

3.2.1 Langage Python

Python est un langage de programmation largement adopté par les professionnels de la donnée. Développé par Guido van Rossum, sa première version est sortie en 1991. Ses usages dépassent le cadre de la Data Science et incluent le développement logiciel, la conception d'algorithmes, ainsi que la gestion des infrastructures web des réseaux sociaux [16].

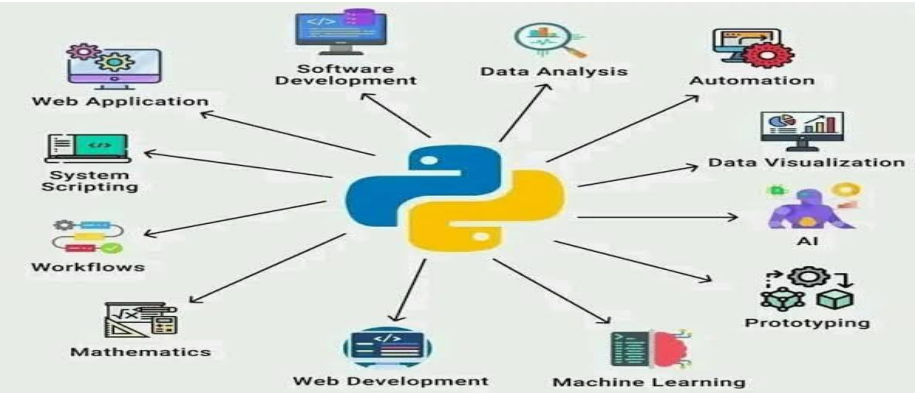

Figure 3.1 Les domaines d'applications de python

Python est à la fois simple et puissant, Il nous permet de créer des scripts très faciles à écrire et possédant de nombreuses bibliothèques, nous pouvons nous attaquer à des projets plus ambitieux. C'est un code de programmation qui se déroule en ligne, c'est-à-dire qu'il n'est pas nécessaire de le compiler avant de l'exécuter. Il est polyvalent, ça veut dire qu'il opérate sur différent systèmes d'exploitation : RaspberryPi,MacOSX,Linux, Android,iOS et meme sur les mini-ordinateurs [24].

Python est un langage à la fois intuitif et puissant, permettant la création de scripts simples à rédiger tout en offrant un large éventail de bibliothèques facilitant la réalisation de projets ambitieux. Son interprétation en ligne élimine la nécessité de compilation préalable, ce qui le rend particulièrement flexible. De plus, sa compatibilité avec divers systèmes d'exploitation, tels que Raspberry Pi, Mac OS X, Linux, Android, iOS et même les mini-ordinateurs, en fait un outil polyvalent.

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

3.3 Travaux d'expérimentation

3.3.1 Présentation

Dans le cadre de nos recherches expérimentales, nous avons mené une série d'études approfondies visant à explorer différents aspects de la reconnaissance automatique du locuteur. La figure associée illustre de manière détaillée la méthodologie adoptée pour cette étude.

Le schéma présenté dans la figure 3.2 décrit précisément l'approche méthodologique suivie pour notre travail.

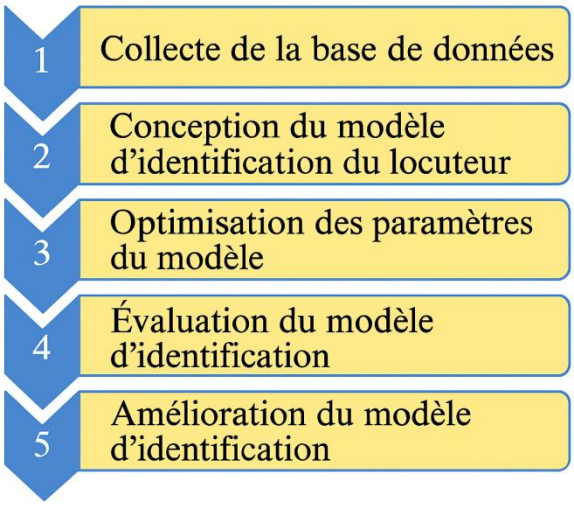

Figure3.2 Démarche méthodologique de notre travail

3.3.2 Collecte de la base de données

Pour constituer notre base de données en dialecte algérien, nous avons suivi une approche inspirée de la base TIMIT du dialecte américain, tout en apportant les ajustements nécessaires. Toutefois, certaines contraintes n'ont pas pu être rigoureusement respectées pour certains enregistrements.

La collecte des données a été guidée par les critères suivants :

- Sélection de phrases équilibrées sur le plan phonétique
- Enregistrements réalisés dans un environnement acoustiquement contrôlé, au format WAV
- Utilisation d'un microphone haute performance positionné à 20 cm du locuteur
- Normalisation de la durée des échantillons d'apprentissage
- Protocole d'enregistrement structuré comprenant :
- 23 phrases par locuteur (14 en dialecte algérien, 5 en français et 4 en anglais)
- Trois fréquences d'échantillonnage : 44 kHz, 16 kHz et 8 kHz

3.3.3 Conception du modèle d'identification du locuteur

Cette section détaille l'approche méthodologique adoptée pour le développement de notre modèle d'identification du locuteur, illustrée par l'organigramme présenté dans la figure 3.3.

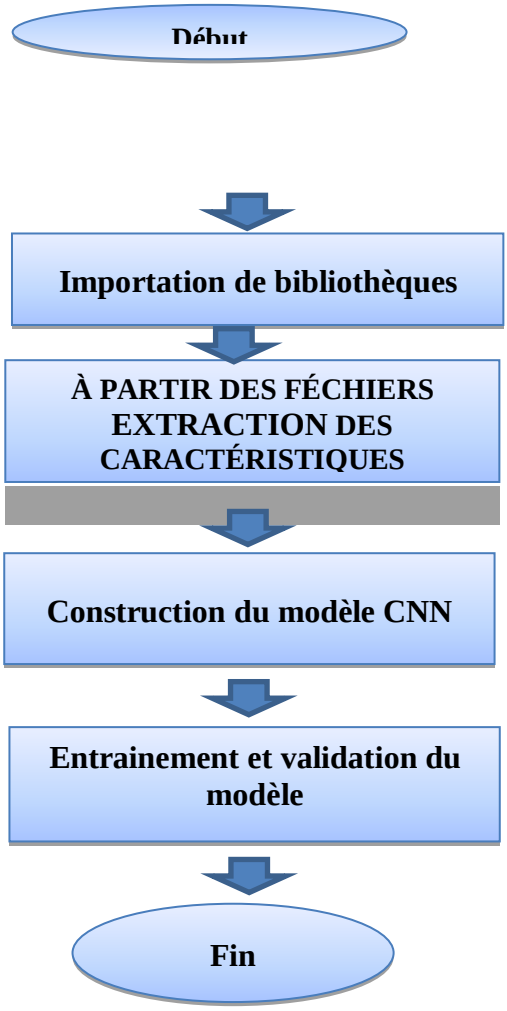

Figure 3.3 Modèle d'apprentissage

L'approche méthodologique suivie dans cette étude s'articule autour de quatre étapes majeures, illustrées dans l'organigramme. Chacune de ces étapes sera examinée en détail dans les sections

suivantes.

3.3.3.1.Importation de la bibliothèque

Les différentes bibliothèques utilisées et leurs fonctions sont présentées dans le tableau 3.1

Bibliothèque	**Fonction**
Numpy	Opérations numériques
Tensorflow	Construction et entrainement du modèle de Deep Learning
Librosa	Traitement audio (lecture,analyse)
Os	Gestion des fichiers et des répertoirs
Sklearn	Fournit des outils pour apprentissage automatique
Matplotlib	Visualisation des données sous forme graphique et de diagrammes
Gammatone	Manipuler les filtres et représentation pour l'audio

Tableau 3.1 : Lesdifférentesbibliothèquesetleursfonctions

3.3.3.2 Extraction des caractéristiques audio avec GTCC

L'extraction des GTCC repose sur une modélisation inspirée du fonctionnement du système auditif humain. Les GTCC utilisent des filtres Gammatone qui reflètent mieux la réponse fréquentielle de l'oreille humaine.

A) Préparation des données

Avant de procéder à l'extraction des GTCC, nous effectuons plusieurs opérations :

- Chargement et normalisation des fichiers audio pour assurer une cohérence des données.

-Application des filtres Gammatone afin de segmenter le signal sonore en bandes fréquentielles adaptées à la perception humaine.

B) Calcul des GTCC

- Transformation du signal audio en une représentation cepstrale basée sur la réponse des filtresGammatone.

- Stockage des coefficients dans une matrice `NumPy`.

C) Encodage des étiquettes

- Conversion des labels textuels en valeurs numériques (`LabelEncoder`).

- Création d'une correspondance entre les labels et leurs indices.

D) Contrôle et vérification

- Suppression des fichiers corrompus ou non exploitables.

- Vérification de la cohérence entre les échantillons et leurs étiquettes.

- Vérification des dimensions des GTCC .

3.3.3.3. Construction du modèle CNN avec GTCC

Le modèle basé sur ConvNet (CNN) utilise directement les GTCC comme entrée :

- Modèle séquentiel structuré avec des couches adaptées à l'analyse de séquences temporelles.

- Utilisation de Conv1D, permettant relations temporelles des coefficients GTCC.

3.3.3.4. 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%.

- ✓ **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 sur-apprentissage. L'optimiseur déployé dans notre modèle : ADAM.

1. 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.

3.3.4. Optimisation des paramètres du modèle

Pour optimiser les performances de notre modèle, nous avons mené plusieurs expérimentations étudiant les phénomènes de sous-apprentissage et de sur-apprentissage sur l'ensemble des enregistrements en dialecte algérien à 16 kHz. Les résultats présentés dans la figure3de notre script montrent notamment un sur-apprentissage important avec des taux de :

- ✓ **Précision de validation variable, atteignant une valeur maximale de 73.91%**
- ✓ **Perte qui varie insuffisant considérablement, avec une valeur maximale 1,40**
- ✓ **Test égal à : 60,87% qui est pour un système de classification.**

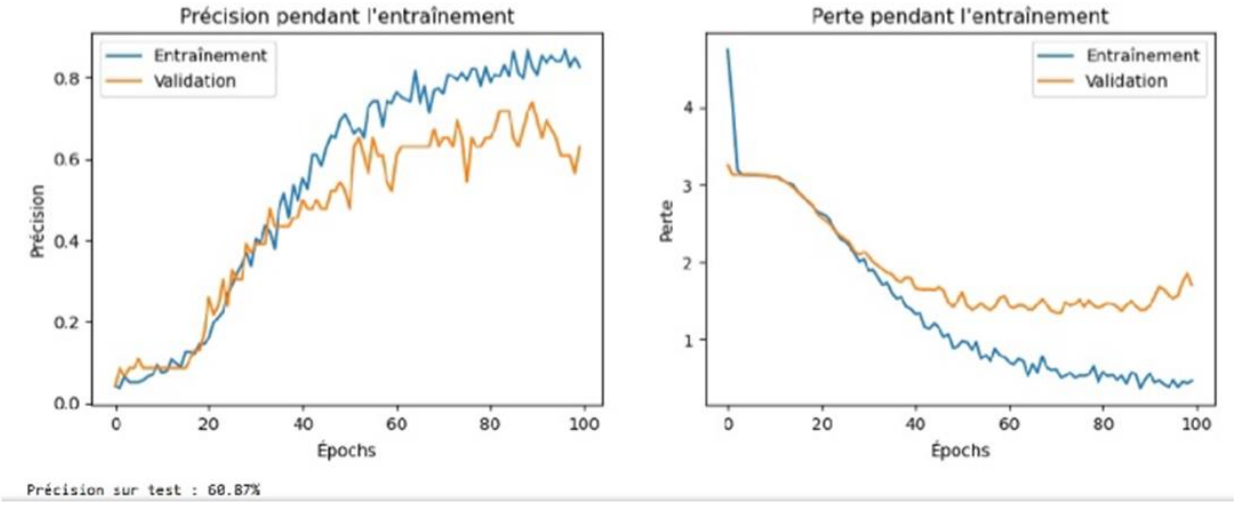

Figure 3.4 : Précision et la perte pendant l'entrainement

Pour optimiser les paramètres, nous avons utilisé un algorithme appelé le **RANDOM SEARCH**. Il permet de déterminer les paramètres adéquats afin d'améliorer les taux de la précision et la perte de la validation, y compris celui de la classification.

Nous avons également ajouté d'autres paramètres qui n'existait pas dans le programme initial il s'agit de :

- ✓ Application de la **régularisation L1 et L2** pour éviter le sur-apprentissage.
- ✓ Application de la fonction **Shuffle** pour mélanger les données aléatoirement, ce qui permet d'éviter que le modèle n'apprenne un ordre artificiel ou des motifs non représentatifs dans les données d'entraînement.
- ✓ Utilisation du taux d'apprentissage adaptatif. Ce paramètre permet d'ajuster le taux d'apprentissage afin d'éviter le sur-apprentissage et le sous-apprentissage.
- ✓ **Early_stopping :** cette fonction permet d'arrêter l'apprentissage au moment opportun afin d'éviter le phénomène de sur-apprentissage.

3.3.4.1 Évaluation du modèle d'identification

Dans cette expérimentation, nous avons appliqué notre modèle à six ensembles d'enregistrements issus de notre base de données, comprenant des données multilingues et monolingues en dialecte algérien, chacune avec trois fréquences différentes : 8 kHz, 16 kHz et 44 kHz. L'évaluation des

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.

a) Analyse des taux de perte et précision

Les figures **3.4,3.5,3.6** ,**3.7,3.8** et **3.9** illustrent respectivement les courbes de perte et de précision en fonction de nombre d'époques, des ensembles d'enregistrement monolingue 8Khz, monolingue 16Khz et monolingue 44Khz et multilingue 8Khz, multilingue 16Khz et multilingue 44 Khz.

Globalement, la précision augmente avec le nombre d'époques, cela se traduit à un meilleur ajustement du modèle pendant l'apprentissage.

Les résultats observés montrent que l'ensemble des enregistrements multilingues présente les meilleurs résultats, ce qui démontre sa robustesse en termes de caractéristiques vocales.

Plus précisément celle dont la fréquence d'échantillonnage 44KHz .Néanmoins le taux de précision de validation qui vaut et taux de classification restent insuffisant pour un modèle d'identification qui doit atteindre au moins un taux de 90 %.

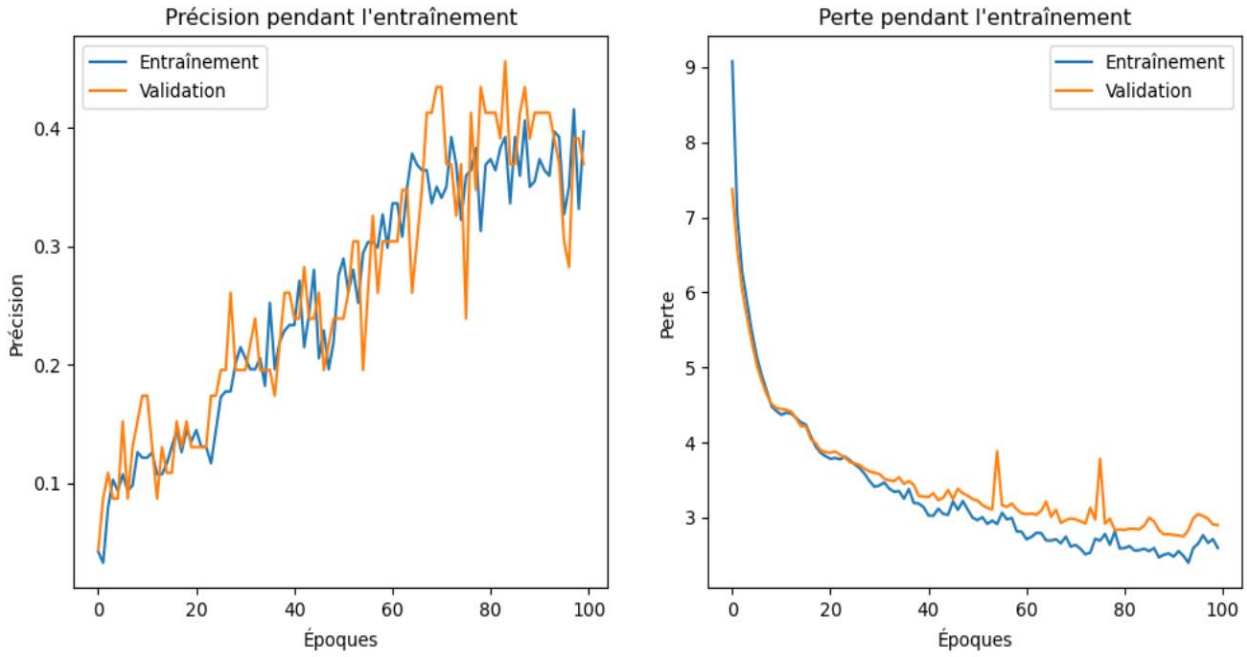

Figure3.5 : Taux de précision et de perte de l'ensemble monolingue à 8Khz

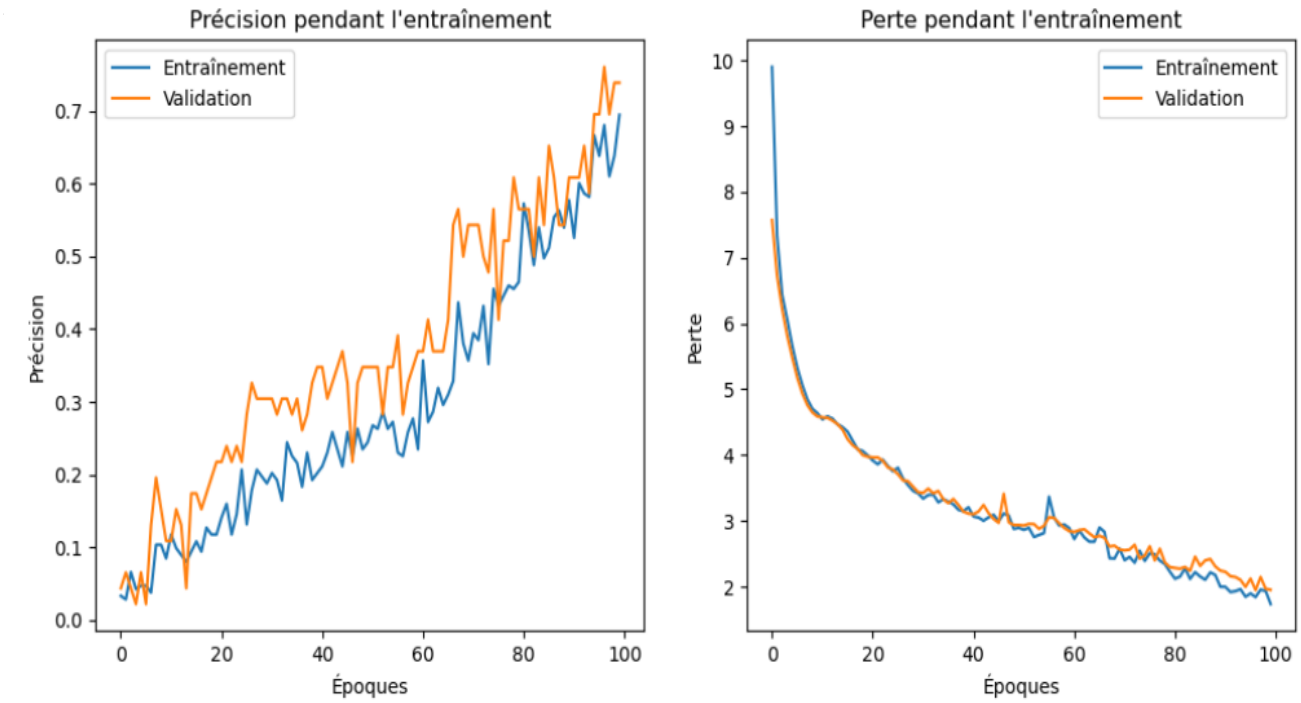

Figure3.6 : Taux de précision et de perte de l'ensemble monolingue à 16Khz

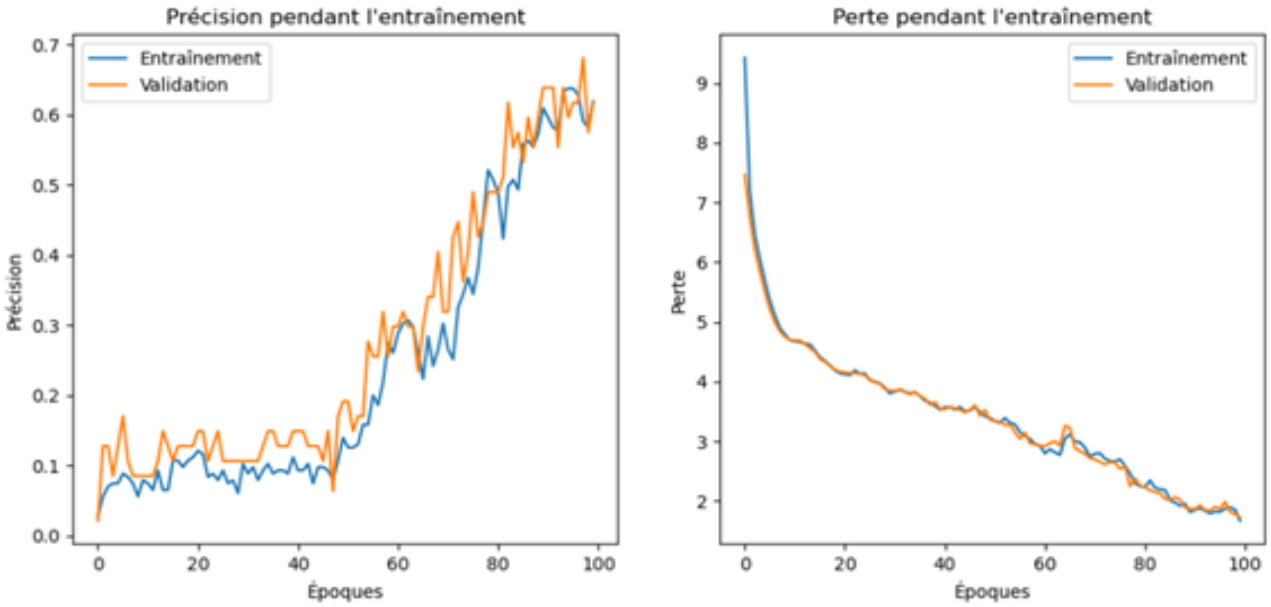

Figure3.7 : Taux de précision et de perte de l'ensemble monolingue à 44Khz

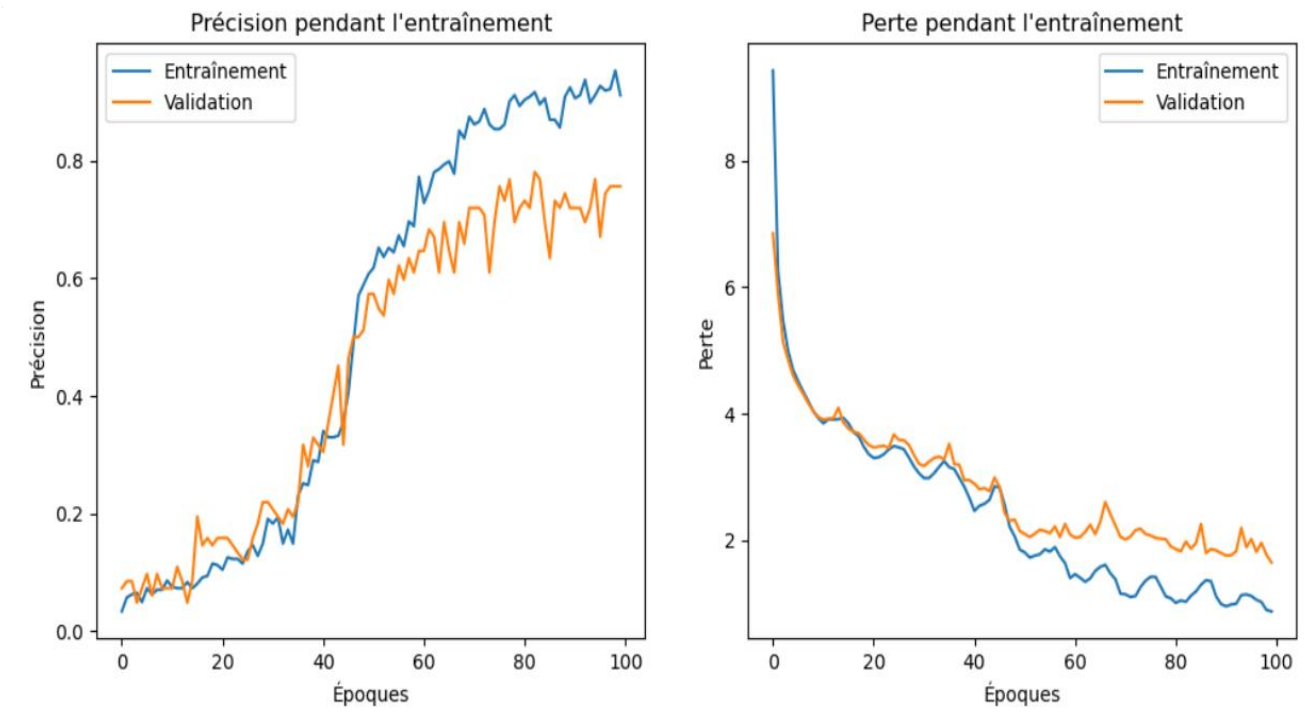

Figure3.8 : Taux de précision et de perte de l'ensemble multilingue à 8Khz

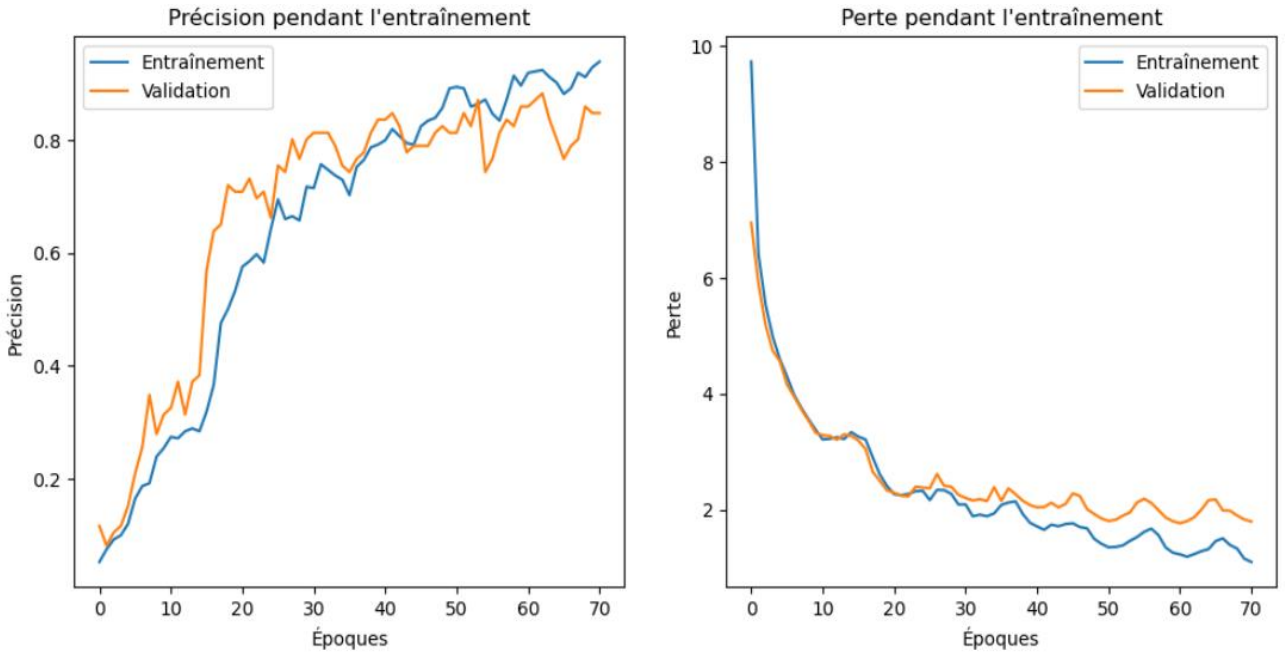

Figure3.9: Taux de précision et de perte de l'ensemble multilingue à 16Khz

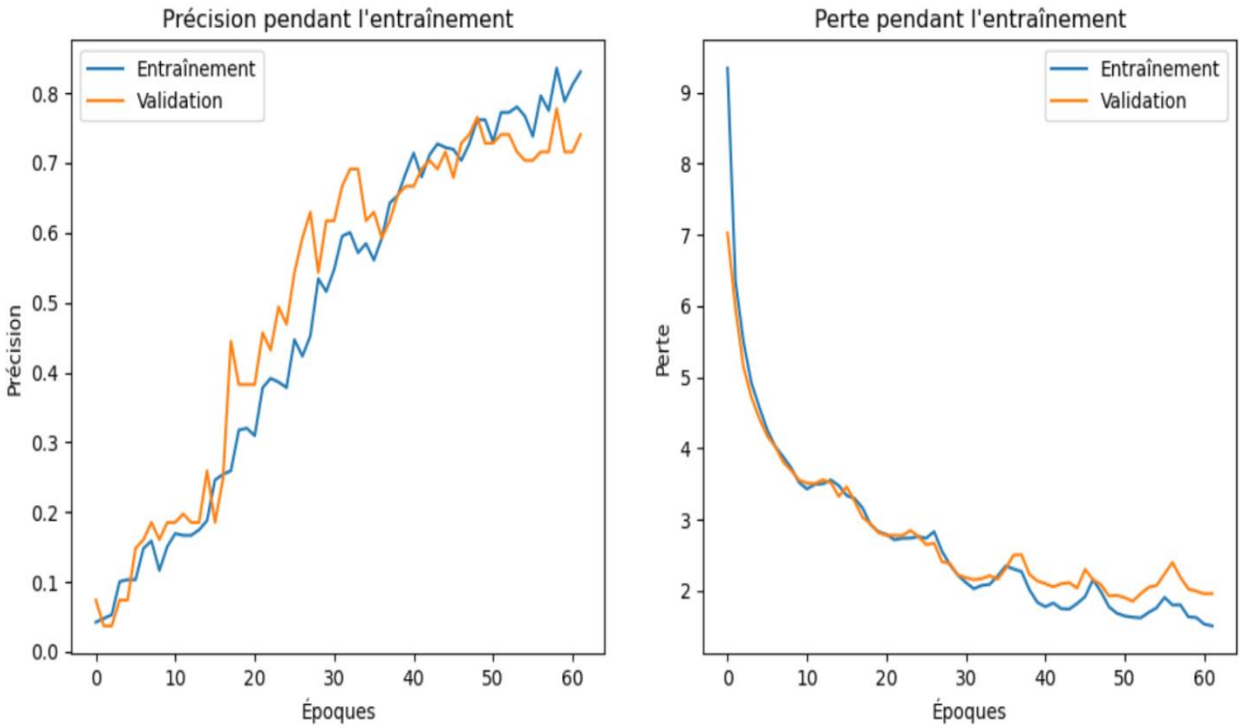

Figure3.10 : Taux de précision et de perte de l'ensemble multilingue à 44Khz

b) MATRICES DE CONFUSION

Analyse de la matrice de confusion et taux de test

La matrice de confusion est un outil d'évaluation utilisé pour analyser les performances d'un modèle de classification. Elle permet de comparer les prédictions du modèle avec les valeurs réelles, en affichant le nombre de vrais positifs, de faux positifs, de vrais négatifs et de faux négatifs.

La matrice de confusion se présente généralement sous la forme d'un tableau à double entrée, où les lignes représentent les classes réelles et les colonnes représentent les classes prédites ainsi :

- ✓ Chaque ligne correspond à l'identité réelle (le "vrai locuteur" ou "vraie personne").
- ✓ Chaque colonne correspond à l'identité prédite par ton système.
- ✓ Les valeurs sur la diagonale (du haut gauche au bas droite) sont les bonnes prédictions.
- ✓ Les valeurs hors diagonale sont les erreurs (confusions entre identités).

Les figures 3.10, 3.11, 3.12 ,3.13, 3.14 et 3.15 illustrent respectivement les matrices de confusion des ensembles d'enregistrement monolingue 8Khz, monolingue 16Khz , et monolingue 44Khz et multilingue 8Khz, multilingue 16Khz , et multilingue 44Khz :

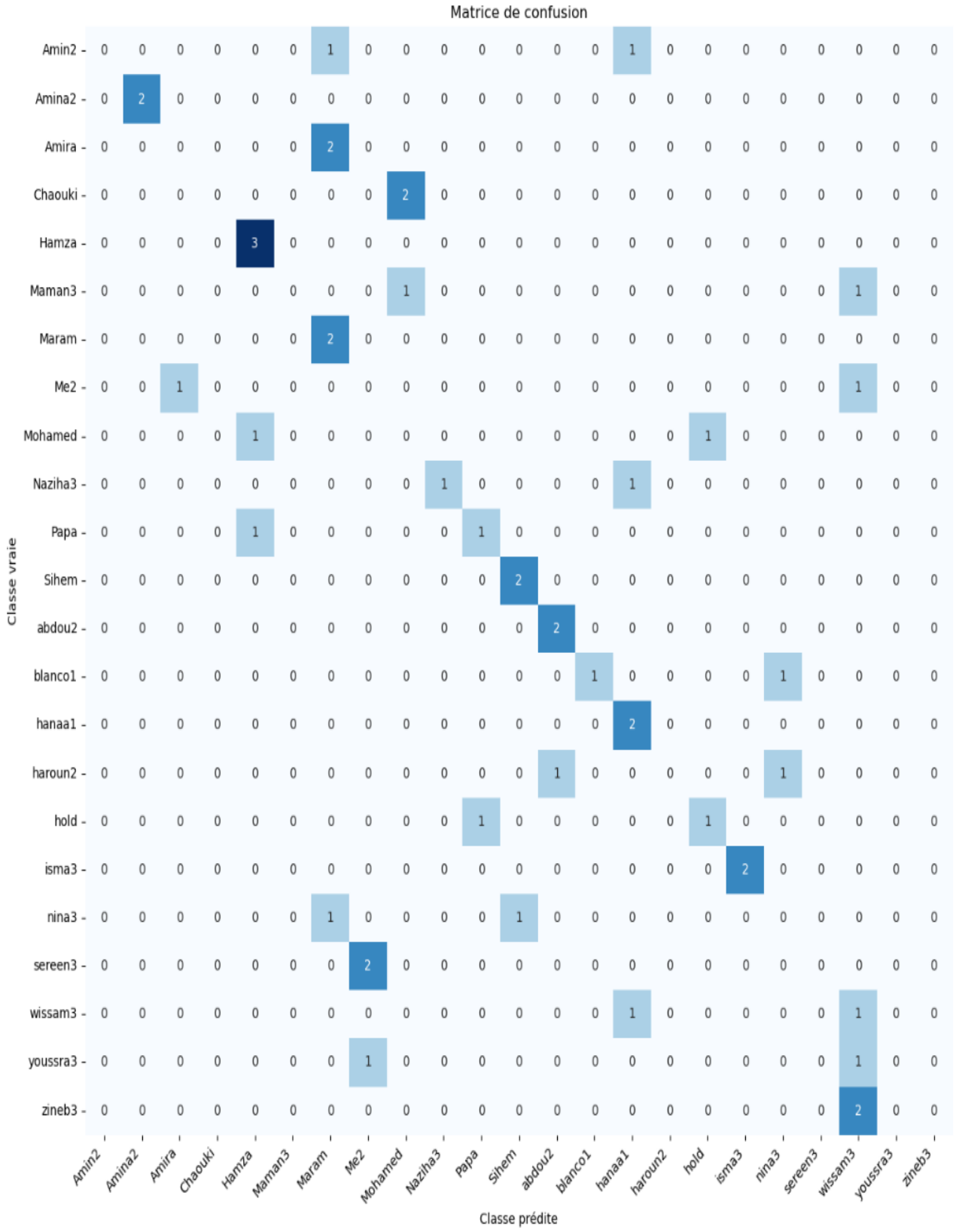

Figure 3.11 : Matricedeconfusionmonolingue8KHz

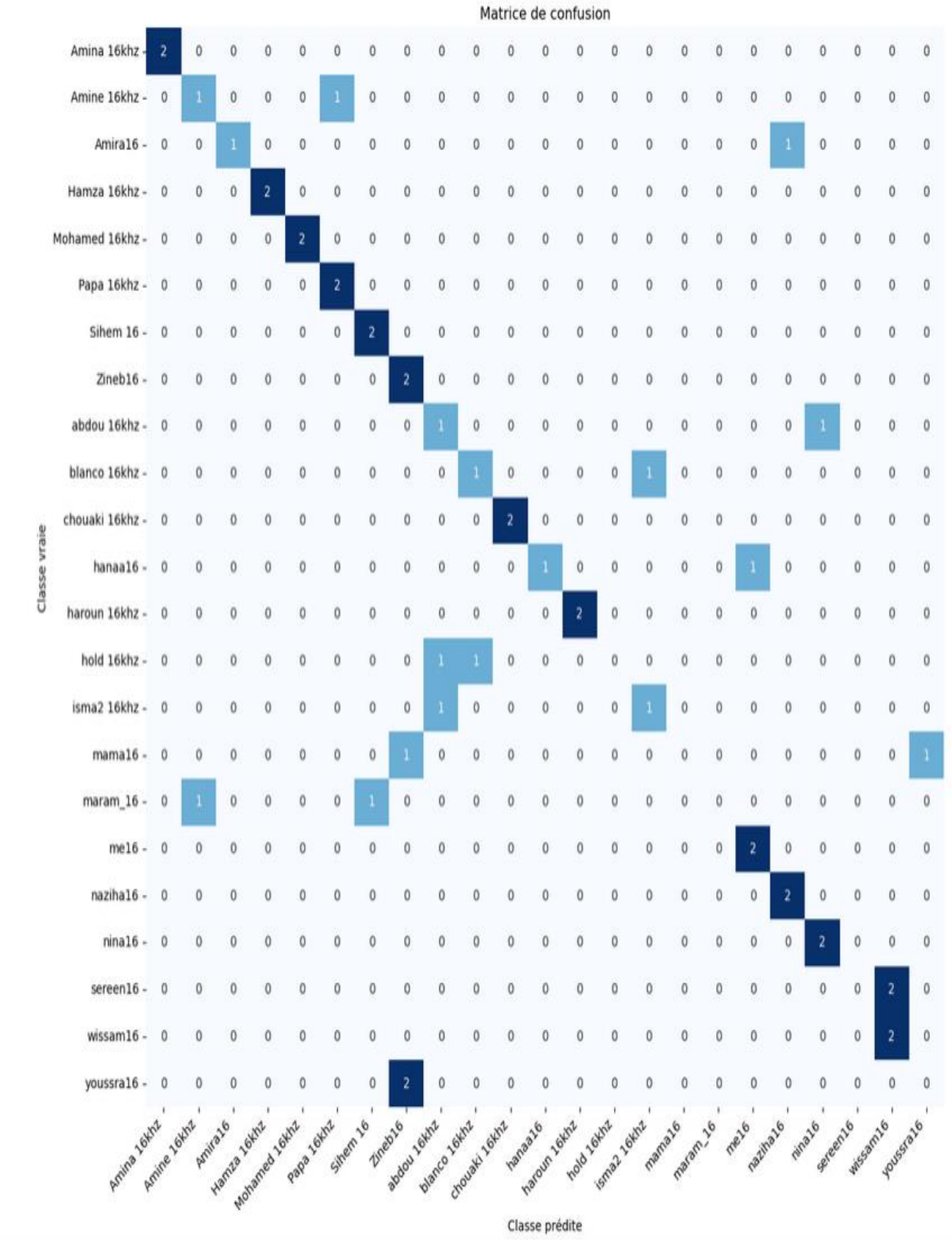

Figure 3.12 : Matricedeconfusionmonolingue16KHz

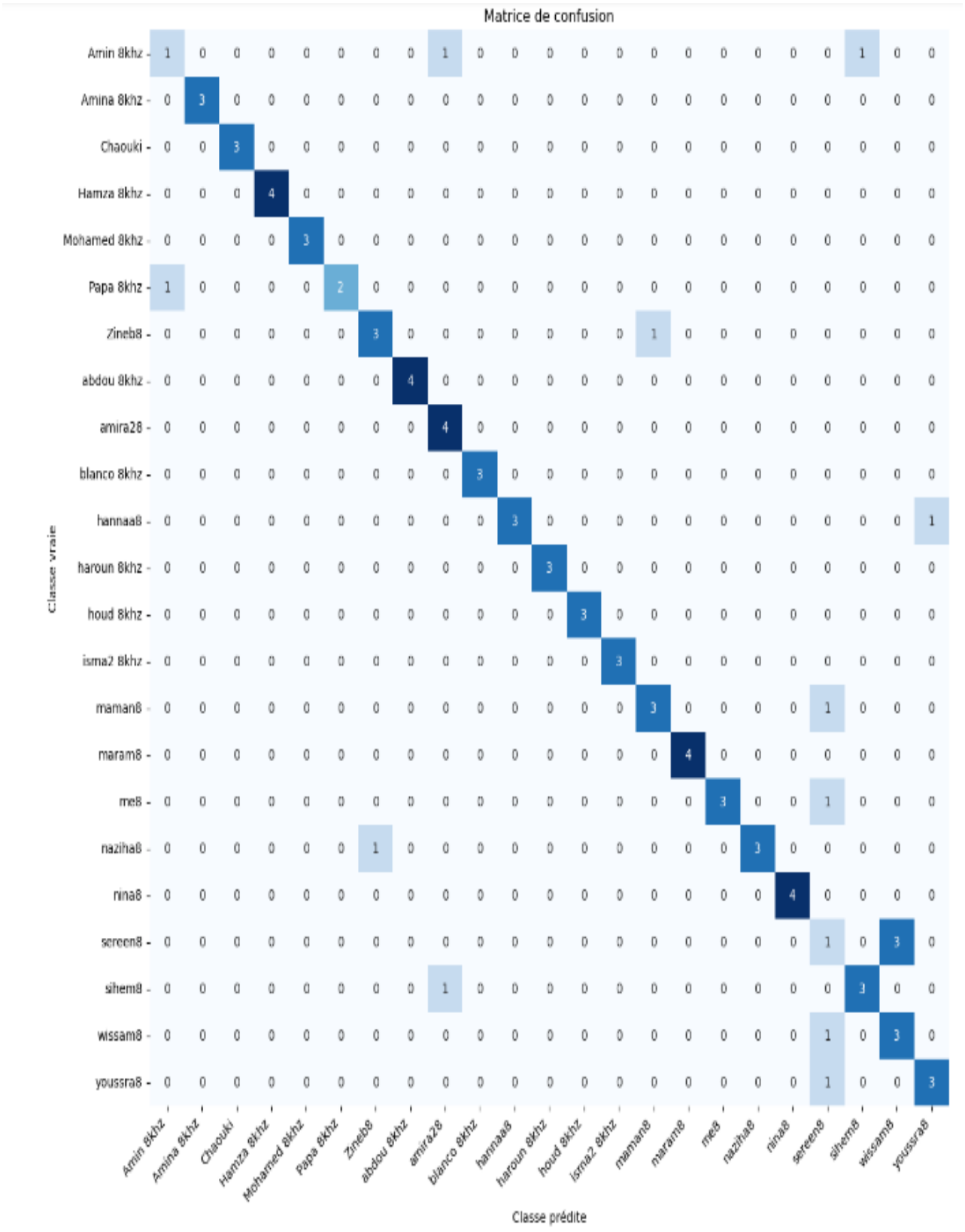

Figure 3.13 : Matricedeconfusionmonolingue44KHz

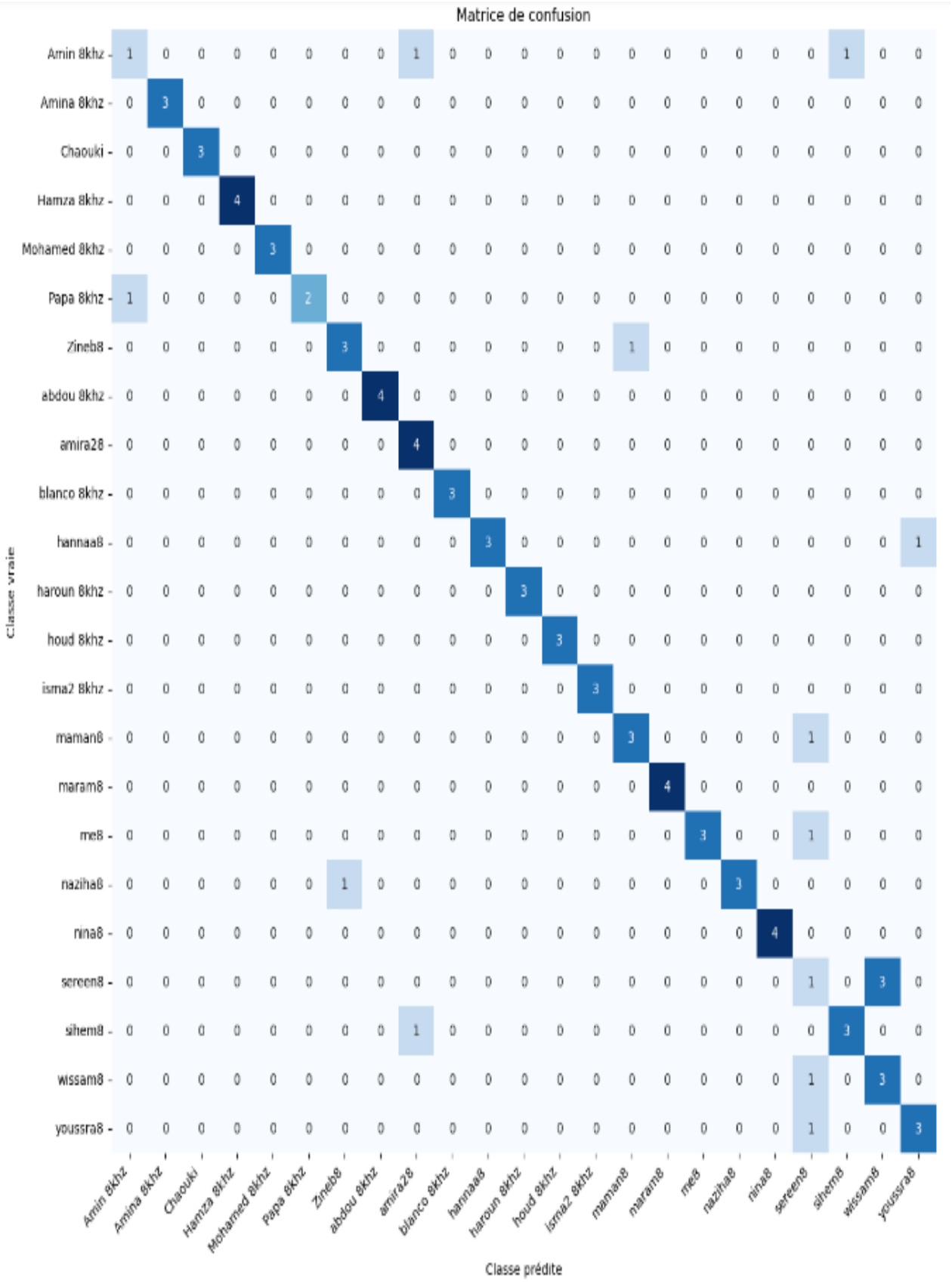

Figure 3.14 : Matricedeconfusionmultilingue8KHz

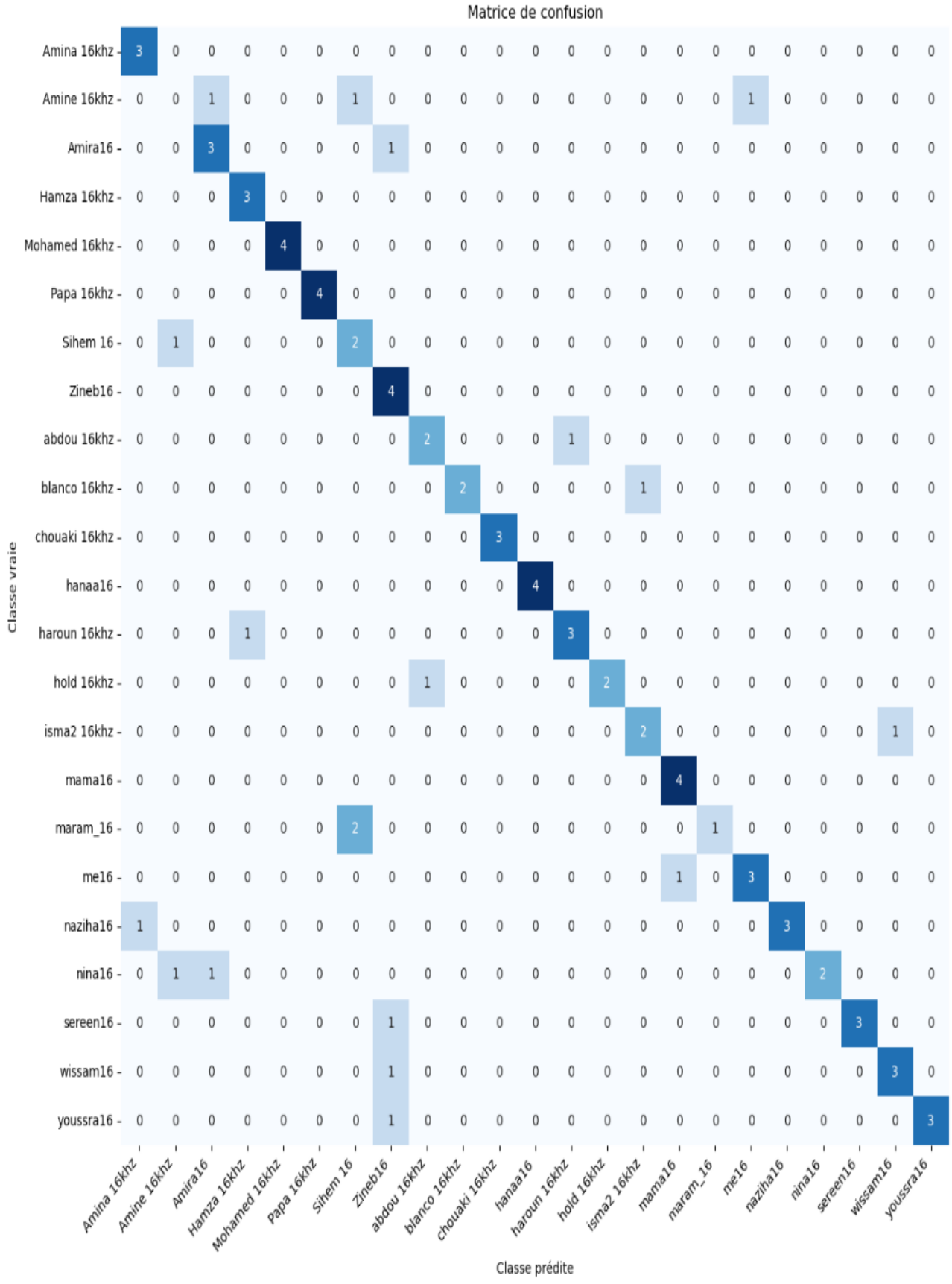

Figure 3.15 : Matricedeconfusionmultilingue16KHz

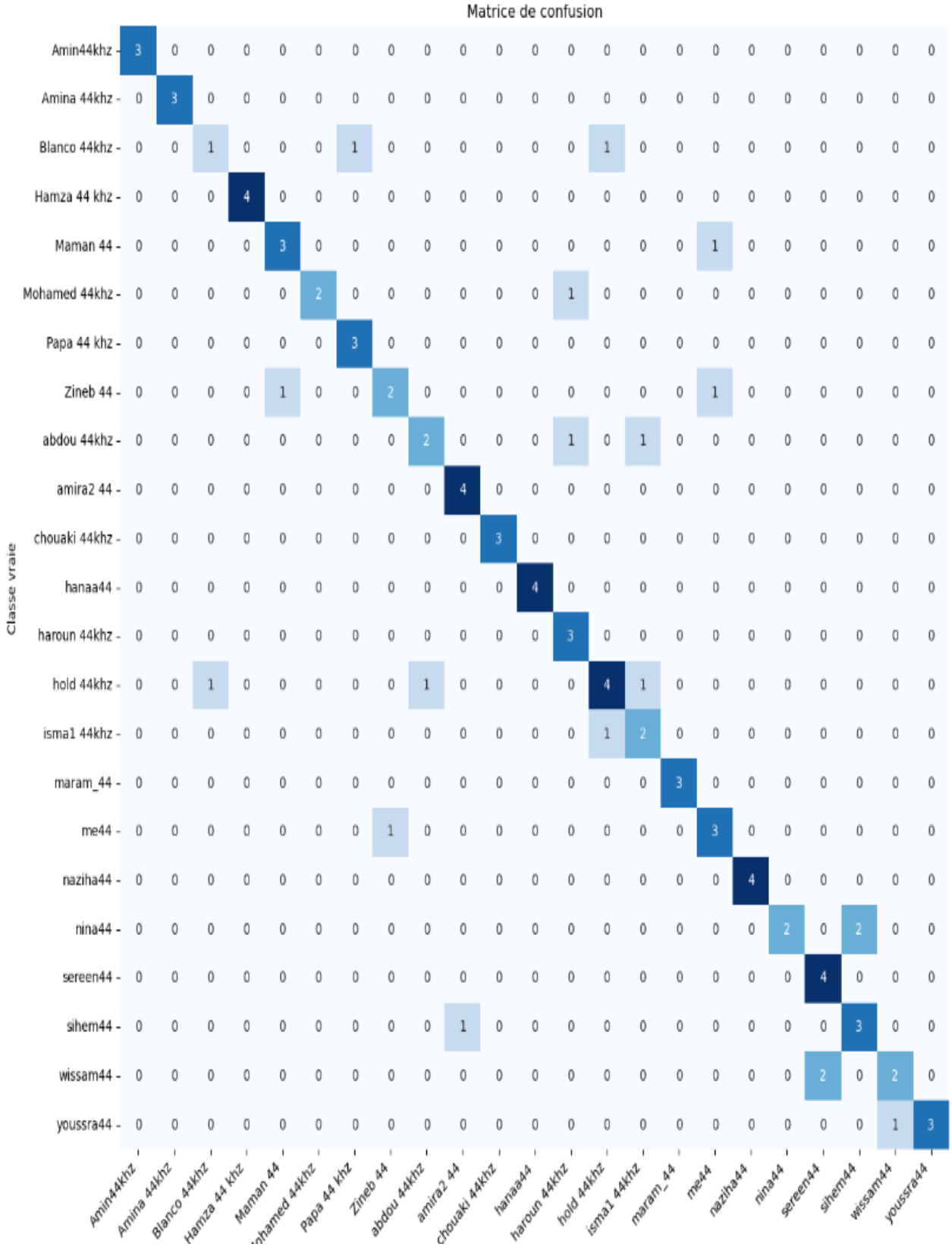

Figure 3.16 : Matricedeconfusionmultilingue44KHz

En comparaison avec la matrice de confusion monolingue à 8 kHz, celles à 16 kHz et 44 kHz présentent moins d'erreurs, car les enregistrements à ces fréquences disposent de d'avantage de caractéristiques nécessaires pour une reconnaissance précise du locuteur.
Les matrices de confusion multilingues affichent également les meilleurs résultats, pour les mêmes raisons évoquées précédemment.

3.3.5. Amélioration du modèle d'identification

Pour booster les performances de notre modèle d'identification, nous avons eu recours à l'augmentation de données (Data Augmentation), une approche qui a fait ses preuves.

L'augmentation des données consiste à améliorer la qualité et la diversité de notre base de données, nous avons développé un script qui applique plusieurs 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 (effectué avec deux amplitudes différentes).

Ces transformations ont permis de quintupler le nombre d'enregistrements, enrichissant ainsi notre base de données.

Afin d'évaluer les performances de notre modèle sur les mêmes ensembles cités préalablement dont les données ont été multipliées. Dans les paragraphes suivants nous présentons les résultats obtenus.

1. Analyse des taux de perte et précision

Les figures 3.17, 3.18, 3.19 ,3.20, 3.21 et 3.22 illustrent respectivement les courbes de perte et de précision en fonction de nombre d'époques, des ensembles d'enregistrement monolingue 8Khz, monolingue 16Khz et monolingue 44Khz augmentées, et multilingue 8Khz, multilingue 16Khz et multilingue 44 Khz augmentées :

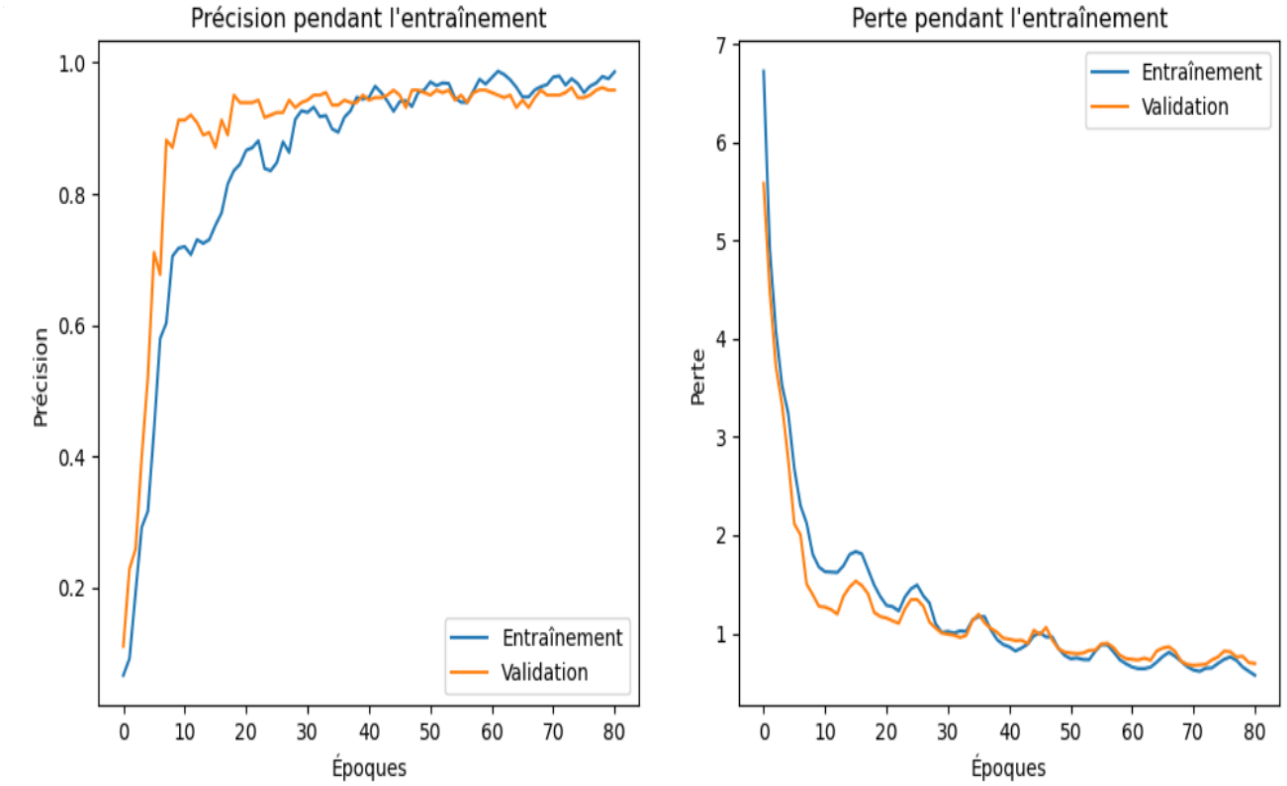

Figure3.17 : Taux de précision et de perte de l'ensemble monolingue à 8Khz Augmenté.

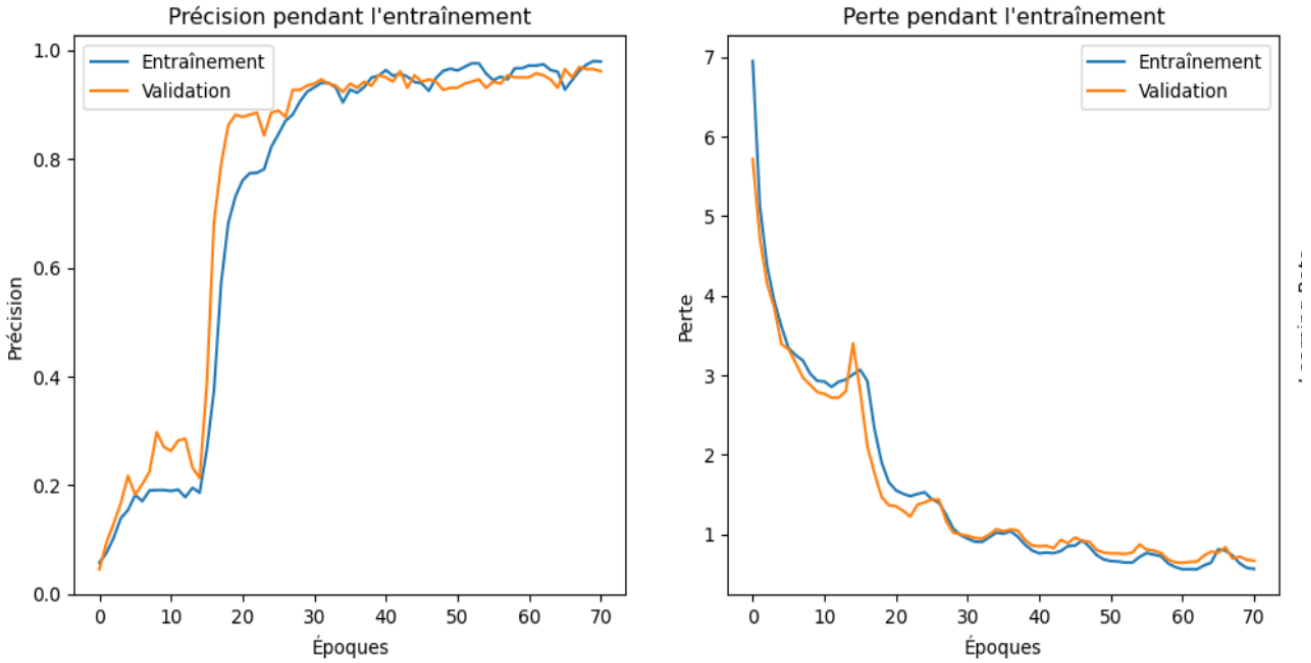

Figure3.18 : Taux de précision et de perte de l'ensemble monolingue à 16Khz Augmentée.

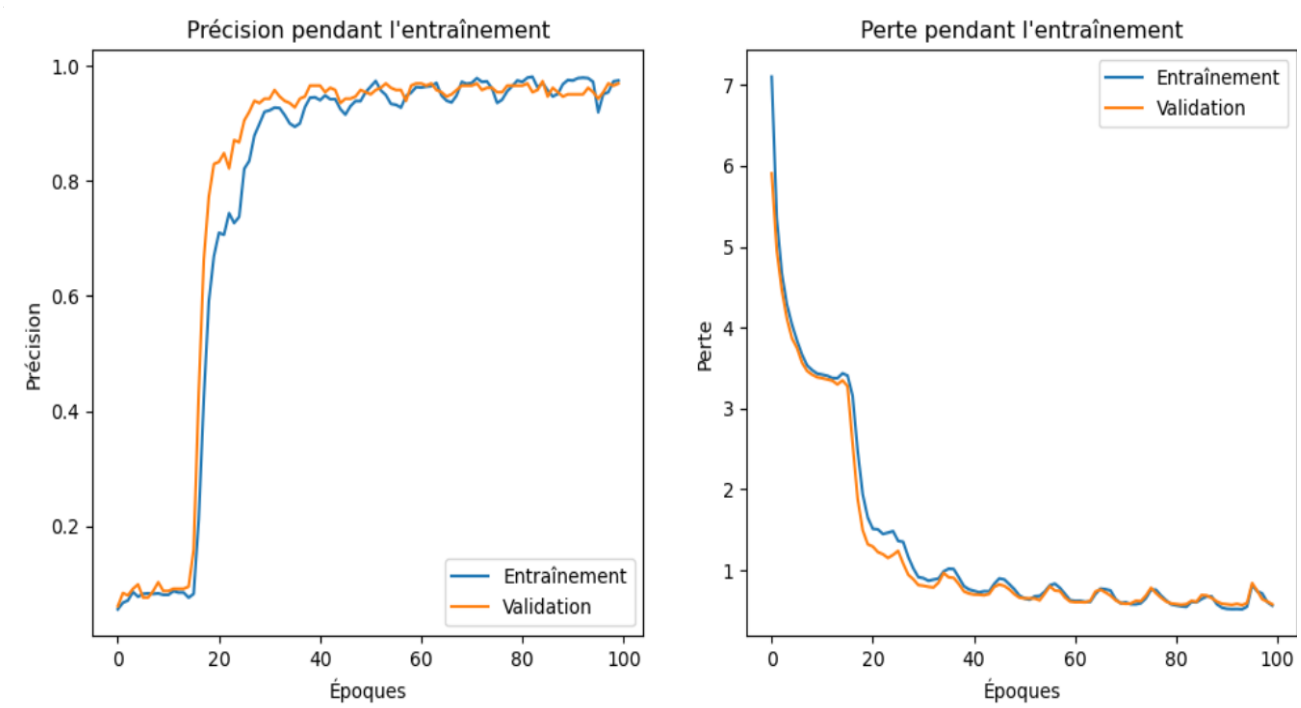

Figure3.19 : Taux de précision et de perte de l'ensemble monolingue à 44Khz Aug

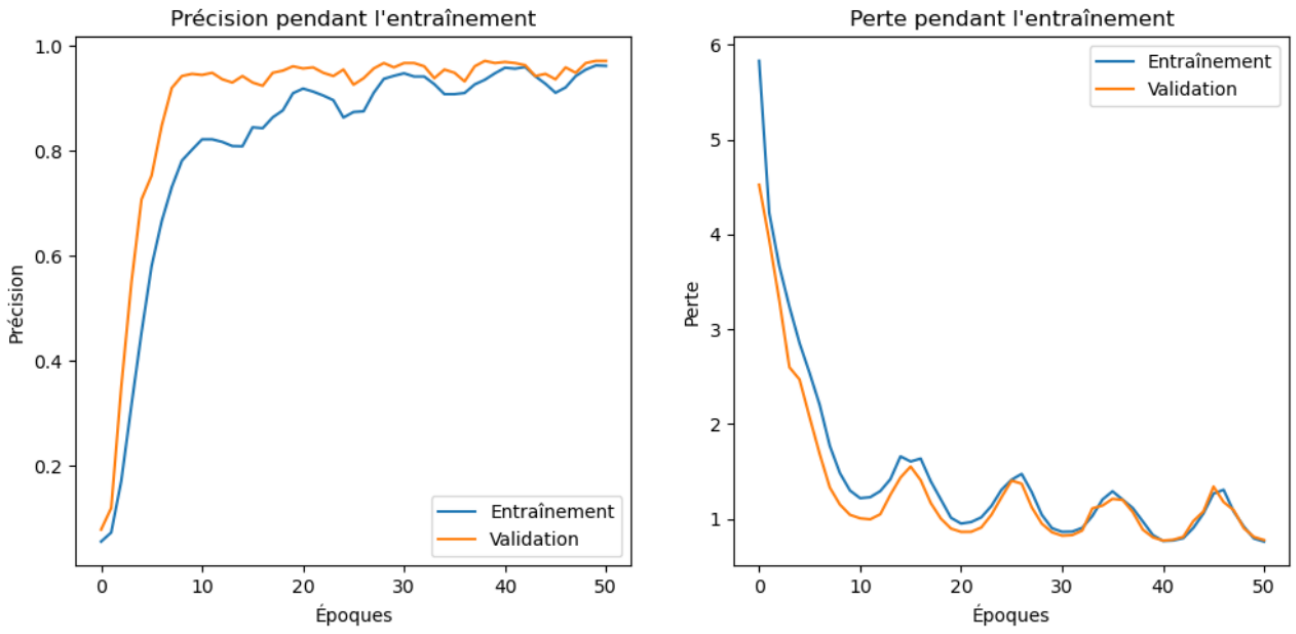

Figure3.20 : Taux de précision et de perte de l'ensemble multilingue à 8Khz Augmentée.

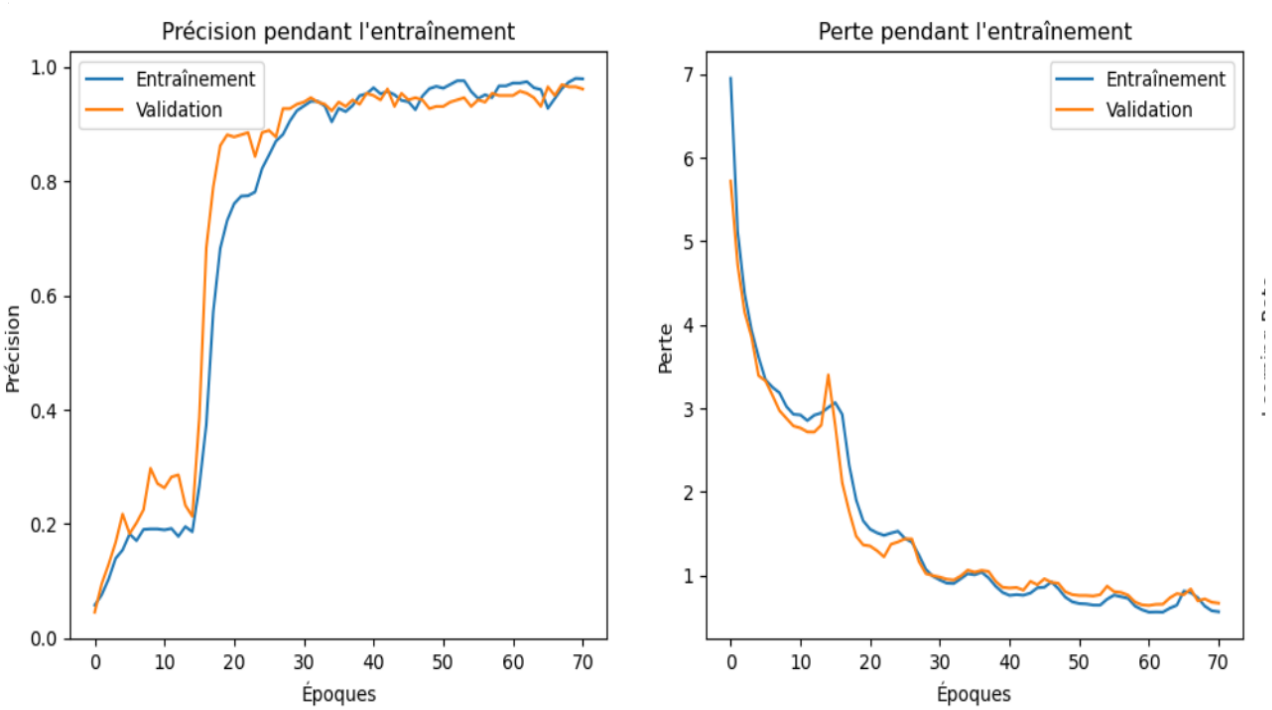

Figure3.21 : Taux de précision et de perte de l'ensemble multilingue à 16Khz Aug

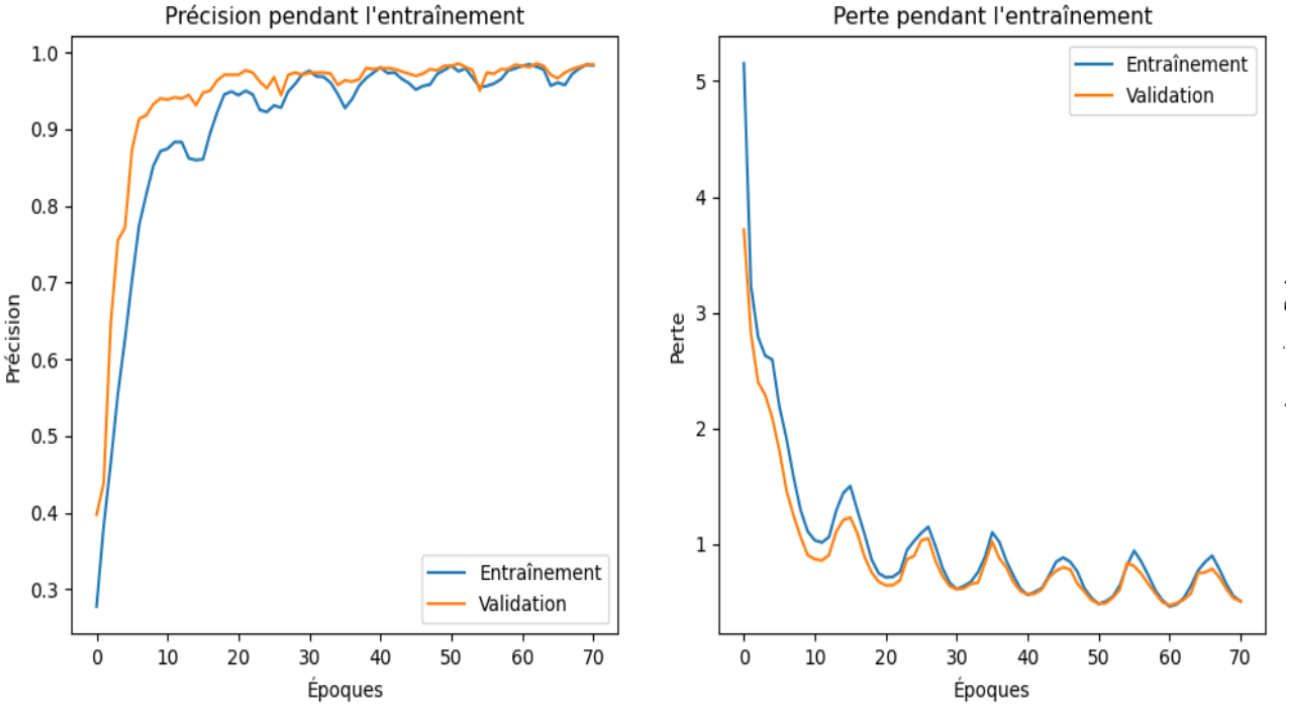

Figure3.22 : Taux de précision et de perte de l'ensemble multilingue à 44Khz Augmentée.

Analyse de la matrice de confusion

Les figures 3.23, 3.24, 3.25 ,3.26, 3.27 et 3.28 illustrent respectivement les matrices de confusion des ensembles d'enregistrement monolingue 8Khz, monolingue 16Khz , et monolingue 44Khz augmentées et multilingue 8Khz, multilingue 16Khz , et multilingue 44Khz augmentées :

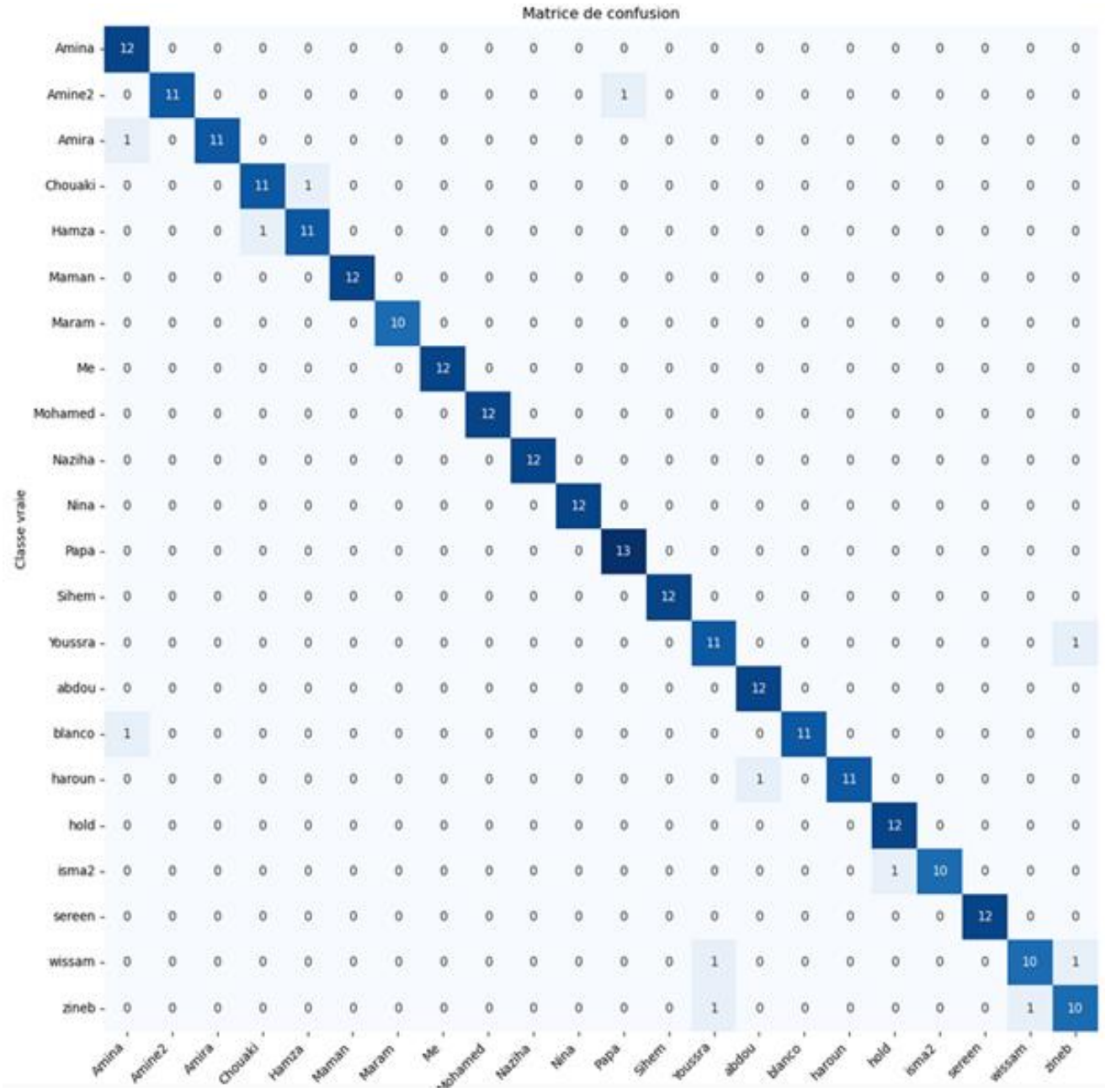

Figure 3.23 : Matricedeconfusionmonolingue8KHzAugmentée

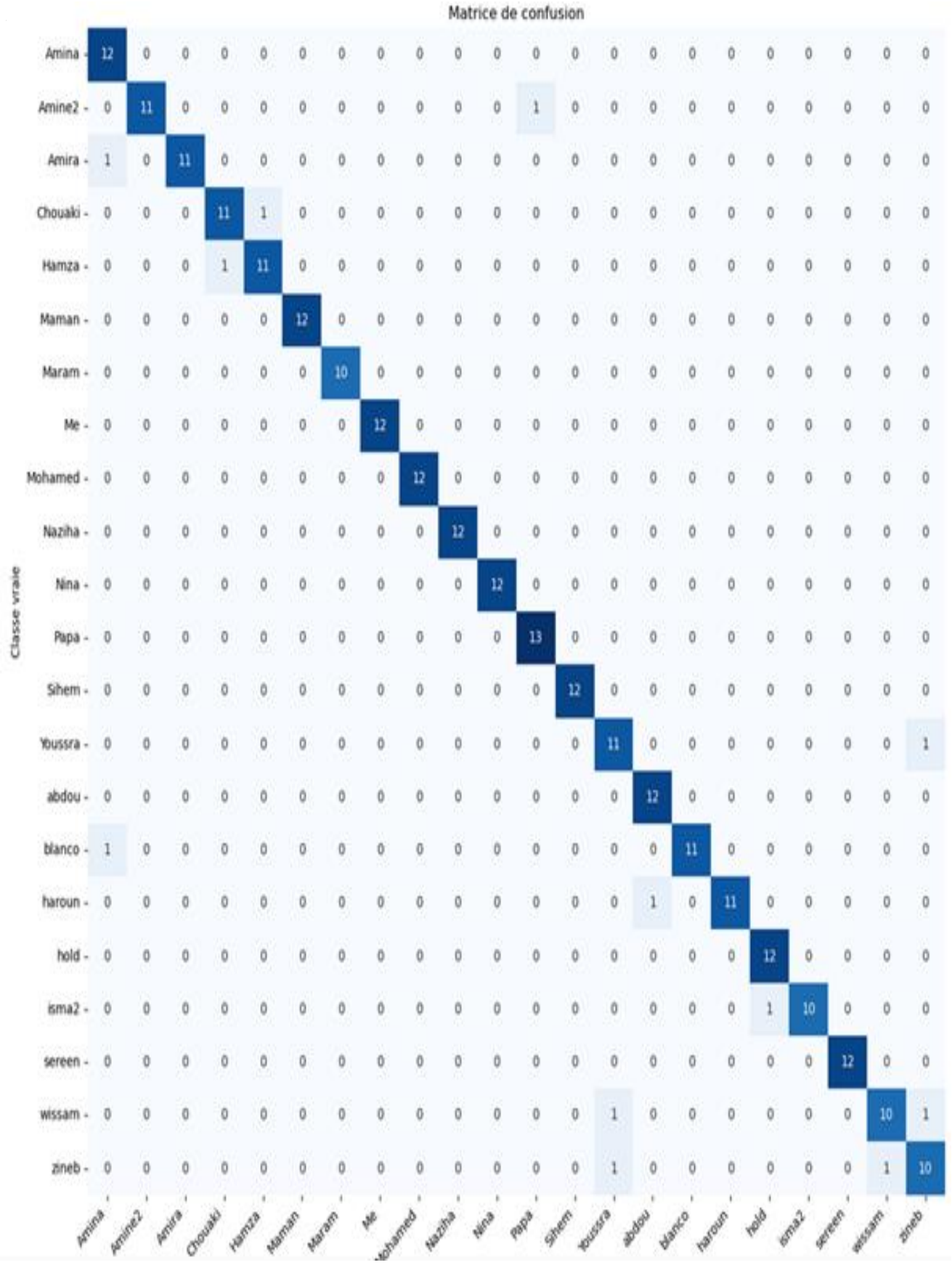

Figure 3.24 : Matricedeconfusionmonolingue16KHzAugmentée.

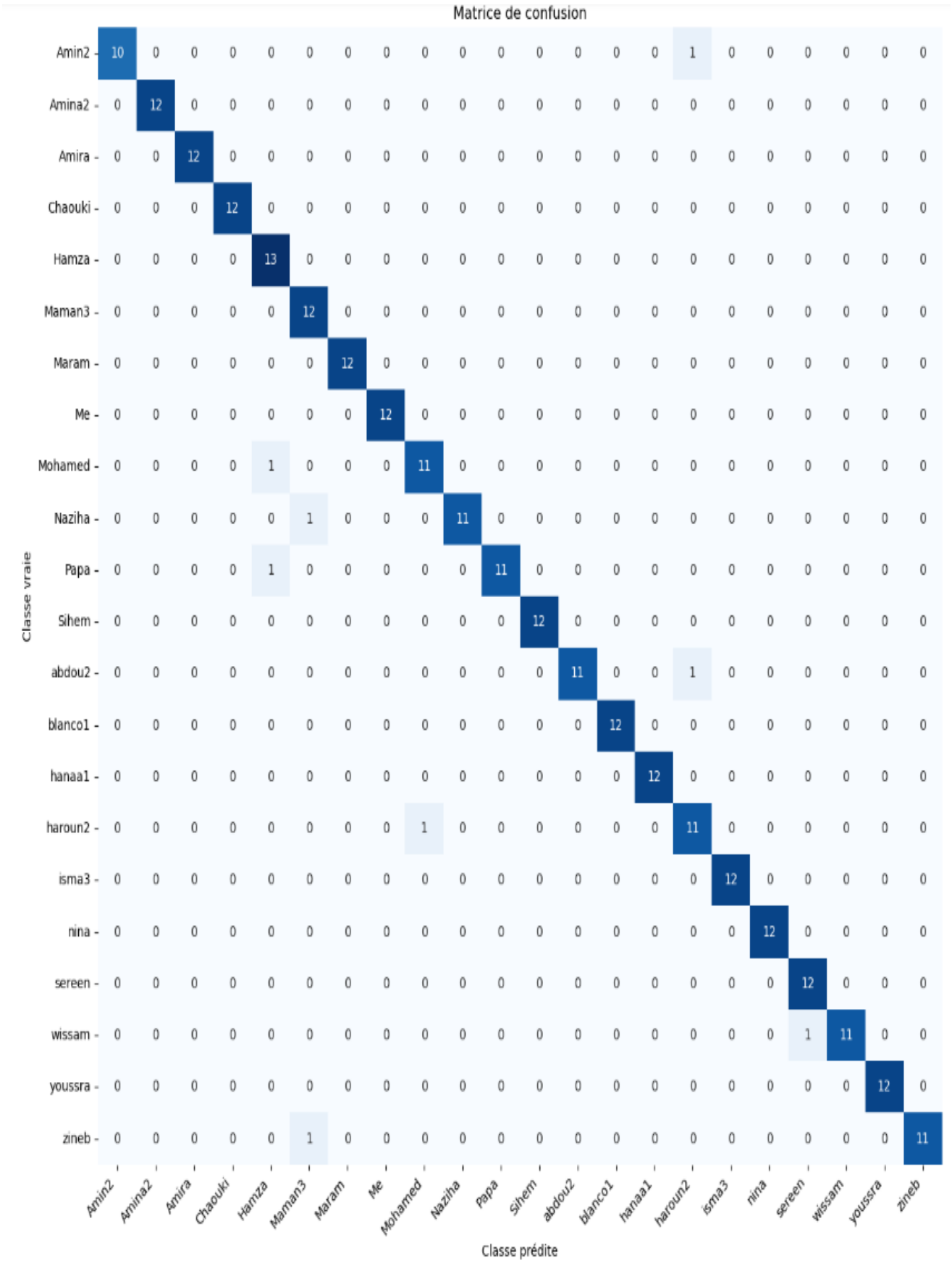

Figure 3.25 : Matricedeconfusionmonolingue44KHzAugmentée.

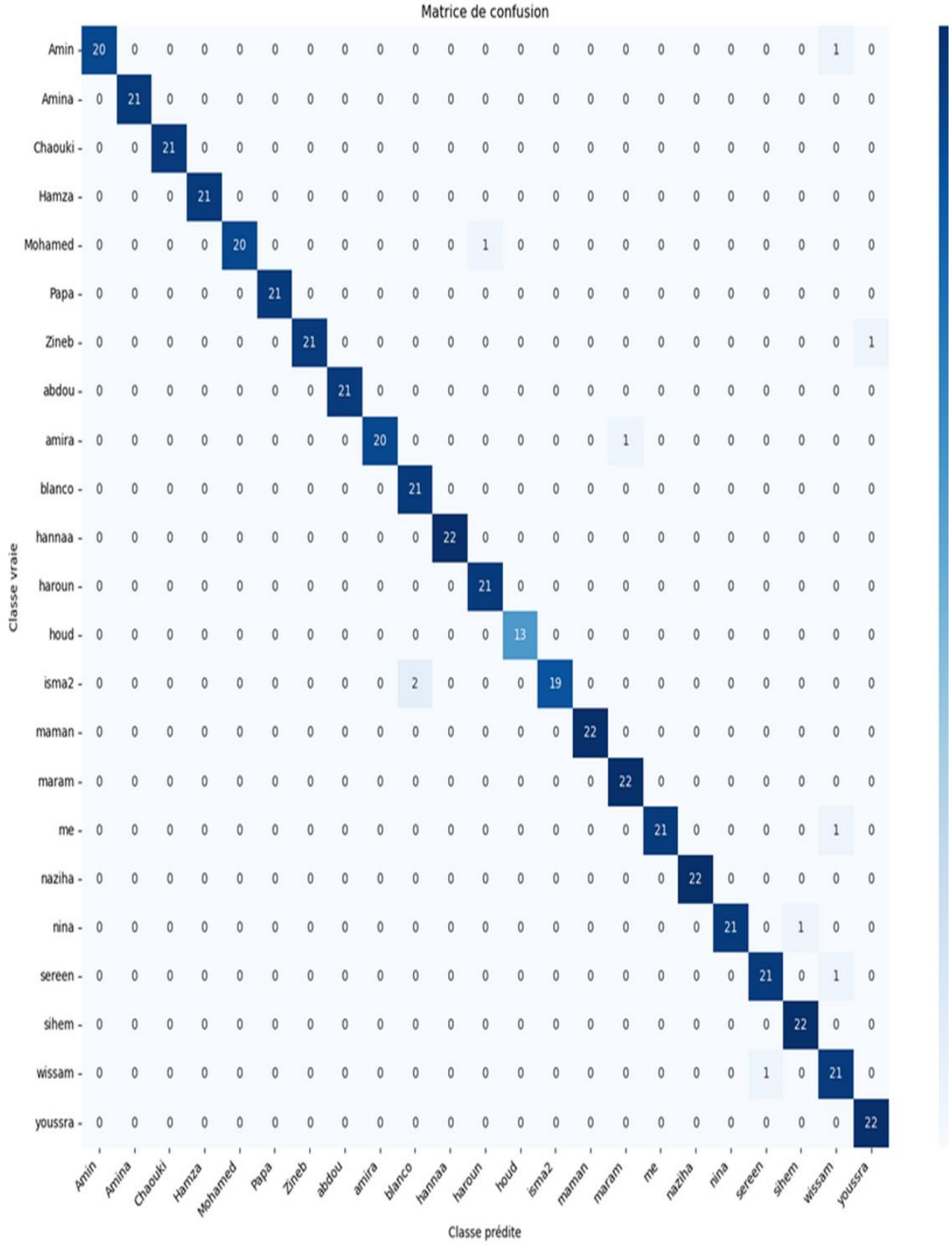

Figure 3.26 : Matricedeconfusionmultilingue8KHzAugmentée.

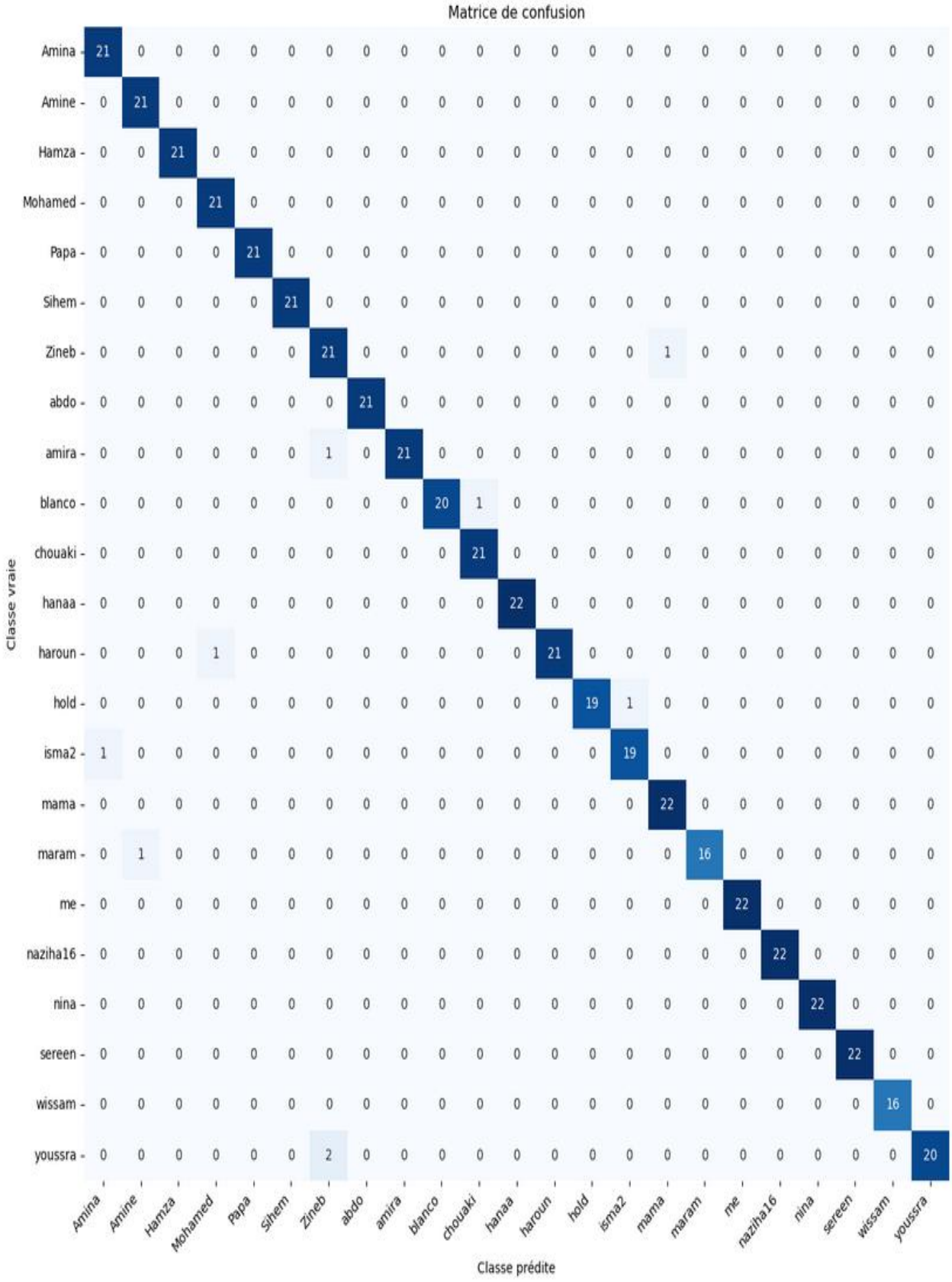

Figure 3.27 : Matricedeconfusionmultilingue16KHzAugmentée.

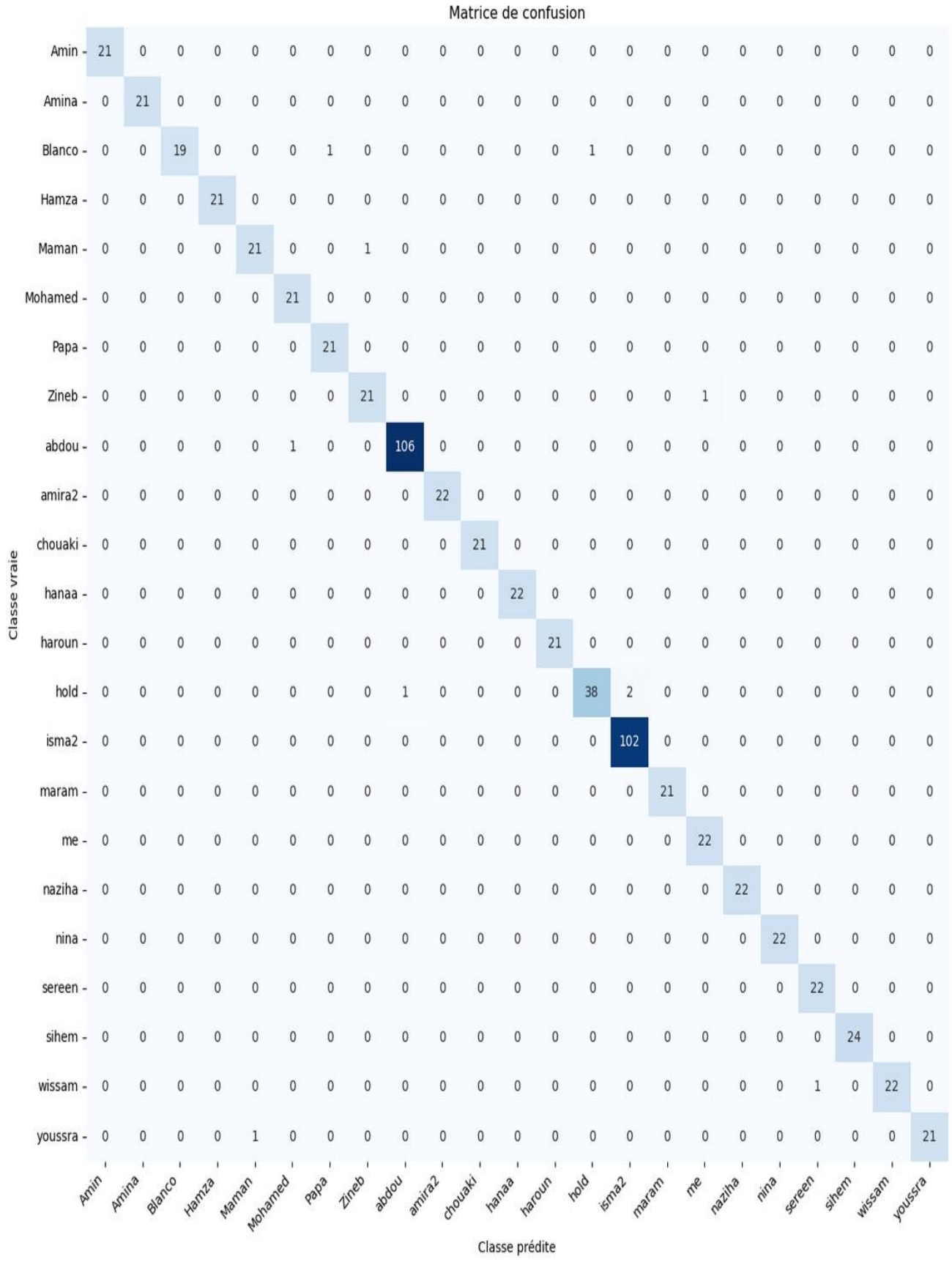

Figure 3.28 : Matricedeconfusionmultilingue44KHzAugmentée.

Après l'augmentation des données, nous remarquons, une très nette amélioration des taux de précision de perte et de classification, ainsi que les matrices de confusion.

Le tableau III. 2 montrent les différents taux de précision de validation et de test pour les différents ensembles d'enregistrement :

	8 KHz		**16 KHz**		**44 KHz**	
	Validation	**Test**	**Validation**	**Test**	**Validation**	**Test**
Monolingue	94	94	95	95	96	95
Multilingue	96	95	96	96	97	97

Tableau3.2 : Taux de précision et test pour les ensembles d'enregistrement monolingue et multilingue après l'augmentation des données.

Nous observons que les six ensembles affichent des taux très satisfaisants, avec un maximum atteignant 97 %, ce qui constitue un résultat très remarquable.

Conclusion :

Dans ce dernier chapitre, nous avons présenté de manière détaillée nos travaux expérimentaux visant à développer et évaluer un système d'identification du locuteur multilingue et multifréquences basé sur l'apprentissage profond CNN combiné avec le GTCC, les résultats obtenus montrent que notre système atteint des performances très satisfaisantes, avec des taux de précision atteignant 97% dans les meilleures configurations. Nous avons également constaté que les configurations multilingues offrent une meilleure robustesse que les configurations monolingues.

CONCLUSION GENERALE ET PERSPECTIVE

Cette recherche a permis de développer un système d'identification automatique du locuteur robuste et efficace, capable de fonctionner dans un contexte multilingue et avec différentes qualités d'enregistrement. Les résultats obtenus, avec des taux de précision atteignant 97%, démontrent la faisabilité et l'efficacité de l'approche proposée. Les contributions scientifiques et techniques de ce travail incluent le développement d'un modèle multilingue robuste, l'adaptabilité aux différentes qualités d'enregistrement et la validation de l'approche CNN-GTCC.

Les résultats obtenus ouvrent de nombreuses perspectives d'applications pratiques, notamment dans les domaines de la sécurité, des télécommunications, de la justice et des assistants vocaux. Cependant, certaines limitations subsistent, telles que la taille de la base de données et la robustesse acoustique. Pour y remédier, des travaux futurs pourraient s'orienter vers l'extension linguistique, l'amélioration de la robustesse acoustique, l'optimisation computationnelle et l'évaluation comparative avec d'autres approches.

Les axes de recherche futurs pourraient inclure :

- ✓ L'intégration d'autres langues et dialectes régionaux.
- ✓ L'évaluation et l'amélioration des performances en conditions bruitées pour renforcer la robustesse.
- ✓ La réduction de la complexité pour des applications temps réel
- ✓ Le développement de prototypes pour de

BIBLIOGRAPHIE

[1] A .Amehray,[Rehaussement de bruitage perceptuel de la parole], thèse de doctorat, école nationale supérieure des télécommunications deBretagne, 2009.

[2] S.K.Singh,P.C.Pandey,«FeatuesandTechniqueForSpeakerRecognition»,Seminar Report, page 5,6,Novembre 03.

[3] 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.

[4] M.DenisJouve,«Reconnaissancedulocuteurenmilieuxdifficiles»,thèsededoctorat, UNIVERSITÉD'AVIGNONETDESPAYSDEVAUCLUSE,18Juillet2017

[5] Y. AZIZA, « modélisation Ar et arma de la parole pour une vérification robuste du locuteurdansunmilieubruitéenmodedépendantdutexte»,MémoiredeMagister,Université Ferhat Abbas, Sétif, 2013.

[6] H. Satori, M. Harti and N. Chenfour, « Système de Reconnaissance Automatique del'arabe basé sur CMUSphinx », mémoire master, DharMehraz Fès Morocco.

[7] OthmanLachhab,«ReconnaissanceStatistiquedelaParoleContinuepourVoix Laryngée etAlaryngée », Université Mohammed V de Rabat (Maroc), 2017.

[8] Vincent Jousse, « Identification nommée du locuteur : exploitation conjointe du signal sonore et de sa transcription », mémoire master, Université du Maine, 2011.

[9] MichaelFMcTear,«Spokendialoguetechnology:towardtheconversationaluser interface. Springer Science & Business Media », article, page 3, 2004.

[10] M. A. Wissmann et K. M. Béring, « Automatique Language Identification », Speech Communication, article, page 4, 2001.

[11] Mr.Haddab,«reconnaissanceautomatiquedulocuteurparlaméthodedutauxpassage par zéro », mémoire master, université Mouloud mamri de Tizi-Ouzou, 2007/2008Jin,Minho,andYoo,ChangD,«SPEAKERVERIFICATIONANDIDENTIFICATION

»,KorêInstitutAvancédesSciencesetTechnologies,RépubliquedeCorée,2004.

[12] SiwarZRIBIBOUJELBENE,«IdentificationduLocuteurparSystèmeHybrideGMMSMO », thèse, TUNISIA, March 22/26/2009.

[13] Dr.ClintSlatton,«ASpeakerVerificationSystem»,thèse,UniversitédeFlorida,2006.

[14] A. Preti, « Surveillance de reseaux professionnels de communication par la reconnaissance du locuteur », Thèse, Université d‟Avignon et des Pays de Vaucluse, France, 2008.

[15] (consulté le 12/06/2023), disponible sur : https://www.editionseni.fr/open/mediabook.aspx?idR=f6e7a7353a3574180124387fa03fdcl,
[16] AmineAbdaoui,«MachineLearning »,article,page5, 1/7/2019.

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

[18] Dr. Ouarda ZEDADRA, « Système de prédiction de la consommation d'énergie basé Deep Learning », Mémoire master, Université de 8 Mai 1945, Septembre 2021.

[19] Pr.BILAMIAzeddine,«ApprentissageIncrémental&MachinesàVecteursSupports »,UniversitéHADJLAKHDAR–BATNA,18/12/2013

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

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

[22] M.AbderrahmaneAdjila,«Détectiond'activitévocaleutilisantl'apprentissageprofond »,Mémoiremaster,UniversitédeGhardaïa, 2019/2020.

[23] Apprendreprogrammationcourspython3,(consultéle16/06/2023)disponiblesur :https://python.doctor/.

[24] consulté le 18/06/2023, disponible sur : https://www.data-bird.co/blog/langagepython#toc-que-peut-on-faire-avec-python-

[25] consultéle22/06/2023disponiblesur:https://datascientest.com/kaggle-tout-ce-quil-asavoir-sur-cette-plateforme.

[26] X. Valero and F. Alias, “GammatoneCepstralCoefficients: BiologicallyInspiredFeatures for Non-SpeechAudio Classification,” IEEE Transactions on Multimedia

[27] Jia-Ming Liu1, Mingyu You1*, Guo-Zheng Li1, Zheng Wang1, Xianghuai Xu2, Zhongmin Qiu2, Wenjia Xie1, Chao An1, Sili Chen1"COUGH SIGNAL RECOGNITION WITH GAMMATONE CEPSTRAL COEFFICIENTS"china2013

[28] S. S. Thomas, M. S. R. Aravind, P. B. S. Kumar, "Comparison of MFCC and GTCC for classification of humanemotions", 2019 International Conference on Intelligent SustainableSystems (ICISS).