الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne démocratique et populaire

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

جامعة سعد دحلب البليدة
Université SAAD DAHLAB de BLIDA

كلية التكنولوجيا
Faculté de Technologie

قسم الإلكترونيك
Département d'Électronique

Mémoire de Master

Filière : Télécommunication

Spécialité : Systèmes des télécommunications

Présenté par

BEGHA Chayma

&

ZIOUANE Abir

Deep learning pour la reconnaissance automatique du locuteur

Proposé par : Mr. ABED Ahcéne

Année Universitaire 2021-2022

Remerciements

Au terme de ce travail, nous tenons à remercier en premier lieu Dieu (Allah) qui nous a donné la force, la volonté et le courage pour terminer cette mémoire

Dieu merci !

Toute notre gratitude et nos vifs remerciements vont à notre promoteur Mr ABED Ahcéne Enseignant à l'Université de Blida pour avoir assuré l'encadrement de ce travail. Pour leurs aides, ses conseils et son suivi durant la réalisation de notre projet.

Un grand remerciement aux membres du jury chacun par son propre nom pour l'honneur et l'intérêt qu'ils nous ont accordé en acceptant d'examiner et d'évaluer notre mémoire.

Nos remerciements Vont particulièrement à nos parents et à toute notre famille pour leurs soutiens et amour inconditionnels et leurs encouragements durant tout notre parcours.

Dédicaces

Avec l'expression de ma reconnaissance, je dédie ce modeste travail a ceux qui quels que soient les termes embrassés, je n'arriverais jamais à leur exprimer mon amour sincère.

A ma très chère mère, quoi que je fasse ou que je dise, je ne saurai point te remercie comme il se doit. Ton affection me couvre, ta bienveillance me guide et ta présence à mes cotes a toujours été ma source de force pour affronter les différents obstacles je t'aime très fort.

A mon cher papa merci pour votre soutien et encouragement je t'adore.

A mon adorable petite sœur Hadil ma joie et le bonheur de ma petite famille.

A ma sœur Ikram et mon frère Djamel qui me manque infiniment aussi mon ange Amira je vous aime.

A mon fiancé Oussama merci pour tes conseils et ton soutiens durant toutes ces années.

A ma chère cousine chaima tu es la meilleure, et à Afef je vous aime.

A toutes mes copines sans exception Sonia Sara Aicha Roumaissa bouchra zahra bahia merci pour votre présence dans ma vie .

A mes tantes cousins et cousines je vous aime.

Je termine ma dédicace et le meilleur pour la fin chère binôme tu n'es pas seulement une binôme tu es une sœur merci pour ta patience ta compréhension tout au long de ce projet j'ai de la chance de t'avoir dans ma vie.

ZIOUANE ABIR

A mon très cher Père

Aucune dédicace ne pourrait exprimer l'amour et le respect que j'ai toujours pour vous, rien au monde ne vaut les efforts fournis jour et nuit pour mon éducation et mon bien être.

A ma très chère Mère

Tu représente pour Moi la source du bonheur et l'exemple de dévouement qui n'a pas cessé de m'encourager .tu as fait plus qu'une mère puisse faire pour que ses enfants suivent le bon chemin dans leur vie et leurs études.

Qu'Allah le tout puissant vous protège et vous procure santé, bonheur et longue vie pour que vous demeuriez le flambeau illuminant le chemin de vos enfants.

A mes Très chers Frères AYOUB et ABDERRAHMEN

Qui n'ont pas cessée de me conseiller, encourager et soutenir

Que Dieu les protège et leurs offre la chance et le bonheur

A ma Très chère cousine Nour

Tu es ma petite sœur et ma moitié qui me donne de l'amour et de la vivacité je t'aime

A ma Meilleure Abir

Ma copine mon bras droit et ma sœur Merci d'être Toujours Prés de moi dans mes joies et mes peines très heureuse de passer ces années avec toi merci d'être celui sur qui je peux toujours compter je t'aime

A ma meilleure Chaima

Les bons amis sont une bénédiction. Vous êtes la plus grande bénédiction de tous. ,avec une amie comme toi à mes côtés, il n'y a vraiment rien que nous ne puissions accomplir. Merci de m'avoir fait ressentir ça.

A ma meilleure Selsabil

Tu es l'une des rares personnes dans ma vie qui m'a aidé à devenir ce que je suis aujourd'hui. Merci pour ça je t'aime.

A mon très cher Ami

Je te remercie grandement pour l'aide que tu m'as apporté pour finir ce travail pénible. Avec toi tout semble plus facile

BEGHA Chayma

ملخص :

شهدت التكنولوجيا الحديثة تطورًا كبيرًا في مجال التفاعل بين الإنسان والآلة، ومن يوم لآخر يصبح دور الآلة أكثر أهمية. تتضمن هذه التقنيات التعرف التلقائي على المتكلم وهو أحد أكثر موضوعات طلبا ، ويطلق عليه أحيانًا القياسات الحيوية للمتكلم. في هذا العمل، نقدم نظاما المعرفة الآلية للمتكلم من خلال التعلم العميق للصوت. باستخدام أحد أكثر أنواع المعاملات الصوتية شيوعًا في أنظمة التعرف الآلي على الكلام وهي معاملات(MFCCs). في هذا العمل نحن مهتمون بالتعلم العميق (التعلم العميق) الذي يستند إلى شبكات عصبية تلافيفية(CNN) وأنظمة Sinc و Constantsinclayer الأكثر استخدامًا في هذه الأنظمة، للحصول في نهاية العملية على هوية المتكلم انطلاقا من الصوت.

كلمات المفاتيح:

التعرف الآلي على المتحدث، المعاملاتMFCC، شبكات الخلايا العصبية التلافيفيةCNN ، التعليم المعمق.

Résumé:

La technologie moderne a connu une grande évolution dans le domaine l'interaction entre l'homme et la machine, de jour en jour le rôle de la machine devient plus important. Parmi ces technologies on trouve la reconnaissance automatique du locuteur qui est l'un des thèmes d'application les plus fertiles, parfois appelée biométrie du locuteur. Dans ce travail, Nous présentons l'identification du locuteur par deep learning à partir de la voix. En utilisant l'un des types de paramètres acoustique les plus utilisés dans les systèmes de reconnaissance automatique du locuteur qui sont les (MFCC). Nous nous intéressons à l'apprentissage profond (deep learning) qui est basé sur les réseaux de neurones convolutifs (CNN) et les systèmes Sinc et Constantsinclayer qui sont les plus utilisés pour identifier les locuteurs par l'apprentissage profond (deep learning) , pour obtenir à la fin d'identification la classification de chacun des systèmes précédents et nous avons conclu lequel est le meilleur pour identifier un locuteur .

Mots clés :

Reconnaissance automatique du locuteur, MFCC, CNN, deep learning,

Abstract :

Modern technology has undergone a great evolution in the field of human-machine interaction, and the role of the machine is becoming more and more important. Among these technologies we find the automatic recognition of the speaker, which is one of the most fertile themes of application, sometimes called biometrics speaker recognition biometrics. In this work, we present speaker identification by deep learning from voice. Using one of the most used types of acoustic parameters in automatic speaker recognition systems which are (MFCC). We are interested in deep learning which is based on convolutional neural networks (CNN) and Sinc and Constantsinclayer systems which are the most used to identify speakers by deep learning, to obtain at the end of identification the classification of each of the previous systems and we concluded which is the best to identify a speaker.

Keywords:

Automatic Speaker Recognition, MFCC,CNN, deep learning,

Table des matières

Liste des Abréviations

API	:	Application Programming Interface
CNN	:	Convolutional Neural Network
dB	:	décibel
FFT	:	Fast Fourier Transform
HMM	:	Hidden Markov Model
IA	:	Intelligence artificielle
IAL	:	Identification automatique du locuteur
LPC	:	Linear Predictive Coding
MFCC	:	Mel-Frequency Cepstrum. Coefficients
PLP	:	Perceptuel Linear Prediction
RAP	:	Reconnaissance Automatique de la Parole
RNN	:	Recurrent Neural Networks
WSJ	:	Wall Street Journal
F_0	:	Fréquence Fondamentale

Liste des Figures

Introduction Générale

L'objectif principal dans ce mémoire est de construire un système de reconnaissance automatique du locuteur. La parole est incontestablement le mode de communication le plus naturel que les humains utilisent pour interagir entre eux. Dans ce contexte, concevoir une machine qui imite le comportement humain, en particulier la capacité d'utiliser la parole d'une manière naturelle et répondre correctement au langage parlé, a attiré l'attention des ingénieurs et des scientifiques durant le dernier siècle [1].

La reconnaissance automatique du locuteur est le processus de reconnaitre une personne à l'aide de sa voix, en se basant sur l'information véhiculée par le signal de parole. Ce domaine regroupe les taches relatives à la vérification du locuteur et son identification.

En vérification du locuteur, le système décide si une identité proclamée caractérise un client ou un imposteur [2]. En identification, le système identifie une personne parmi un ensemble connu par le système.

Les différents types de paramètres acoustiques couramment cités dans la littérature sont les coefficients : LPCC, PLP, MFCC er LPC. Généralement les coefficients MFCC sont les paramètres acoustiques (caractéristiques) les plus utilisés dans les systèmes RAP.

Pour arriver à identifier un locuteur, on peut utiliser des différents modèle d'apprentissage profond (deep learning) où la machine aura une base de données remplie par des voix des différents locuteurs Pour arriver à identifier un locuteur, on peut utiliser des différents modèle d'apprentissage profond (deep learning) où la machine peut d'une manière automatique savoir à quel locuteur appartient l'extrait de parole. En intégrant a ce système une base de données remplie par des voix de différents locuteurs, la machine fait l'analyse redondant de cette base de données pour chaque voix entrée et elle donne des résultats d'identification de locuteur avec une précision qui varie selon le modèle utilisé lors de l'apprentissage.

Ce mémoire est constitué de trois chapitres :

Le premier chapitre représente une vue générale sur la reconnaissance du locuteur et ces différentes taches aussi comment fonctionne le système de reconnaissance de locuteur ainsi on a parlé sur les avantages de RAP.

Le deuxième chapitre aborde les principales techniques de la reconnaissance automatique de la parole et les différentes applications, nous avons décrit l'architecture du système après nous parlons sur l'extraction des paramètres MFCC.

Le Troisième chapitre, explique comment identifier le locuteur en utilisant trois modèles du deep learning sur Matlab .

Chapitre 1

Reconnaissance automatique du locuteur

Introduction

La parole est un moyen de communication très efficace et naturel utilisé par l'humain. Depuis longtemps, il rêve de pouvoir s'adresser par ce même moyen à des machines ce qui les rend plus intelligentes.

Cependant, les recherches consacrés dans l'essai de créer une telle machine intelligente qui peut reconnaître le mot parlé, comprend sa signification ou reconnaître la personne qui parle, sont loin d'être fonctionnelles. Dans ce chapitre nous allons présenter quelques généralités sur les systèmes de reconnaissance automatique de parole et particulièrement les systèmes de reconnaissance automatique du locuteur (RAL). Nous essayons de donner une idée générale sur la RAL, de discuter les problèmes de cette dernière, ainsi que les méthodes utilisées pour résoudre ces problèmes.

Système de reconnaissance de la parole

La reconnaissance automatique de la parole pose de nombreux problèmes d'un point de vue théorique. Leurs complexités faites que seuls des sous-problèmes ont pu être à ce jour résolus. Ces solutions partielles correspondent à des contraintes plus ou moins fortes, et les systèmes existants supposent une coopération plus ou moins grande des utilisateurs.

Pour classer les systèmes de reconnaissance automatique, on a généralement recours aux critères suivants :

- Le mode d'élocution : des syllabes ou mots isolés aux mots connectés, jusqu'à une parole dite, "continue" c'est-à-dire sans pauses artificielles.
- Taille du vocabulaire et difficulté de la grammaire (la complexité du langage autorisé).
- La dépendance plus ou moins grande vis-à-vis du locuteur.
- L'environnement protégé ou non (la robustesse aux conditions d'enregistrement).

En outre, il n'est pas inintéressant de les différencier selon deux points qui ont aussi leur importance :

- La compréhension est-elle requise ou non ? (Un système de compréhension cherche à accéder à la signification de l'énoncé parlé).

- Le discours est-il naturel, ou la syntaxe des phrases doit-elle être contraignante ?

Les systèmes réalisés de la RAP, sont conçus pour des applications spécifiques. Cela conduit à une restriction de l'univers du dialogue homme-machine.

L'objectif essentiel des études sur la reconnaissance et la compréhension automatique de la parole est « de permettre, à terme, un dialogue le plus naturel possible entre l'homme et la machine, dans le cadre d'une application spécifique»[3].

1.1.1 Domaine d'applications de la RAP

Les applications de la reconnaissance de la parole sont nombreuses. Elle libère complètement l'usage de la vue et des mains, et laisse l'utilisateur libre de ses mouvements. La vitesse de transmission des informations est supérieure, dans la RAP à celle que permet l'usage du clavier. Enfin tout le monde ou presque sait parler, alors que peu de gens sont à l'abri des fautes de frappe et d'orthographe, etc.

Ces avantages sont tellement importants que l'on trouve déjà sur le marché des dispositifs d'utilisation limitée, mais néanmoins efficaces.

Citons certaines applications qui ont déjà vu le jour :

- Saisie vocale de données.
- Donne des ordres tout en pilotant une automobile ou un avion.
- Aide aux handicapés.
- Chambre d'hôpital avec possibilités de commandes vocales pour le malade.
- Commande vocale de machines ou de robots.
- Commande vocale d'une montre portable, etc.

1.1.2 Complexité du problème

Pour comprendre le problème de reconnaissance automatique de la parole, il est bon d'en comprendre les différents niveaux de complexités et les différents facteurs qui en font un problème difficile.

Le signal de parole est des plus complexes, et sujet à beaucoup de variabilités. Ayant été produit par un système phonatoire humain complexe, il n'est pas facile de le caractériser à partir d'une simple représentation bidimensionnelle de propagation des ondes. On peut

distinguer certaines de ses caractéristiques comme les sons élémentaires ou phonèmes, la hauteur, le timbre, l'intensité, la vitesse... Mais en réalité, la voix est beaucoup plus complexe qu'on ne peut percevoir par l'oreille. L'onde sonore varie non seulement avec les sons prononcés, mais également avec les locuteurs. La très grande variabilité que peut présenter un même discours selon la façon de parler du locuteur qui peut chanter, créer, murmurer, être enroué ou enrhumé, et aussi selon le locuteur lui-même (homme, femme, enfant, voix nasillarde, différences de timbre), sans parler des accents régionaux, rend très délicate la définition d'invariants. Il faut pouvoir séparer ce qui caractérise les phonèmes, qui devrait être une constante quel que soit le locuteur et sa prononciation, de l'aspect particulier à chaque locuteur. Qu'est ce qui permet à notre cerveau de distinguer un mot d'un autre, indépendamment de celui qui nous parle ?

De plus, la mesure du signal de parole est fortement influencée par la fonction de transfert du système de reconnaissance (les appareils d'acquisition et de transmission),ainsi que par le milieu ambiant. Ainsi l'obstacle majeur d'avoir une grande précision de la reconnaissance, est la grande variabilité des caractéristiques d'un signal vocal. Cette complexité du signal de parole provient de la combinaison de plusieurs facteurs, la redondance du signal acoustique, la grande variabilité inter et interlocuteur, les effets de la coarticulation en parole continue, et les conditions d'enregistrement [4].

1.1.3 Paramétrisation du signal de parole

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

Analyser un tel signal est une tâche difficile vu le grand nombre de paramètres associés. Néanmoins, trois principaux paramètres s'imposent : la fréquence fondamentale, le spectre fréquentiel et l'énergie. Ces paramètres sont appelés traits acoustiques et sont énumérés ci-après [5, 6]

1.1.3.1 Fréquence fondamentale (F0)

C'est une caractéristique acoustique propre à chaque personne. Elle est fonction de plusieurs paramètres physiologiques tels que le volume de la glotte et la longueur de la trachée. Elle se définit par la cadence du cycle d'ouverture et de fermeture des cordes

vocales pendant la phonation des sons voisés. La fréquence fondamentale varie d'un locuteur à un autre selon le genre et l'âge comme suit [7] :

- de 80 Hz à 200 Hz pour une voix d'homme.
- de 150 Hz à 450 Hz pour une voix de femme.
- de 200 Hz à 600 Hz pour une voix d'enfant.

1.1.3.2 Spectre fréquentiel

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

1.1.3.3 Energie

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

1.1.3.4 Spectrogramme

Le spectrogramme est un diagramme associant à chaque instant t d'un signal, son spectre de fréquence. Les spectrogrammes sont utilisés pour identifier des sons, comme des cris d'animaux et des sons d'instruments musicaux. Ils sont largement utilisés dans le domaine de la reconnaissance de la parole.

On peut dire qu'Un spectrogramme affiche la force d'un signal dans le temps aux différentes fréquences d'une forme d'onde. Les spectrogrammes peuvent être des graphiques bidimensionnels avec une troisième variable représentée par des couleurs ou des graphiques tridimensionnels avec une quatrième variable de couleur.

1.1.3.5 Amplitude

L'amplitude est une mesure de son changement dans une seule période (telle que le temps ou la période spatial). L'amplitude d'un signal non périodique est son amplitude comparée à une valeur de référence. Il existe différentes définitions de L'amplitude qui sont toutes fonctions de L'amplitude des différences entre les valeurs extrêmes de la variable. Dans les textes plus enceins, la phase d'une fonction périodique est parfois appelée l'amplitude [8].

Reconnaissance automatique du locuteur RAL

Une première façon de catégoriser les systèmes de reconnaissance du locuteur consiste à les classer selon les tâches qu'ils ont à accomplir.

1.1.4 Vérification automatique du locuteur

Il s'agit de vérifier que le locuteur qui parle est bien le locuteur prévu ou proclamé. Différentes terminologies, utilisées dans la littérature, pourraient avoir la même définition que la vérification du locuteur (VL), telle que la vérification de la voix, l'authentification de la voix et l'authentification du locuteur.

Elle effectue une comparaison (elle s'appelle également la décision binaire) entre les paramètres d la voix d'entrée et ceux des voix proclamées qui sont enregistrées dans une base de donnée.

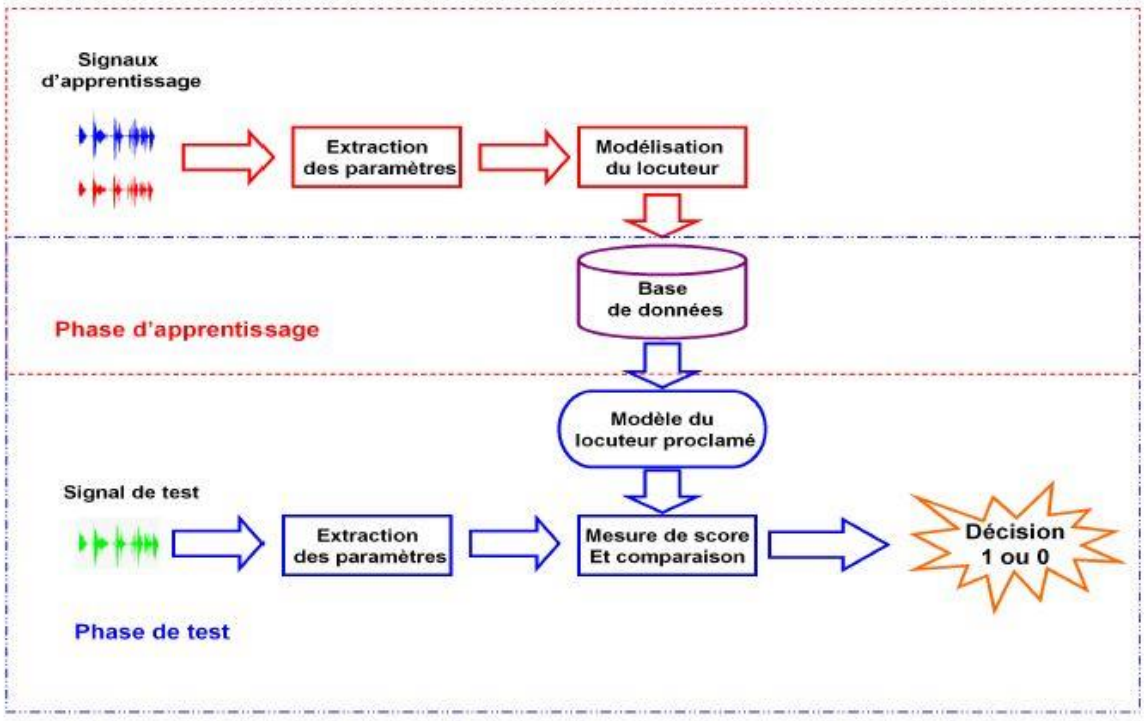

Figure 1.1 Système de vérification du locuteur

La figure 1.1 montre la structure de base d'un système de VAL .Il y a trois modules principaux : prétraitement du signal d'entrée, modélisation du locuteur et la phase de test.

Le prétraitement du signal d'entrée est employé, pour la mise en forme du signal à fin d'extraire les paramètres pertinents. Cette ´étape est appelée aussi analyse acoustique.

Après cette opération, nous obtenons des vecteurs acoustiques. Ces vecteurs acoustiques sont utilisés pour extraire un modèle propre pour chaque locuteur et le sauvegarder dans une base de références. La phase de test est effectuée en faisant une comparaison entre le modèle du locuteur proclamé enregistre dans la base de données et le modèle du signal de parole à l'entrée du système. Si le score calcule est au-dessus d'un certain seuil, l'identité proclamée est vérifiée. Si on utilise un seuil élevé, le système obtient une sureté élevée et empêche l'acceptation des imposteurs, mais en faisant ainsi on risque de rejeter des identités des personnes véritables, et vice versa.

1.1.5 Identification automatique du locuteur

C'est le processus de trouver l'identité d'un locuteur inconnu `à l'aide de sa voix parmi un ensemble de L locuteurs connus par le système. La structure de base d'un système d'identification est montrée dans la figure 1.2.

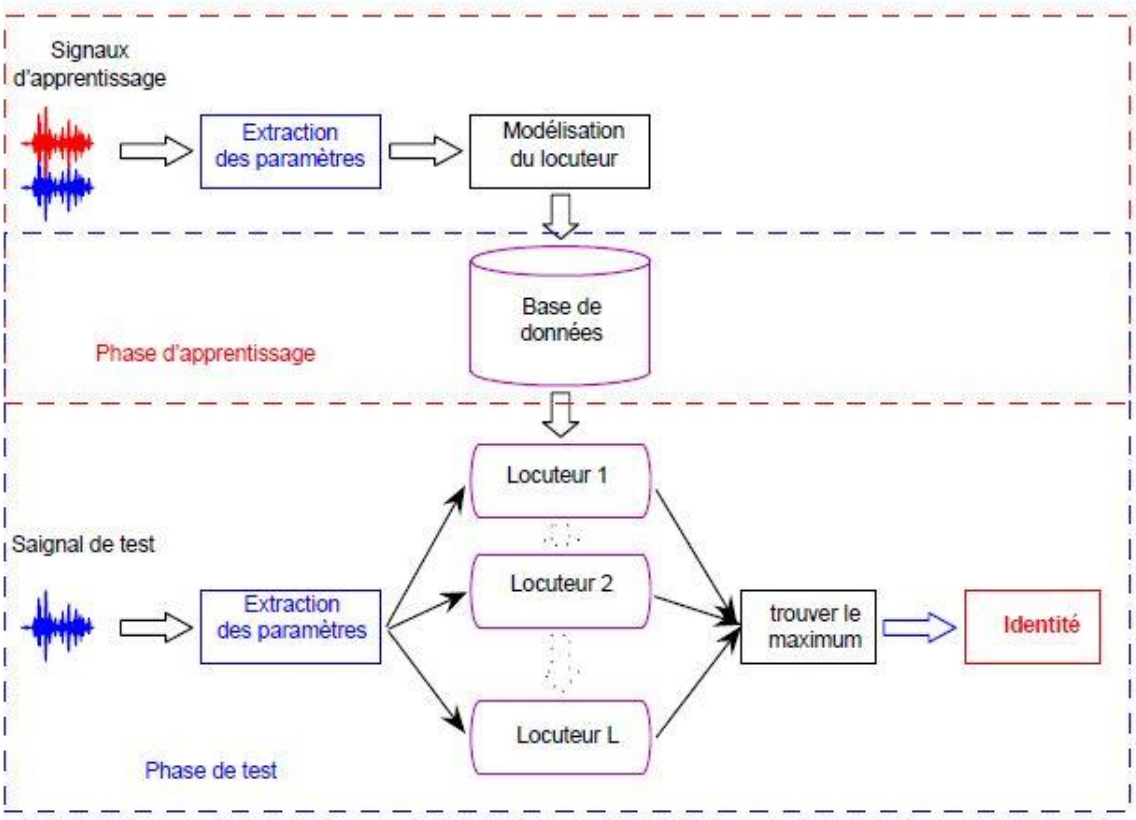

Figure 1.2 Système d'identification du locuteur

Nous notons que l'étape d'apprentissage dans le système d'identification est la même que dans un système de vérification. Dans la phase de reconnaissance, les modèles de L locuteurs sont comparés en parallèle et le plus-probable est rapporté.

On distingue deux types d'identification du locuteur, l'identification en ensemble fermé et l'identification en ensemble ouvert. Dans le premier cas, elle consiste à identifier le locuteur, qui parle, parmi un ensemble(fermé) de L locuteurs. Dans le second cas, la tâche est analogue à l'identification en ensemble fermé, en rajoutant l'hypothèse que le locuteur qui parle peut n'être aucun des L locuteurs. Cette tâche peut être vue comme une étape d'identification en ensemble fermé suivie d'une étape de vérification en utilisant comme identité proclamée, l'identité trouvée lors de l'identification On peut classifier les systèmes d'identification du locuteur selon son indépendance au texte prononcé, en systèmes indépendants du texte et systèmes d´pendants du texte, selon qu'il existe ou non des contraintes sur le contenu linguistique de l'énoncé quel utilisateur doit prononcer pour procéder à la reconnaissance.

Le fonctionnement d'un système de reconnaissance du locuteur se décompose en deux phases : une phase d'apprentissage et une phase de test.

1.1.5.1 Phase d'apprentissage

Un locuteur prononce l'ensemble du vocabulaire, souvent plusieurs fois, pour créer en machine le dictionnaire de références acoustiques. Pour l'approche analytique, l'ordinateur demande à l'utilisateur d'énoncer des phrases souvent dépourvues de toute signification, mais qui présentent l'intérêt de comporter des successions de phonèmes bien particuliers.

1.1.5.2 Phase de test

En phase de test, le fonctionnement diffère l´légèrement selon que l'on considère un système d'identification ou un système de vérification. En identification, le système doit retourner l'identité du locuteur inconnue à partir d'un ´énoncé de test. En vérification, le système retourne une d´décision d'acceptation ou de rejet d'une identité proclamé, en basant sur un énoncé de test.

Mécanisme de Production de la parole

La parole est un mode de communication multi sensoriel. La moitié de l'information disponible dans le son est également visible sur le visage (mouvement des lèvres, ouverture et fermeture de bouche).Nous accompagnons aussi nos mots de gestes (hochement de tête, froncement de sourcils…..). C'est un dispositif complexe.

Le son est toujours le produit de l'action d'une source d'énergie sur un excitateur, modifié par un résonateur. Pour la voix humaine, les poumons font office de source d'énergie, les cordes vocales d'excitateur, et le conduit qui va du pharynx aux lèvres de résonateur.

1.1.6 Production de parole (le son)

La production de la parole fait intervenir différents organes. La source de la parole provient des poumons qui émettent un flux d'air.

Ce flux d'air va traverser le larynx pour faire vibrer ou non les cordes vocales. Il va ensuite traverser le conduit vocal (cavité nasale et buccale) et les articulateurs tels que les lèvres et la langue (Figure 1.3). Cet ensemble agit comme un filtre, considéré comme linéaire, dont la réponse impulsionnelle comporte des fréquences de résonance caractérisées par des pics, appelés formants, dans le spectre du signal de sortie.

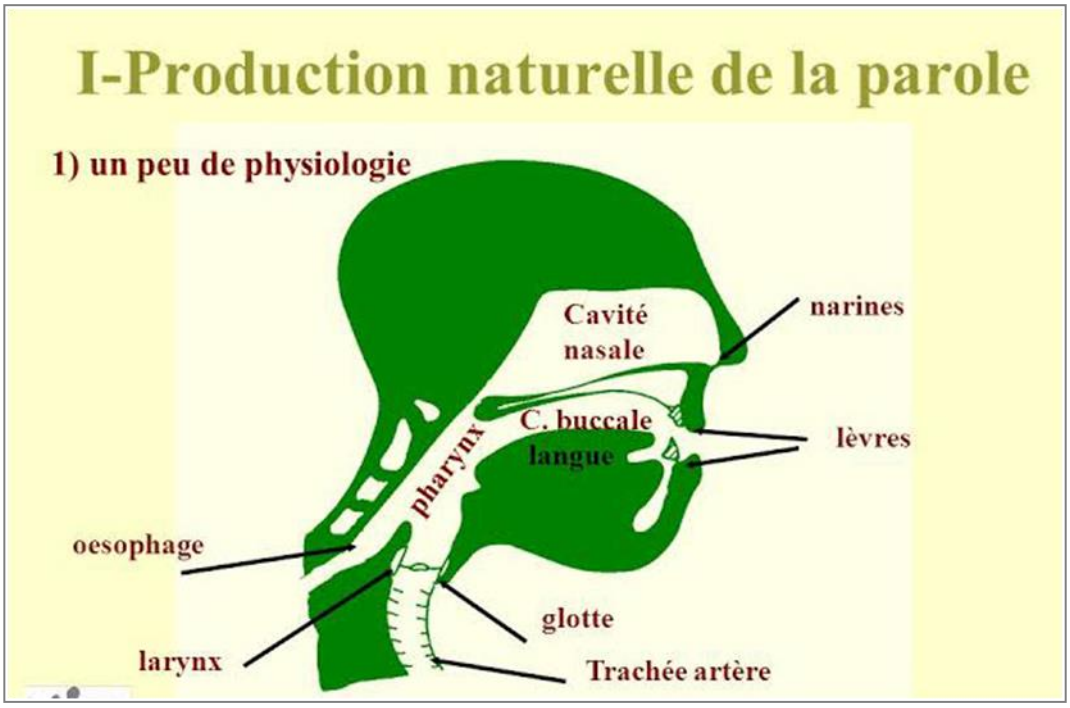

Figure 1.3 Modèle physiologique de la production de la parole

Le signal résultant est globalement non stationnaire mais peut être considéré comme stationnaire sur de très courtes périodes, de l'ordre de $20ms$ (signal pseudo-stationnaire). Sur un segment de parole de cette longueur la voix est habituellement et schématiquement séparée en deux classes distinctes :

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

Dans le premier cas, la source d'excitation est modélisée par un train d'impulsions périodique, de fréquence dite de voisement F_0, qui correspond à la fréquence de vibration des cordes vocales, la fréquence fondamentale ou pitch ; dans le second cas, la source est modélisée par un bruit blanc. Cette représentation binaire de la production de la parole.Elle est reprise sur la figure 1.4.

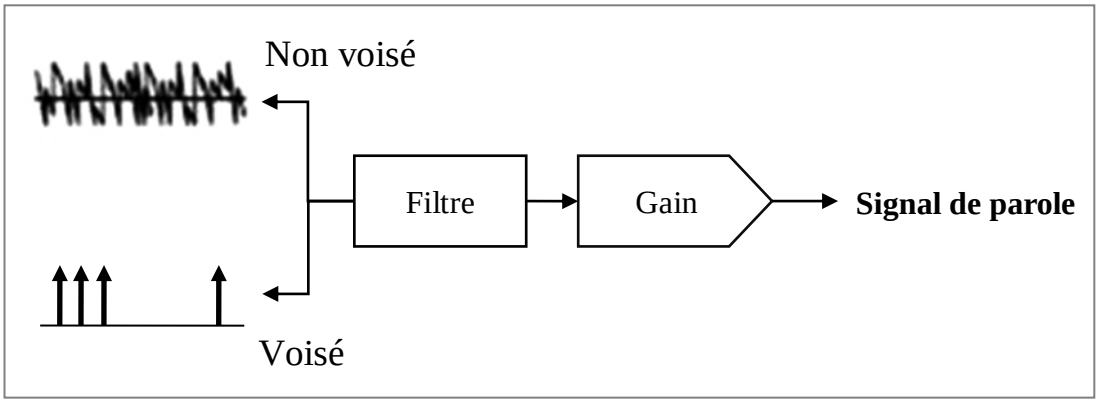

Figure 1.4 Modèle de production de la parole

1.1.7 Variabilités du signal de parole

Le signal vocal de deux prononciations à contenu phonétique égal est distinct pour un même locuteur (variabilité intra-locuteur) ou pour des locuteurs différents (variabilité interlocuteur). Ces deux types de variabilités sont expliquées ci-après [9, 10].

La figure 1.5 montre deux signaux à contenu phonétique égal, prononcé par le même locuteur.

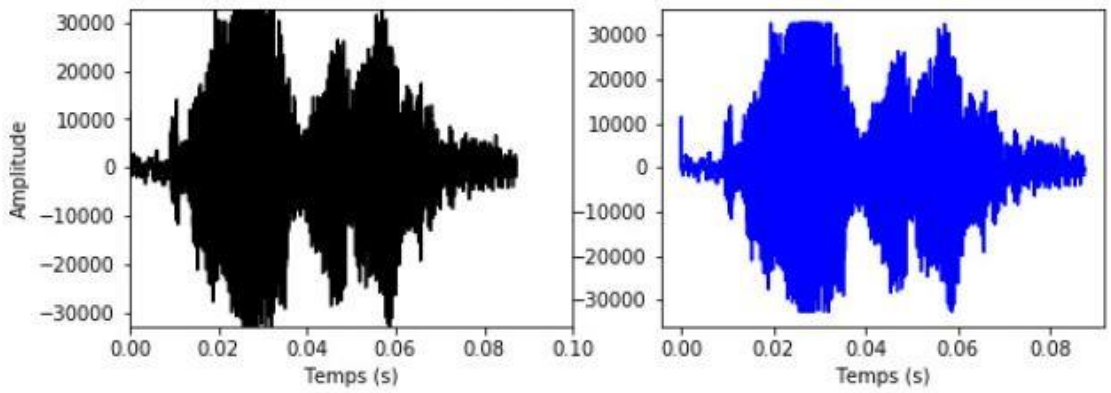

Figure 1.5 Signaux de parole à contenu phonétique égal produit par le même locuteur

1.1.7.1 Variabilité inter-locuteurs

Le signal de parole ne véhicule pas, seulement, un message, il porte aussi des informations sur l'individu qui l'émet. Il varie en fonction du locuteur. Cette variabilité utile pour différencier les locuteurs, est principalement due à des différences fonctionnelles et anatomiques (les fonctions de l'appareil phonatoire et de l'oreille) entre locuteurs (chacun a son propre appareil phonatoire). Autre origine de la variabilité interlocuteurs revient aux différences de prononciation qui existent au sein d'une même langue et qui constituent les accents régionaux.

1.1.7.2 Variabilité intra-locuteur

La variabilité intra-locuteur identifie les différences dans le signal produit par une même personne. Cette variation peut résulter de l'état physique ou moral du locuteur (Rhume par exemple). L'humeur ou l'émotion du locuteur peut également influencer son rythme d'élocution, son intonation.

La variabilité intra-locuteur peut rendre l'identification du locuteur plus difficile, avec la variabilité de parole qui est due aux conditions d'enregistrement du signal de parole (bruit ambiant, microphone utilisé, lignes de transmission).

Application des systèmes IAL

Les applications de IAL (identification automatique du locuteur) sont multiples et peuvent variées selon leurs types. Les évolutions des techniques de l'ASR ont permis aux systèmes d'évoluer et d'être de plus en plus efficaces. Aussi, La plupart des systèmes ASR

sont des systèmes dépendants ou indépendants du locuteur. Les systèmes dépendants du locuteur nécessitent une phase d'apprentissage où de nombreuses heures de parole sont généralement indispensables. Cependant, les systèmes indépendants du locuteur ne nécessitent aucune phase d'apprentissage des données et sont souhaitables pour de nombreuses applications où l'apprentissage est difficile à mener. Cette section décrit brièvement les quatre grands types de systèmes qui existent en reconnaissance de la parole [11]

1.1.8 Commandes vocales

Les systèmes à commandes vocales offrent une interaction entre l'utilisateur et la machine grâce à des commandes vocales, qui s'utilisent généralement dans les systèmes embarqués. Ces commandes représentent des mots isolés que l'utilisateur prononce dans le but d'interagir avec le système. Ce type de système IAL est celui adopté par la présente étude.

1.1.9 Systèmes de compréhension

Principalement, ils permettent de dialoguer avec une machine. Ainsi, l'utilisateur prononce une suite de mots-clés que le système est capable de reconnaître. A la différence des systèmes à commandes vocales, ce type de système utilise en plus un dispositif de compréhension des mots pour les interpréter et répondre en conséquence. Les systèmes de compréhension utilisent un vocabulaire restreint et un mode indépendant du locuteur. L'utilisation de ces systèmes se limite habituellement à l'interrogation d'une base de données et de standards téléphoniques automatisés [12].

1.1.10 Systèmes de dictée automatique

Le rôle de ces systèmes est la transcription d'un texte dicté par un utilisateur de la meilleure manière possible. Toutefois, le texte transcrit doit respecter les règles orthographiques et grammaticales propres à la langue considérée. Ce type de système ne prend pas en charge la compréhension du texte à transcrire et qui engendre des erreurs de transcription. Ces systèmes sont fréquemment utilisés pour transcrire des compte rendu ainsi que des rapports. Dans ces cas, l'utilisateur doit adapter sa locution car il est conscient qu'il s'adresse à un ordinateur. Avec l'objectif d'obtention de meilleures

précisions en temps réel, ces systèmes à grand vocabulaire sont devenus très dépendants de l'utilisateur. En outre, une phase d'apprentissage nécessitant du temps, est vitale pour permettre au système d'apprendre des modèles spécifiques de la voix de son utilisateur [13].

1.1.11 Systèmes de transcription grande vocabulaire

L'objectif de ces systèmes est de transcrire des documents audio non préparés par extraction du maximum d'informations de l'enregistrement. Le signal audio étant un signal riche en informations de différentes natures (informations sur les utilisateurs, les frontières des phrases, les zones de musique ou encore les hésitations des utilisateurs) pouvant enrichir la transcription en mots. Un système de transcription grand vocabulaire se compose de plusieurs modules : un module permettant la transcription en mots du signal, un autre module permettant l'extraction des informations additionnelles disponibles dans le signal audio (tel que le module de la reconnaissance automatique du locuteur). Ces systèmes de transcription traitent de multiples documents de diverse nature pouvant être des enregistrements de réunions, d'émissions de télévision, de journaux radiophoniques et de compte-rendu [14].

Conclusion

Dans ce chapitre nous avons présenté la reconnaissance du locuteur est l'une des tâches pionnières de l'intelligence artificielle consistant à reproduire la capacité d'un être humain à extraire des informations de la parole produite par un autre être humain. Cette tâche, trop complexe pour être reproduite par un système informatique unique, a été subdivisée en plusieurs sous-problèmes en fonction du type d'informations à extraire et à reconnaître. Les problématiques les plus étudiées sont la reconnaissance du locuteur, de son état émotionnel, de la langue employée et du langage parlé. Le problème de la RAL est généralement abordé selon deux approches, analytique et globale. L'approche analytique permet d'aborder le problème de la reconnaissance du locuteur continue, quant à l'approche globale inspirée par les méthodes de reconnaissance des formes, elle privilégie l'aspect information acoustique sur l'aspect linguistique. Cette approche est une manière de contourner les difficultés de l'analyse linguistique. La décision du système s'effectue après évaluation d'un taux de ressemblance entre la forme à reconnaître et une série de formes

préalablement mémorisées. Chacune de ces approches a ses avantages et ses inconvénients, mais les approches analytiques s'avèrent supérieures aux approches globales quant à la qualité des résultats, lorsque le nombre des énoncés potentiels est élevé et/ou lorsque la redondance acoustique est faible. Que La reconnaissance soit globale ou analytique, le processus de reconnaissance du locuteur commence par un prétraitement acoustique du signal vocal dont le but de réduire le flux d'information et d'éliminer les redondances présentées dans celui-ci. Malgré Ces problèmes et difficultés, les systèmes de reconnaissance automatique du locuteur deviennent chaque année de plus en plus performants.

Chapitre 2

Principales Techniques de la Reconnaissance Automatique du Locuteur

Introduction

On désigne habituellement par reconnaissance du locuteur tout processus de décision consistant à utiliser des caractéristiques du signal de parole pour déterminer l'identité du locuteur d'un énoncé particulier. Si en plus le procédé utilisé ne fait pas appel à une intervention humaine, que se soit au niveau des prétraitements, des différentes étapes algorithmiques ou de l'interprétation des résultats, on parle alors de reconnaissance automatique du locuteur. A l'instar des autres thématiques du traitement automatique de la parole, la reconnaissance du locuteur est située à un point de croisement entre plusieurs disciplines, parmi lesquelles on peut citer le traitement du signal, la reconnaissance des formes, les statistiques, les probabilités, la théorie de la décision, mais aussi la phonétique et la linguistique. La reconnaissance du locuteur, qui regroupe à la fois : l'identification du locuteur et sa vérification, est le processus d'identifier automatiquement qui parle utilisant l'information individuelle incluse dans la voix de la personne. Cette technique permet d'employer la voix du locuteur pour vérifier son identité et contrôler l'accès aux services tels que les achats par téléphone, les services d'accès aux bases de données, les services d'information, la messagerie, la commande de sécurité pour des informations confidentielles et l'accès aux ordinateurs à distance, en plus des applications judiciaires.

Reconnaissance Automatique du Locuteur

Le but de la reconnaissance du locuteur est de reconnaître l'identité d'une personne l'aide de sa voix. Les applications de la RAL sont liées principalement aux problèmes d'authentification ou de confidentialité.

Les différentes tâches de la reconnaissance du locuteur sont regroupées en trois catégories principales suivantes :

- l'identification de locuteur (IL)
- la vérification du locuteur (IV)
- l'indexation des locuteurs ou le suivi de locuteurs [15].

2.1.1 Architecture du système

On peut supposer qu'une personne veuille exprimer une certaine pensée à une autre personne ou à une machine. Elle doit composer une phrase significative sous une forme d'ordre de mots pour exprimer cette pensée .Quand les mots sont choisis ,la personne envoie les signaux de commande appropriés aux organes de production de la parole qui forment une expression de la parole et prononce la phrase désirée. Cette phrase est représentée par un signal acoustique.

De point de vue fonctionnement, un système de RAL se décompose en deux processus différents : le premier processus analyse le son articulé et le transforme en une suite de vecteurs acoustiques, qui caractérisent le signal de parole. L'autre cherche la correspondance en termes d'identité de ces vecteurs acoustiques par rapport aux modèles construits durant la phase d'apprentissage. Dans ce projet nous nous intéressons à implémenter un système d'identification du locuteur en utilisant les réseaux de neurones profonds tels que : les réseaux de neurones convolutifs CNN [16].

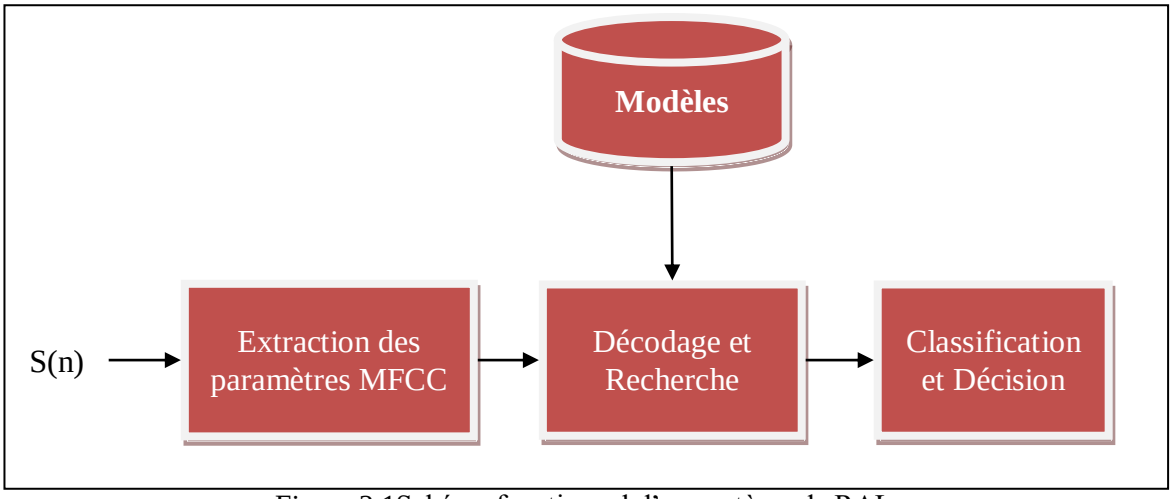

Figure 2.1Schéma fonctionnel d'un système de RAL

Le signal acoustique d'entrée, est transformé en une suite de vecteurs de coefficients acoustiques. Ces vecteurs sont généralement donnés par une représentation cepstral et sont calculés sur chaque trame. En particulier, les MFCC (Mel Frequency Cepstral Coefficients) et les PLP (Perceptuel Linear Prediction) sont souvent utilisés pour représenter les caractéristiques spectrales à court terme. Le bloc de classification décode les vecteurs acoustiques dans une représentation symbolique, selon le sens de maximum de vraisemblance pour estimer le modèle qui produit l'ordre des vecteurs acoustiques

d'entrée. Le bloc final de ce système est un processus de vérification et de décision employé pour mesurer la confiance pour chaque mot identifié. Chacune de ces opérations implique beaucoup de détails et de calcul numérique étendu.

Généralement, les différentes étapes pour construire un système de RAP sont les suivantes :

1. Choisir l'ensemble des paramètres acoustiques et les traitements associés qui représentent le mieux les propriétés du signal acoustique
2. Choisir la tâche de reconnaissance (identité)
3. Apprentissage de l'ensemble des paramètres acoustiques et les modèles des locuteurs.
4. Calculer et évaluer les performances du système résultant de la RAL.

2.1.2 Représentation du signal vocal

A partir d'une opération complexe rassemblant tous les organes de l'appareil phonatoire on obtient le signal de parole. L'excitation de la cavité orale ou nasale par une source acoustique produit une onde acoustique qui véhicule le signal de parole. Ce signal peut être considéré comme une concaténation des réalisations acoustiques, qui sont produites par des actions et des mouvements de l'appareil phonatoire. Chaque réalisation élémentaire, dans le signal de parole, est vue comme un phonème.

La forme du conduit vocale est propre pour chaque locuteur, plusieurs locuteurs d'âge, de sexe et de morphologie différents, d'où pour un même phonème on peut trouver beaucoup de réalisations acoustiques.

Généralement, le signal de parole est considéré comme une association de plusieurs entités élémentaires stationnaires. Cependant, un seul mot peut être prononcé de différentes façons. Les origines de ces variabilités peuvent être vues comme des variabilités interlocuteur et intra-locuteur. En outre, la transmission du signal acoustique, l'aire, le microphone et le câblage influent également sur le signal de parole. Toutes ces variabilités et difficultés rendent la tâche de reconnaître le mot désiré très complexe.

2.1.3 Prétraitement des signaux de parole

La figure2.2 montre un schéma général du processus de traitement appliqué. Le signal est prétraité et segmenté dans une suite de trames de25 à $32\ ms$. Généralement, ces trames sont recouvertes entre eux. Si T_s est le décalage entre deux trames consécutives, alors $1/T_s$ est la fréquence des trames (en Hertz). Le signal de parole peut être considéré quasi stationnaire. Afin d'obtenir le spectre à court terme de chaque trame, les paramètres spectraux obtenus sont habituellement transformés pour fournir aux trames une représentation faiblement corrélé et de dimensions réduites. Chaque trame est représentée par un vecteur x de coefficients acoustiques contenant les paramètres d'analyse. Le vecteur de paramètres acoustiques est habituellement augmenté en ajoutant d'autres termes tels que l'énergie ou les coefficients dynamiques.

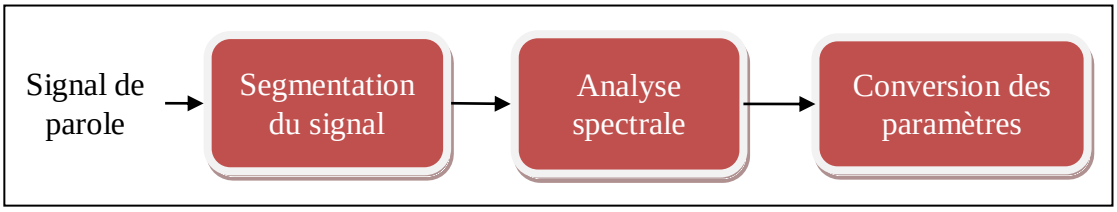

Figure 2.2 Schéma générale de l'étape d'analyse d'un signal de parole

On obtient un signal de parole par l'application de la convolution entre la source et le conduit vocal. Aussi pour inverser l'opération on doit appliquer la déconvolution sur ce signal dans un espace où la convolution devient une multiplication.

Pour relever les hautes fréquences qui ont une énergie plus faible par rapport à celle des basses fréquences, on fait passer le signal de parole (signal pré-accentué) par un filtre numérique dont sa réponse impulsionnelle de premier ordre est donné par:

$$h(z) = 1 - az^{-1} \tag{01}$$

Le signal pré-accentué est lié au signal d'origine x par la formule suivante :

$$x_p(t) = x(t) - ax(t-1) \tag{02}$$

Avec : $0.9 < a < 1$

A ce niveau, deux principales méthodes existent pour l'analyse spectrale dans les systèmes de RAP : le banc de filtres et le codage prédictif (LPC). Cette dernière a été classiquement utilisée pour plusieurs raisons : elle est basée sur une méthode puissante de

production de la parole, ce modèle est approprié aux sons voisés et en plus acceptable pour les sons non voisés. En outre, pour une parole de bonne qualité, la prédiction linéaire fournit de bons résultats que les méthodes basées sur le banc de filtres. Cependant, ces derniers sont l'outil principal d'analyse lors de ces dernières années puisqu'ils montrent de meilleurs résultats en présence du bruit [17].

2.1.4 Extraction des paramètres MFCC

L'extraction des paramètres d'un signal de parole est faite par transformer ce dernier en suite des vecteurs acoustiques. Cette forme est beaucoup plus adéquate à la modélisation statistique et vectorielle. La reconnaissance du locuteur s'applique sur des signaux de parole représentés sous la forme spectrale. Pour cette raison on peut considérer le signal pré-accentué de parole comme un signale quasi-stationnaire après l'analyser par une fenêtre comme la fenêtre de Hamming avec une courte durée de l'ordre vaut 25ms :

$$h(n) = \begin{cases} 0.54 - 0.46\cos 2\pi n & 0 < n < N-1 \\ 0 & ailleurs \end{cases} \quad (03)$$

Pour faire la représentation spectrale du signal de parole, on applique la transformée de Fourier sur chaque trame du signal obtenu après le fenêtrage, en général on utilise des algorithmes tel que FFT (Fast Fourier Transform).[18].

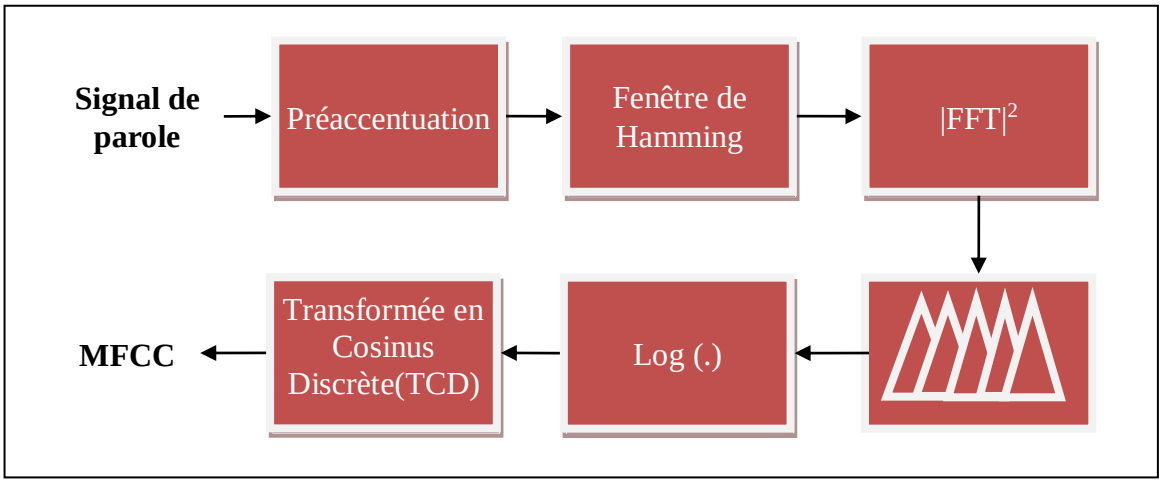

Figure 2.3 Extraction des paramètres MFCC

Le spectre obtenu contient plusieurs fluctuations, mais on s'intéresse qu'à l'enveloppe du spectre. Aussi pour réduire la taille des vecteurs spectraux, le spectre du signal doit être lissé, pour éliminer les fluctuations et le rendre lisse il faut le multiplier par un banc de filtres (série de filtres à bande passante équidistante dans l'échelle Mel). Ce banc de filtres

est défini selon la forme de chaque filtre qui le compose, et la localisation de ses fréquences (centre, a droit ou à gauche), ces filtres sont souvent sous forme triangulaire . La localisation des fréquences centrales des filtres est donnée par :

$$f_{mel} = 2595\ log_{10}\left(1 + \frac{f}{700}\right) \quad (04)$$

Avec :f la fréquence en Hz

A la fin, on calcule l'enveloppe spectrale en DB a partir du logarithme de cette enveloppe, puis on applique une transformée en cosinus discrète pour obtenir les coefficients cepstraux selon des logarithmes des énergies issues du banc de filtres. Le calcul mathématique des coefficients est définit par l'expression suivante :

$$C_n = \sum_{j-1}^{K} S_j cos\left[(j - 0.5)\frac{n.\pi}{K}\right] \quad (05)$$

Avec :

- $i = 1,2, \ldots . L$;
- K : nombre de coefficients spectraux calculés précédemment ($k = 23$) ;
- S_j : coefficients spectraux ;
- L : nombre de coefficients cepstraux que nous voulons calculer ($\mathrm{L} = 12$).

A ce niveau là, l'évolution de temps n'apparait pas dans les paramètres MFCC.
L'information dynamique dans le signal de parole n'est pas la même pour chaque locuteur. On obtient cette information par les dérivées cepstraux.

- Les premières dérivées sont les coefficients Δ qui représentent la vitesse de variation des vecteurs dans le temps
- Les dérivées deuxièmes sont les coefficients ΔΔ qui concernent l'accélération de la parole

Ces coefficients sont exprimés par :

$$\Delta C_m = \frac{\sum_{p=-1}^{L} p^2 C_{m+p}}{\sum_{p=-1}^{L} |p|} \quad (06)$$

$$\Delta\Delta C_m = \frac{\sum_{p=-1}^{L} p^2 C_{m+p}}{\sum_{p=-1}^{L} |p|^2} \quad (07)$$

En général, le nombre de coefficients est pris égal à 13, et parfois réduit à 12, en considérant deux points essentiels :

1. le premier coefficient représentant l'énergie de la trame et ne pouvant réellement contribuer à la reconnaissance.
2. les 12 coefficients représentant l'enveloppe cepstral plus ou moins lissée, avec une suppression des hautes variations fréquentielles.

2.1.5 Phase d'apprentissage du système

L'apprentissage du système est un sous-ensemble de l'apprentissage automatique, qui consiste essentiellement en un réseau neuronal à trois couches ou plus. Ces réseaux neuronaux tentent de simuler le comportement du cerveau humain, même s'ils sont loin d'en avoir la capacité, ce qui leur permet "d'apprendre" à partir de grandes quantités de données. Si un réseau neuronal à une seule couche peut toujours faire des prédictions approximatives, des couches cachées supplémentaires peuvent aider à optimiser et à affiner la précision.

L'apprentissage du système est à la base de nombreuses applications et services d'intelligence artificielle (IA) qui améliorent l'automatisation, en exécutant des tâches analytiques et physiques sans intervention humaine. La technologie d'apprentissage du système est à l'origine de produits et de services de tous les jours (comme les assistants numériques, les télécommandes de télévision à commande vocale et la détection des fraudes à la carte de crédit) ainsi que de technologies émergentes (comme les voitures à conduite autonome).

2.1.5.1 Apprentissage du système et apprentissage automatique

Les algorithmes d'apprentissage automatique exploitent des données structurées et étiquetées pour faire des prédictions, ce qui signifie que des caractéristiques spécifiques sont définies à partir des données d'entrée du modèle et organisées en tableaux. Cela ne signifie pas nécessairement qu'ils n'utilisent pas de données non structurées ; cela signifie

simplement que si c'est le cas, elles subissent généralement un prétraitement pour les organiser dans un format structuré.

L'apprentissage du système élimine une partie du prétraitement des données qui est généralement impliqué dans l'apprentissage automatique. Ces algorithmes peuvent ingérer et traiter des données non structurées, comme du texte et des images, et ils automatisent l'extraction de caractéristiques, ce qui élimine une partie de la dépendance à l'égard des experts humains. Par exemple, disons que nous avons un ensemble de photos de différents animaux de compagnie et que nous voulons les classer par "chat", "chien", "hamster", etc. Les algorithmes d'apprentissage du système peuvent déterminer quelles caractéristiques (par exemple, les oreilles) sont les plus importantes pour distinguer chaque animal d'un autre. En apprentissage automatique, cette hiérarchie de caractéristiques est établie manuellement par un expert humain.

Ensuite, grâce aux processus de descente de gradient et de rétro propagation, l'algorithme d'apprentissage du système s'ajuste et s'adapte à la précision, ce qui lui permet de faire des prédictions sur une nouvelle photo d'un animal avec une précision accrue.

Les modèles d'apprentissage automatique et d'apprentissage du système sont également capables de différents types d'apprentissage, qui sont généralement classés en apprentissage supervisé, apprentissage non supervisé et apprentissage par renforcement. L'apprentissage supervisé utilise des ensembles de données étiquetées pour catégoriser ou faire des prédictions, ce qui nécessite une certaine forme d'intervention humaine pour étiqueter correctement les données d'entrée. En revanche, l'apprentissage non supervisé ne nécessite pas d'ensembles de données étiquetées. Il détecte plutôt des modèles dans les données, les regroupant en fonction de leurs caractéristiques distinctives. L'apprentissage par renforcement est un processus dans lequel un modèle apprend à devenir plus précis dans l'exécution d'une action dans un environnement, sur la base d'un retour d'information, afin de maximiser la récompense.

2.1.5.2 Fonctionnement d'apprentissage du système

Les réseaux neuronaux d'apprentissage du système, ou réseaux neuronaux artificiels, tentent d'imiter le cerveau humain par une combinaison d'entrées de données, de pondérations et de biais. Ces éléments fonctionnent ensemble pour reconnaître, classer et décrire avec précision les objets présents dans les données.

Les réseaux neuronaux profonds se composent de plusieurs couches de nœuds interconnectés, chacune s'appuyant sur la couche précédente pour affiner et optimiser la prédiction ou la catégorisation. Cette progression des calculs dans le réseau est appelée propagation vers l'avant. Les couches d'entrée et de sortie d'un réseau neuronal profond sont appelées couches visibles. La couche d'entrée est celle où le modèle d'apprentissage du système ingère les données à traiter, et la couche de sortie est celle où la prédiction ou la classification finale est effectuée.

Un autre processus appelé rétro-propagation utilise des algorithmes, comme la descente de gradient, pour calculer les erreurs dans les prédictions, puis ajuste les poids et les biais de la fonction en remontant les couches dans le but d'entraîner le modèle. Ensemble, la propagation avant et la rétro-propagation permettent à un réseau neuronal de faire des prédictions et de corriger les erreurs en conséquence. Au fil du temps, l'algorithme devient progressivement plus précis.

Ce qui précède décrit le type le plus simple de réseau neuronal profond dans les termes les plus simples. Cependant, les algorithmes d'apprentissage profond sont incroyablement complexes, et il existe différents types de réseaux neuronaux pour répondre à des problèmes ou des ensembles de données spécifiques. Par exemple,

- Les réseaux neuronaux à résolution (CNN), utilisés principalement dans les applications de vision par ordinateur et de classification d'images, peuvent détecter des caractéristiques et des motifs dans une image, ce qui permet d'effectuer des tâches telles que la détection ou la reconnaissance d'objets. En 2015, un réseau neuronal convolutif a battu un humain dans un concours de reconnaissance d'objets pour la première fois.
- Les réseaux neuronaux récurrents (RNN) sont généralement utilisés dans les applications de reconnaissance du langage naturel et de la parole, car ils exploitent des données séquentielles ou des séries chronologiques

2.1.6 Phase de reconnaissance du locuteur

La reconnaissance du locuteur peut aider à déterminer qui parle dans un clip audio. Le service peut vérifier et identifier les locuteurs par les caractéristiques uniques de leur voix, en utilisant la biométrie vocale.

Vous fournissez des données d'entraînement audio pour un seul locuteur, ce qui crée un profil d'inscription basé sur les caractéristiques uniques de la voix du locuteur. Vous pouvez ensuite comparer des échantillons de voix audio avec ce profil pour vérifier que le locuteur est la même personne (vérification du locuteur). Vous pouvez également comparer des échantillons de voix audio à un groupe de profils de locuteurs inscrits pour voir s'ils correspondent à n'importe quel profil du groupe (identification du locuteur).

2.1.6.1 Vérification du locuteur

La vérification du locuteur rationalise le processus de vérification de l'identité d'un locuteur inscrit à l'aide de phrases de passe ou d'une saisie vocale libre. Par exemple, vous pouvez l'utiliser pour la vérification de l'identité des clients dans les centres d'appels ou pour l'accès aux installations sans contact.

2.1.6.2 Fonctionnement de la vérification du locuteur

La vérification du locuteur peut être dépendante ou indépendante du texte. La vérification en fonction du texte signifie que les locuteurs doivent choisir la même phrase de passe à utiliser pendant les phases d'inscription et de vérification. La vérification indépendante du texte signifie que les locuteurs peuvent s'exprimer en langage courant dans les phrases d'inscription et de vérification.

Pour la vérification en fonction du texte, la voix du locuteur est enregistrée en prononçant une phrase de passe parmi un ensemble de phrases prédéfinies. Les caractéristiques de la voix sont extraites de l'enregistrement audio pour former une signature vocale unique, et la phrase de passe choisie est également reconnue. Ensemble, la signature vocale et la phrase de passe sont utilisées pour vérifier le locuteur.

La vérification indépendante du texte n'a aucune restriction sur ce que le locuteur dit pendant l'inscription, à part la phrase d'activation initiale pour activer l'inscription. Ellen'a aucune restriction sur l'échantillon audio à vérifier, car elle extrait uniquement les caractéristiques de la voix pour évaluer la similarité.

Les API ne sont pas destinées à déterminer si l'audio provient d'une personne vivante ou d'une imitation ou d'un enregistrement d'un locuteur inscrit.

2.1.6.3 Identification du locuteur

L'identification du locuteur vous aide à déterminer l'identité d'un locuteur inconnu au sein d'un groupe de locuteurs inscrits. L'identification du locuteur vous permet d'attribuer la parole à des locuteurs individuels et de tirer profit de scénarios avec plusieurs locuteurs, par exemple :

- La prise en charge de solutions pour la productivité des réunions à distance.
- La personnalisation de dispositifs multiutilisateurs.

2.1.6.4 Fonctionnement l'identification du locuteur

L'inscription pour l'identification du locuteur est indépendante du texte. Il n'y a pas de restrictions sur ce que le locuteur dit dans l'audio, à part la phrase d'activation initiale pour activer l'inscription. Comme pour la vérification du locuteur, la voix du locuteur est enregistrée lors de la phase d'inscription, et les caractéristiques de la voix sont extraites pour former une signature vocale unique. Dans la phase d'identification, l'échantillon de voix d'entrée est comparé à une liste spécifique de voix enregistrées (jusqu'à 50 dans chaque demande).

Réseaux de neurones profonds

L'apprentissage profond ou le deep learning est une sorte de l'intelligence artificielle extrait de la technologie du machine learning c'est quand la machine à la capacité d'apprendre par elle-même, sans la programmer ou mettre des règles précises pour s'exécuter.

Le deep learning est basé sur un principe de fonctionnement imitant le cerveau humain, où il utilise des réseaux neuronaux artificiels qui se composent des dizaines et des centaines de couches neurones, chaque couche est formée des informations reçues de la couche précédente. Comme exemple dans un système dédié à la lecture le système apprend les lettres puis les mots puis les phrases…

L'installation des modules se fait avec les réseaux de neurones profonds, cette technologie connais trois modèles qui sont implémentés et utilisés. Ce sont les réseaux neuronaux convolutifs CNN, Sinclayer et le ConstantSinc

2.1.7 Réseaux de neurones convolutifs CNN

Les réseaux convolutifs ou en anglais Convolutional Neural Network est une des classes de réseau neuronal convolutif (CNN). Ces derniers ont joué un grand rôle dans l'évolution du domaine de la reconnaissance du locuteur. On les trouve souvent pour analyser la parole

Les réseaux de neurones convolutifs sont les modèles les plus utilisés dans la classification des locuteurs. Chaque réseau se compose de 4 types de couches :

- La couche de convolution (CONY) qui traite les données d'un champ récepteur.
- La couche de pooling (POOL), qui permet de compresser l'information en réduisant la taille de l'image intermédiaire (souvent par sous-échantillonnage).
- La couche de correction (ReLU), souvent appelée par abus 'ReLU' en référence à la fonction d'activation (Unité de rectification linéaire).
- La couche "entièrement connectée" (FC), qui est une couche de type perceptron.

2.1.8 Réseau Sinc

La forme d'onde de chaque phrase vocale a été divisée en morceaux de 200 ms (avec un chevauchement de 10 ms), qui ont été introduits dans l'architecture SincNet. La première couche effectue des convolutions sincères comme décrit dans la Sec. 2, en utilisant 80 filtres de longueur $L = 251$. L'architecture utilise ensuite deux couches convolutionnelles standard, chacun utilise 60 filtres de longueur 5. Toutes les couches convolutionnelles (y compris la couche d'entrée SincNet).

Ensuite, trois couches entièrement connectées composées de 2048 neurones et normalisées avec la normalisation par lots ont été appliquées. Toutes les couches cachées utilisent des non-linéarités de type leaky-ReLU.

Les paramètres de la couche sinc ont été initialisés en utilisant des fréquences de coupure à l'échelle de Mel. Ttandis que le reste du réseau a été initialisé avec le schéma d'initialisation "Glorot" bien connu.

La classification du locuteur au niveau de l'image a été obtenue en appliquant un classificateur softmax, qui fournit un ensemble de probabilités postérieures sur les locuteurs ciblés. Une classification au niveau de la phrase a été simplement dérivée en

faisant la moyenne des prédictions des cadres et en votant pour le locuteur qui maximise la moyenne postérieure.

2.1.9 Constant Sinc Layer

SincNet pour la reconnaissance du locuteur où la première couche est constituée de filtres sinc. La couche Sinc apprend des filtres passe-bande compacts adaptés à la modélisation du locuteur. Elle est paramétrée par les fréquences de coupure de ces filtres passe-bande. Le gain des filtres sinc est appris par les couches ultérieures (convolutionnelles et entièrement connectées) de l'architecture SincNet. SincNet a été développé pour la reconnaissance du locuteur dans un scénario pratique où peu de données d'entraînement étaient disponibles alors que les énoncés de test étaient très courts.

SincNet a été développé comme une architecture efficiente pour le traitement de la forme d'onde de la parole brute pour la reconnaissance du locuteur. Montre l'architecture SincNet qui consiste en six couches cachées, à savoir la couche Sinc, deux couches convolutives 1D et trois couches entièrement connectées. La couche Sinc effectue des convolutions basées sur le Sinc sur des trames superposées du signal du domaine temporel

Conclusion

Ce chapitre a tenté d'expliquer le processus de la reconnaissance automatique du locuteur, commençant par le rôle de la parole dans la communication puis on a cité les applications de la reconnaissance du locuteur et le principe de fonctionnement des systèmes RAL. A la fin on a expliqué les types des réseaux neuronaux convolutifs et leur rôle dans l'apprentissage profond du système d'identification du locuteur tel que les CNN, le Sinc, et ConsatSINC.

Chapitre 3

Identification du locuteur par deep learning

Introduction

Le deep learning ou apprentissage profond est un sous-domaine de l'intelligence artificielle (IA). qui utilise les réseaux de neurones pour analyser différents facteurs avec une structure similaire au système neural humain. Le deep learning est d'une grande utilité dans l'univers des technologies de l'information et de la communication.

Dans ce chapitre on va réaliser un système d'identification du locuteur où on va installer une base de données contient 20 locuteurs et on va utiliser plusieurs méthodes d'apprentissage profond pour entrainer le système à identifier un locuteur à travers sa voix à l'aide du logiciel Matlab 2021 qui incorpore un toolbox de deep learning.

Contexte expérimentale

3.1.1 Logiciel utilisé

Les différents modules sont implémentés à l'aide du logiciel Matlab. C'est un logiciel commercial de calcul interactif. Il permet de réaliser des simulations numériques basées sur des algorithmes d'analyse numérique. Il peut donc être utilisé pour la résolution approchée d'équations différentielles, d'équations aux dérivées partielles ou de systèmes linéaires, etc

Matlab est un langage simple et très efficace, optimisé pour le traitement des matrices, d'où son nom. Pour le calcul numérique, Matlab est beaucoup plus concis que les "vieux" langages(C, Pascal, Fortran, Basic). Il est enrichi avec des « toolbox » qui sont des ensembles de fonctions supplémentaires, profilées pour des applications particulières (traitement de signaux, analyses statistiques, optimisation, etc.)

Nous avons utilisé la version 2021 qui incorpore le toolbox, deep learning. Ce toolbox fonctionne avec le toolbox audiobox. Ce dernier offert les différents modules pour la lecture du signal de parole, sa représentation et l'extraction des paramètres acoustique tels que les MFCC. La figure montre la fenêtre principale du logiciel avec une partie du programme implémenté.

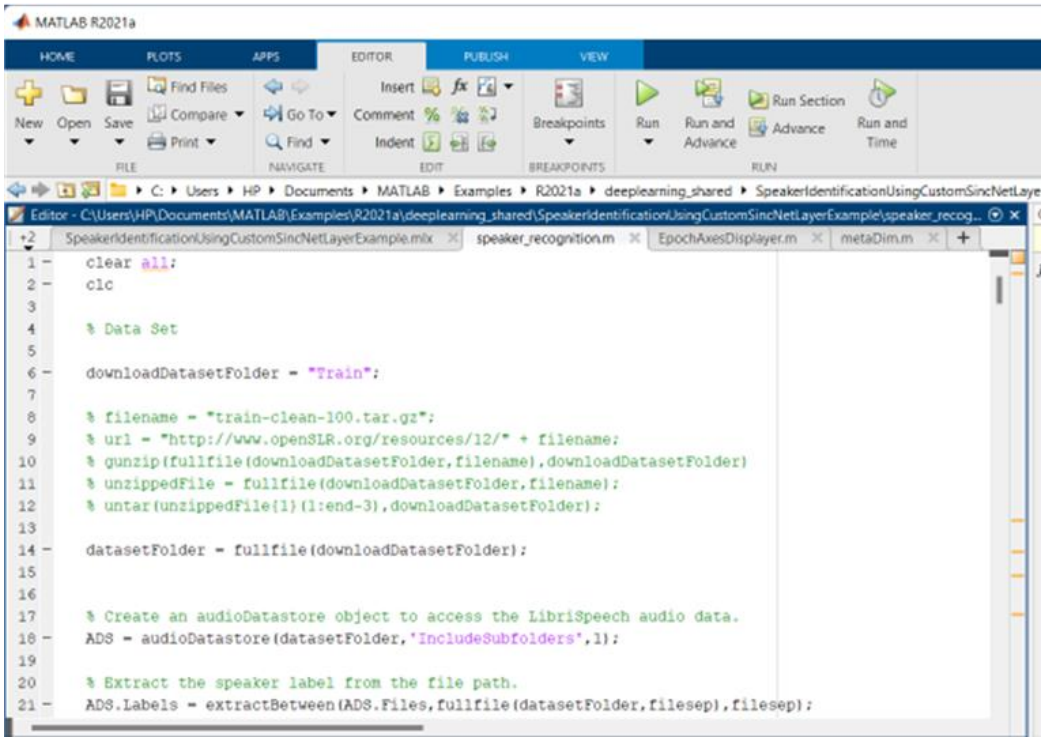

Figure 3.1 Fenêtre principale du Matlab 2021

3.1.2 Corpus de parole

Le corpus LibriSpeech est dérivé de livres audio qui font partie du projet LibriVox, et contient 1000 heures de parole échantillonnée à 16 kHz. Ce corpus est mis à une disposition gratuite pour le téléchargement. Il offert en plus des données d'entraînement de modèles de langage préparées séparément et des modèles de langage pré-construits. Il est remarqué que les modèles acoustiques formés sur LibriSpeech donnent un taux d'erreur plus faible sur les ensembles de test du Wall Street Journal (WSJ) que les modèles formés sur le WSJ lui-même.

3.1.3 Extraction des paramètres acoustiques

Cette étape permet d'extraire l'information pertinente relative à la tâche de classification donnée, à partir d'un signal fortement redondant. La figure 3.2 montre la représentation d'un signal de parole dans le domaine temporelle.

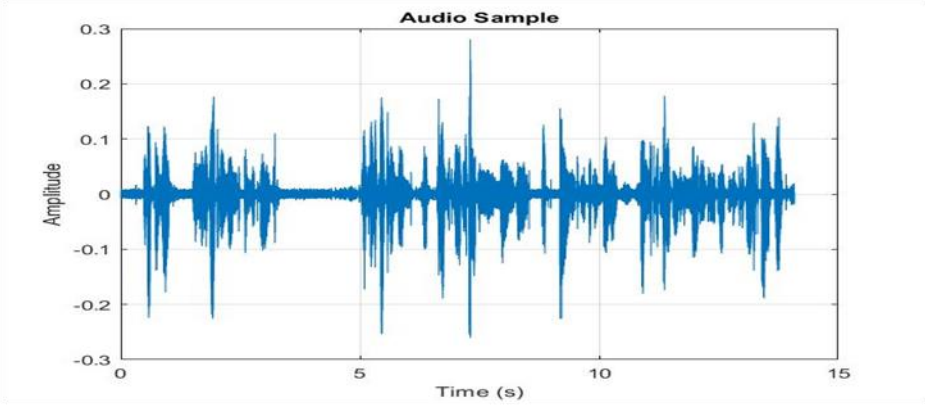

Figure 3.2Signal de parole dans le domaine temporelle

Dans ces expériences, nous avons utilisé les paramètres MFCC (Mel Frequency Cepstrum Coefficients) qui séparent l'effet de l'excitation de celle du conduit vocal à partir d'un modèle d'audition. Les différentes étapes pour calculer ces paramètres peuvent être résumées comme suit :

- Le signal de parole est fenêtré, en premier lieu, par une fenêtre glissante de durée $32ms$ avec un recouvrement de 50% ;
- Calcule de l'énergie de chaque trame ;
- Un passage dans l'échelle Mel, utilisant un banc de filtre en tenant compte de la réponse acoustique de l'oreille humaine.
- Revenir au domaine temporel par l'application de la transformation en cosinus discrète.

Les expériences sont faites en utilisant 20 locuteurs, (10 masculins et 10 féminins), du corpus Librspeech, employant 42 coefficients pour chaque trame (13 MFCC, 1 E) avec ses premiers et seconds dérivés (14 Δ et 14 $\Delta\Delta$). Cela pour voir l'effet de l'information dynamique incluse dans les paramètres acoustiques.

Après nous avons construit un vecteur de 336 paramètres en concaténant 8 trames successives. Ce vecteur est utilisé comme entrée dans les réseaux de neurones implémentés.

3.1.4 Evaluation des performances

L'évaluation des performances de l'identification du locuteur est faite de la manière suivante : le signal de parole à tester est, en premier lieu, passé par le module de l'analyse

acoustique pour le transformer ensuit en vecteurs acoustiques $\{x_1, x_2, \dots, x_t\}$. Si l'identité trouvée et l'identité réelle sont les mêmes, le segment est correctement identifié. L'évaluation des performances finale est donnée par le pourcentage entre le nombre des segments correctement identifié et le nombre total des segments testés.

$$Accuracy = \frac{\#\ correctementsidentifiés}{\#\ total} \tag{08}$$

Identification du locuteur par deep learning

Comme mentionné précédemment, nous avons implémenté trois modèles de deep learning pour l'identification du locuteur, ces trois modèles sont CNN_IAL, SINC_IAL et ConstantSincLayer_IAL. On a appliqué ces modèles sur 15 époques pour identifier le locuteur où chaque modèle a donné une précision d'indentification différente.

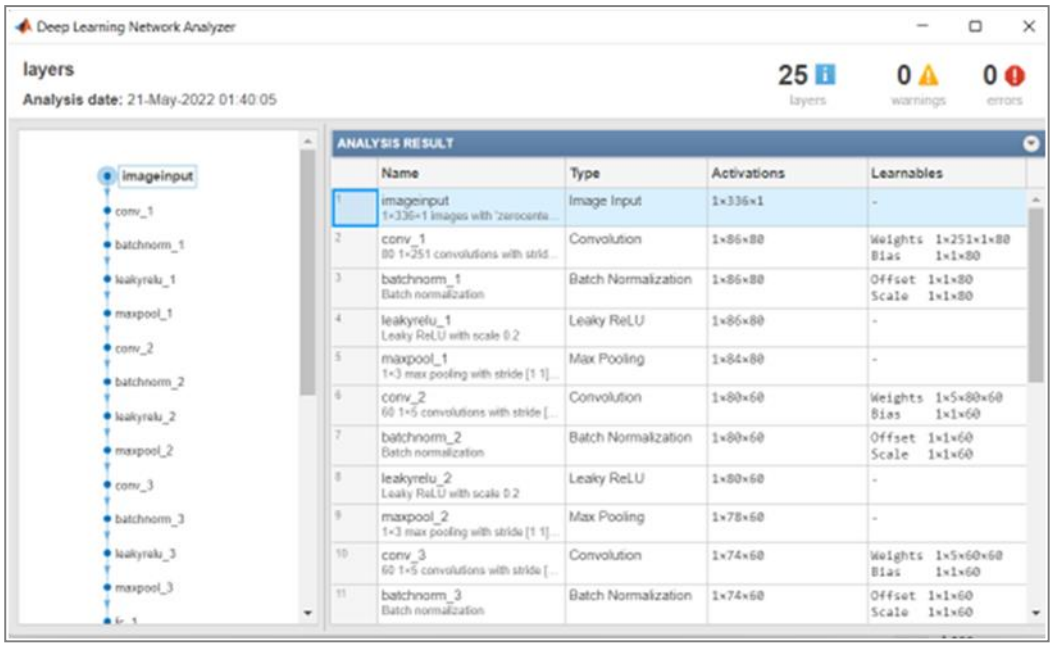

Figure 3.3 Architecture de CNN

3.1.5 Implémentation du système CNN_IAL

Cette étape représente l'apprentissage par CNN (Convolutional Neural Network) dont il est basé sur l'apprentissage d'un nombre de filtre par des signaux de paroles extraits des différents locuteurs (20 locuteurs) sur 15 époques afin qu'il puisse identifier le locuteur qui a parlé.

Ce modèle a donné une précision d'indentification de 54.83% avec un taux de perte de 1.5 (figure3.4).

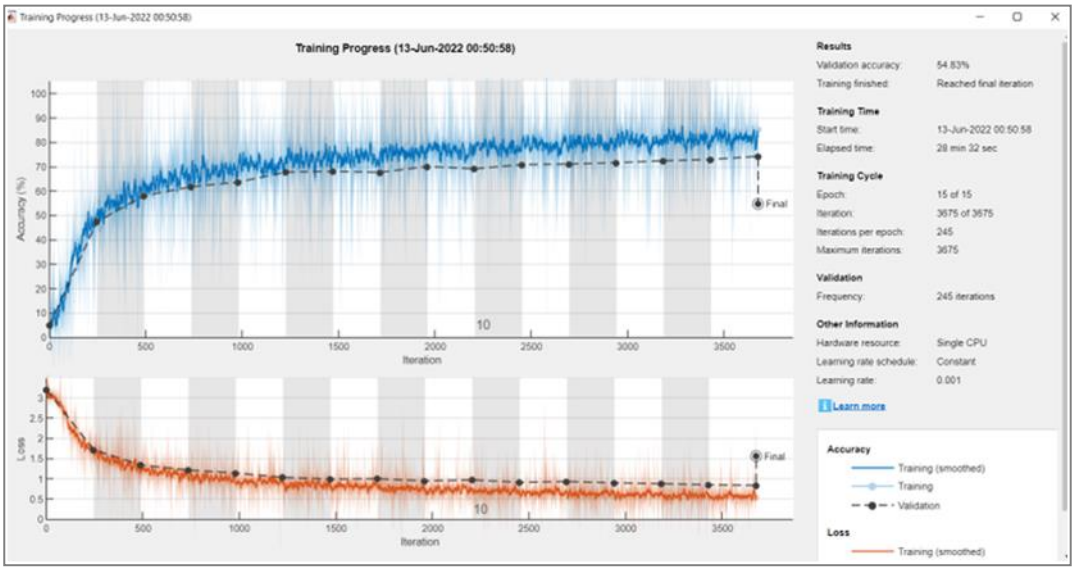

Figure 3.4 Apprentissage des systèmes d'identification du locuteur par CNN

La figure 3.5 montre les réponses fréquentielles des filtres utilisées.

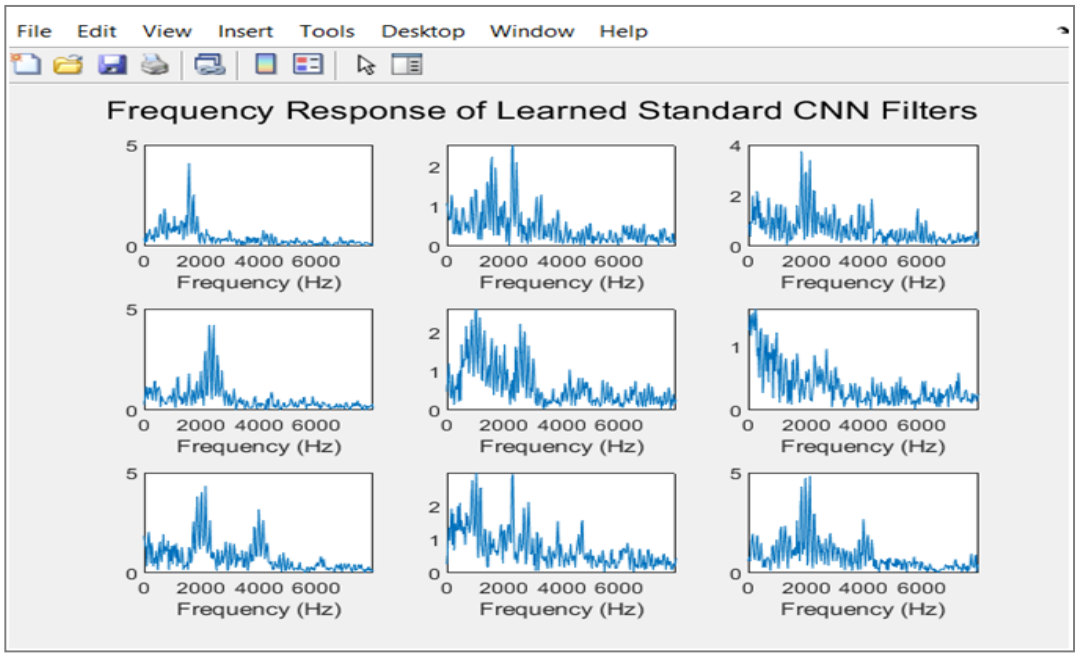

Figure 3.5Réponse fréquentielle des filtres CNN

3.1.6 Implémentation du système Sinc_IAL

Ce modèle d'apprentissage consiste à convoluer un signal de parole avec un ensemble de fonction sinus cardinal (sinc) en cherchant des filtres plus significatifs. On a appliqué cette méthode d'apprentissage pour identifier le locuteur et on a obtenu un taux de classification correcte de 55.72% et une pertes qui vaut 1.5 après 15 époques (figure 3.6).

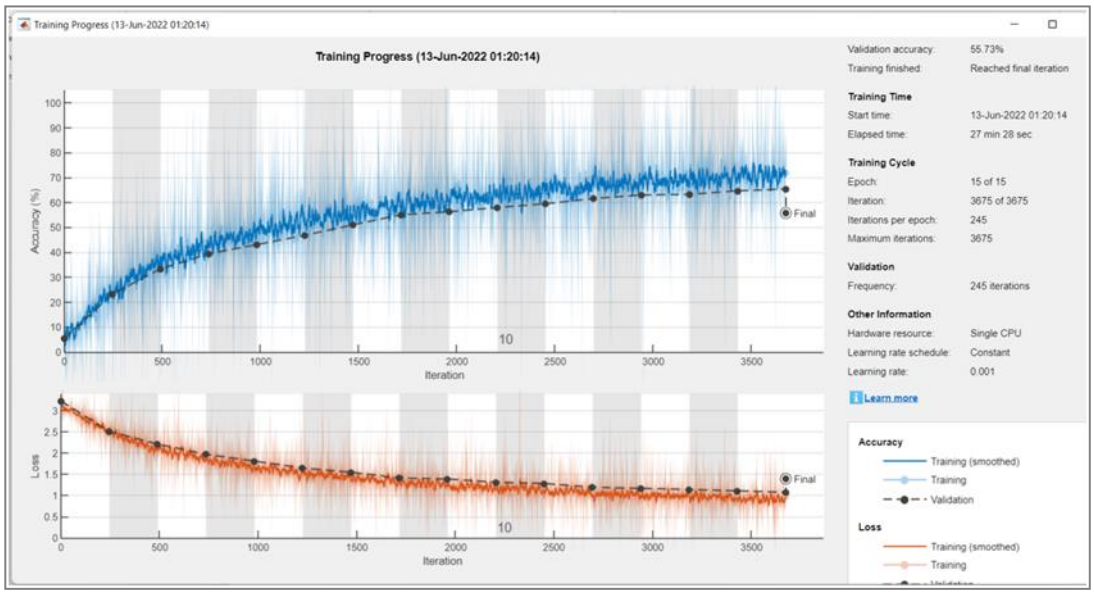

Figure 3.6Apprentissage des systèmes d'identification du locuteur par sinc

La figure 3.7 montre les réponses fréquentielles des filtres d'apprentissages utilisés dans le réseau Sinc.

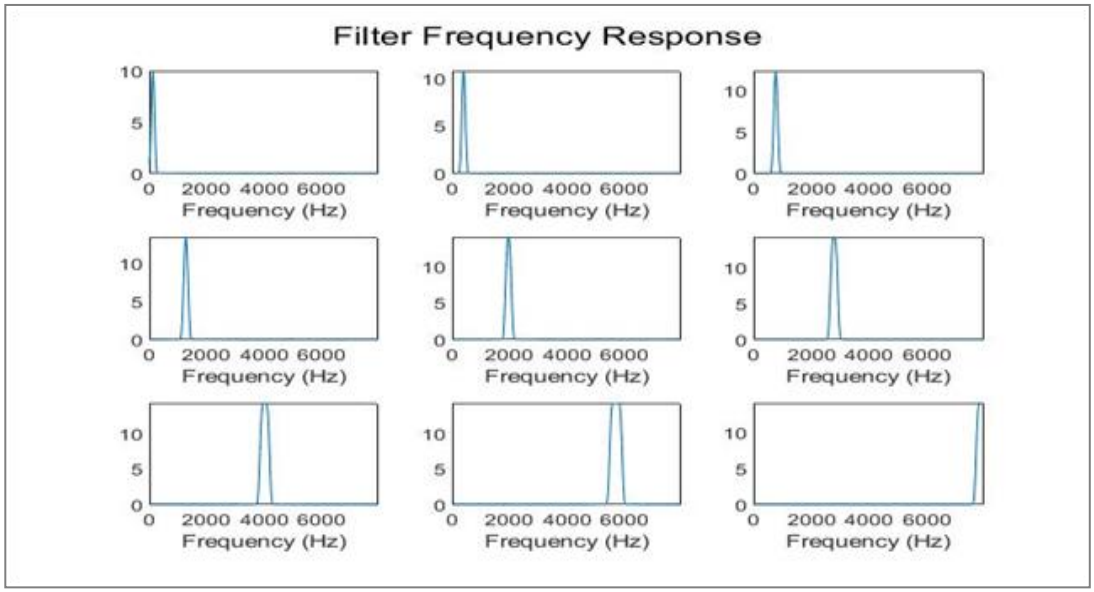

Figure 3.7 Réponse fréquentielle des filtres SINC

3.1.7 Implémentation du système ConstantSinclayer_IAL

Ce modèle d'apprentissage est pratiquement le même que le modèle précédent (sinc) où il utilise des fonctions de sinc lors de la convolution mais avec des largeurs fixes. Les résultats d'identification du locuteur pour ce modèle montrent un taux de classification correcte de 61.54% et un taux de pertes de 1 (figure 3.8).

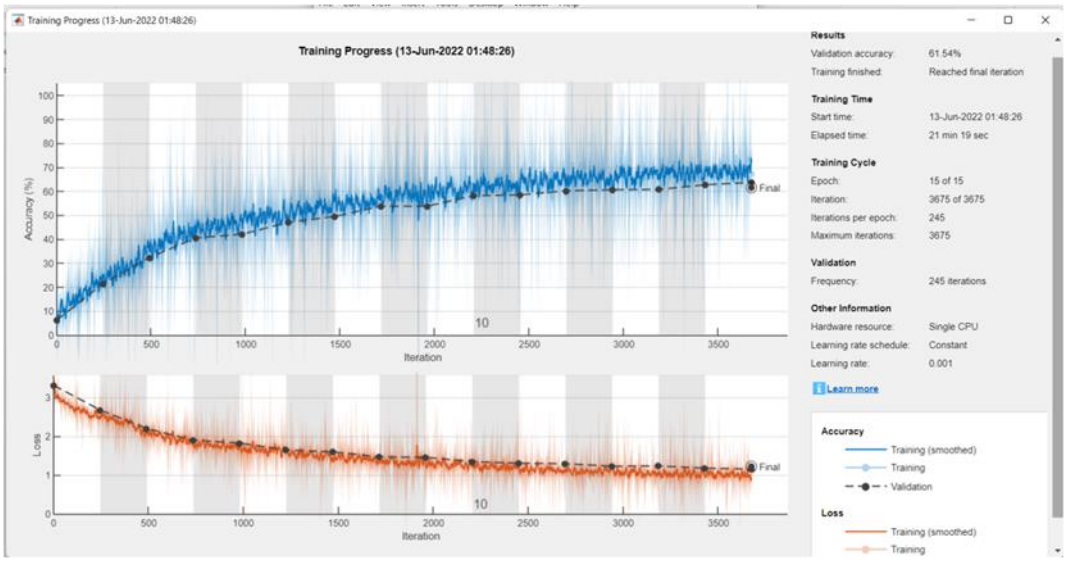

Figure 3.8Apprentissage des systèmes d'identification du locuteur par constantSinclayer

On obtient aussi la réponse de fréquence des filtres (figure 3.9).

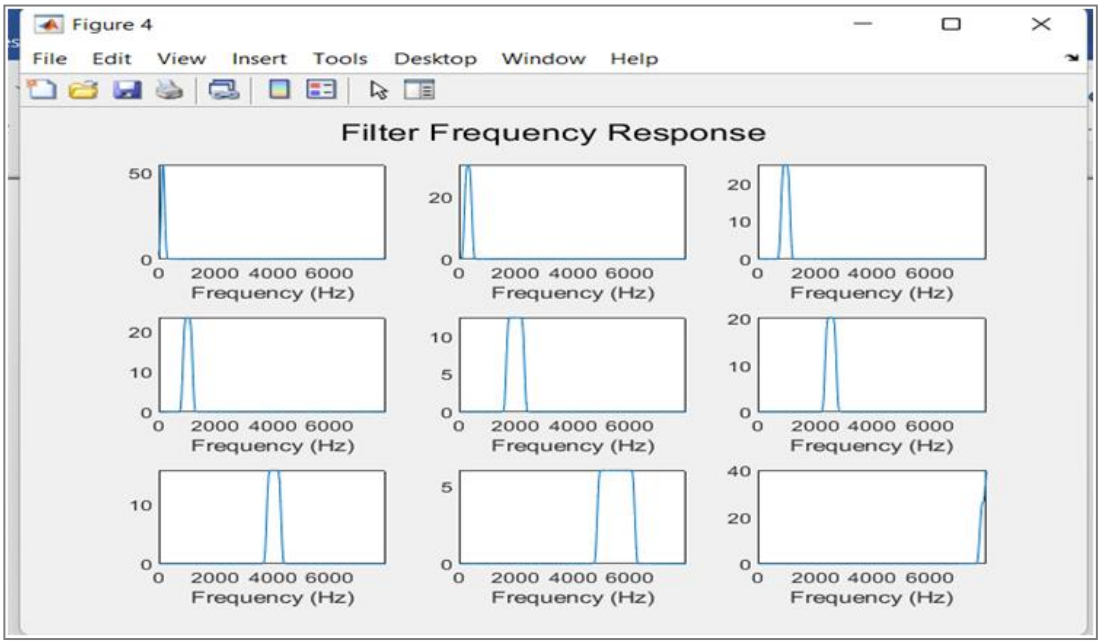

Figure 3.9 Réponse fréquentielle des filtres ConstantSincLayer

3.1.8 Résultats de classification

Pour connaitre le meilleur modèle d'apprentissage d'identifications du locuteur dans notre projet on a fait la comparaison entre les trois modèles selon le taux de classification correct en fonction de nombre d'époques. Le résultat montre que la précision du premier modèle (CNN) est la plus élevée (figure 3.10).

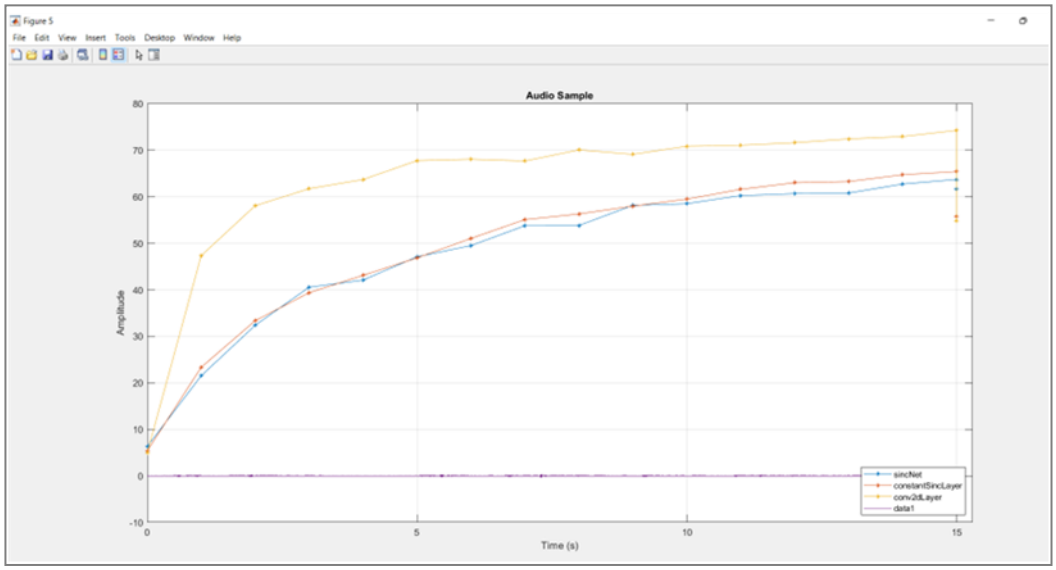

Figure 3.10 Comparaison entre les trois systèmes de deep learning

Conclusion

Dans ce chapitre on a présenté quelques généralités sur l'identification du locuteur par deep learning sur matlab où on a utilisé trois modèles d'apprentissage. On a créer un système d'identification qui a une base de données contient 20 locuteurs. Après l'apprentissage, on a introduit un extrait vocal pour faire le teste et le système a lancé une analyse de base de données, et les résultats ont montrés que le système a arrivé d'identifier le locuteur par les modèles malgré la variance de la précision.

Conclusion Générale

Le travail que nous avons fait au long de cette étude, se base sur la reconnaissance de la parole et l'identification du locuteur par deep learning en extraction des paramètres MFCC.

La reconnaissance automatique du locuteur est le processus de reconnaître une personne à l'aide de sa voix, en se basant sur l'information véhiculée par le signal de parole. Ce domaine regroupe les taches relatives à la vérification du locuteur et sa identification. En identification, le système identifie une personne parmi un ensemble connu par le système.

Dans cette étude on a utilisé l'apprentissage profond (deep learning) selon trois modèles lesquels ont donné des résultats d'identification avec une précision différente :

- Réseau neuronal convolutif (CNN)
- SincNetLayer
- ConstantSincLayer

Ce mémoire a été organisé comme suit :

Le premier chapitre présente une généralité sur la reconnaissance automatique des locuteurs. Cette présentation introduit les différentes tâches nécessaires en reconnaissance automatique du locuteur (identification et vérification), sa structure générale, et aussi les problèmes rencontrés dans ces systèmes.

L'identification automatique du locuteur, qui est le cadre de notre travail, est introduite aussi dans le deuxième chapitre.

Le dernier chapitre est consacré à l'apprentissage profond du système puis extraire les coefficients MFCC pour les 3 modèles afin de tester si le processus de l'identification à bien marcher et connaître quel modèle a donné le résultat le plus précis.

Les résultats de notre projet ont montré que la méthode d'apprentissage CNN sur 15 itérations a donné la meilleure performance pour l'identification du locuteur.

Références Bibliographiques

[1] Daniel Jurafsky and James H Martin. Speech & language processing. Pearson Education India, 2000. 1, 18, 29, 33

[2] R.Boite et M.Kunt, traitement de la parole,lausanne,Ed. Presses Polytechniques Romandes,1987.

[3] Laurent Buniet. Traitement automatique de la parole en milieu bruité : étude de modèles connexionnistes statiques et dynamiques. PhD thesis, Université Henri Poincaré-Nancy 1, France, 1997. 13

[4] L. R. Rabiner et B. H. Juang, "An introduction to hidden Markov models", IEEE Transactions on Acoustics, Speech, and Signal Processing, volume 3, pages 4–16, Janvier 1986.

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

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

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

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

[9] J.M. Pierrel, "Dialogue oral homme-machine ", Edition Hermes, Paris, 1987.

[10] Vincent Jousse. Identification nommée du locuteur : exploitation conjointe dusignal sonore et de sa transcription. PhD thesis, Université du Maine, 2011.22

[11] Michael F McTear. Spoken dialogue technology: toward the conversational user interface. Springer Science & Business Media, 2004. 23

[12] Randy Allen Harris. Voice interaction design: crafting the new conversational speech systems. Elsevier, 2004. 23

[13] Christian Raymond. Décodage conceptuel : coarticulation des processus de transcription et compréhension dans les systèmes de dialogue. PhD thesis, Uni versité d'Avignon et des Pays de Vaucluse, 2005. 18, 23

[14] George Saon and Jen-TzungChien. Large-vocabulary continuous speech recognition systems: A look at some recent advances. IEEE Signal Processing Magazine, 29(6):18–33, 2012. 23

[15] L.R Rabiner and B.H. Juan, Fundamentals of speech recognition. Englewood Cliffs, N.J., USA:Prentice-Hall, 1993.

[16] Mémoire d'ingéniorat 2007 « Application des MFCCs à la reconnaissance des phonèmes arabes '' , Université Saad dahleb Blida .