MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE ABDELHAMID IBN BADIS - MOSTAGANEM

Faculté des Sciences Exactes et d'Informatique

Département de Mathématiques et informatique

Filière : Informatique

MEMOIRE DE FIN D'ETUDES

Pour l'Obtention du Diplôme de Master en Informatique

Option : **Ingénierie des Systèmes d'Information**

Présenté par :

AMAR Karima

HAMMOU Nadjet

THÈME:

Reconnaissance vocale du genre basée sur l'apprentissage profond

Soutenu le:

Devant le jury composé de :

M^me HOCINE Nadia	MCA	Université de Mostaganem	Président
M^r SEHABA Karim	PROF	Université de Mostaganem	Examinateur
M^r MOUMENE Med El Amine	MCA	Université de Mostaganem	Encadreur

Année Universitaire 2022-2023

Résumé

Identifier le genre à partir de la parole a toujours été une tâche difficile. Il s'agit d'une condition très courante et nécessaire dans tous les domaines, y compris le secteur de la santé, les laboratoires médico-légaux et tout domaine industriel. La parole et les images sont des données importantes pour la reconnaissance du genre. Les mots sont les moyens par lesquels le genre peut être facilement identifié. C'est un signal physiologique qui représente des informations à plusieurs niveaux tels que le contenu linguistique (la langue, les mots, l'accent, etc.), le contenu paralinguistique (le sexe, l'âge, la langue, etc.) et l'émotion.

Le deeplearning est une technique d'apprentissage permettant à un programme de reconnaître le langage parlé. Ce système d'apprentissage et de classification, basé sur des réseaux de neurones artificiels numériques est une technique courante en IA, permettant aux machines d'apprendre et reconnaitre des objets, cette dernière est l'approche la plus prometteuse pour notre sujet. Dans ce projet donc, on s'est posé comme objectif principal la mise au point d'un système de classification basé sur l'apprentissage profond pour la reconnaissance du genre à l'aide de la parole.

Mots-clés:IA**,** Apprentissage profond, Genre

Abstract

Gender identification based on spoken language has always been a challenging task, and it plays a crucial role in various fields, including healthcare, forensic laboratories, and industries. Both speech and image data are valuable sources for gender detection, but words are particularly effective in identifying gender. Speech is a physiological signal that conveys information on multiple levels, including linguistic content (language, words, accent, etc.), paralinguistic content (gender, age, language, etc.), and emotion.

Deep learning, a powerful learning technique, enables programs to recognize spoken language. It leverages digital artificial neural networks, a common approach in AI, to facilitate learning and object recognition. For our project, the primary objective was to develop a gender recognition classification system using deep learning, which shows great promise in this domain.

Keywords: AI**,** Deep learning, Gender

Dédicace

Je dédie ce modeste travail à :

Mes **parents**. Aucun hommage ne pourrait être à la hauteur de l'amour dont ils ne cessent de me combler. Que dieu leur procure bonne santé et longue vie.

A celui que j'aime beaucoup et qui m'a soutenue tout au long de ce projet mon **fiancé**, et bien sûr A mon frère **Fadi**, sans oublier mes **tantes** que j'aime.

A toute ma famille, et mes amis, A mon binôme **Karima**,

Mes **chats** adorés Simba, Kitty, Katty, Rex, Mirou, Mira, Gucci.

Nadjet

Dédicace

Je dédie d'abord et avant tout ce projet de fin de stage à ma **maman** qui m'a encouragé pendant mes études et qui m'a beaucoup aidé.

A mes très chères **Frères**, mes **grands-parents** et à toute ma famille et mes chères amies.

Ainsi que ma binôme **Nadjet**.

Karima

Remercîment

Nous remercions le bon DIEU qui nous a donné la vie et la santé pour pouvoir réaliser ce mémoire.

En préambule à ce mémoire, nous souhaitons adresser nos remerciements les plus sincères aux personnes qui nous ont apporté leur aide et qui ont contribué à l'élaboration de ce modeste travail.

Nous exprimons notre profonde gratitude et notre reconnaissance à nos parents pour leurs contributions, leurs soutiens et de leurs encouragements pour l'accomplissement de ce modeste travail.

Nous remercions nos meilleures amies et surtout **Fouad Mokrani** qui a constamment fait preuve d'un soutien inébranlable et qui est resté à nos côtes jusqu'au dernier moment, mérite notre reconnaissance la plus sincère pour tout ce qu'il a accompli et bien sur notre précieux **Youcef Mokhtari**.

Nous étions enchantées de travailler sur ce projet et on espère que le présent rapport reflète cet enthousiasme. On tient enfin à remercier les membres du jury qui nous feront l'honneur d'évaluer la contribution de ce travail.

Liste des figures

Liste des abréviations

Abréviation	Expression Complète
ASR	Automatic Speech Recognition
ANN	Artificial Neural Networks
CNN	Convolutional neural networks
DTW	Dynamic time warping
HMM	Hidden Markov Models
IA	Intelligence artificielle
IHM	Interfaces homme-machine
MLP	Multilayer perceptron
MCC	Coefficient de corrélation de Matthews
MFCC	Mel-Frequency Cepstral Coefficients
RAP	Reconnaissance Automatique de la Parole
ROC	Receiver Operating Characteristics
ReLU	RectifiedLinear Unit

Table des matières

Introduction Générale

La voix humaine est porteuse de la parole. Elle permet principalement la communication mais elle contient également des caractéristiques non linguistiques uniques aux locuteurs. Elle peut indiquer diverses caractéristiques, le genre, l'âge ou bien l'émotion.

Dans l'étude de la parole, la forme est aussi importante que le fond, car la manière dont on raconte fait partie du sens, par exemple dans un crime, une publicité qui vise les consommateurs.

C'est donc par la parole, qui nous semble être l'outil le plus pertinent, que nous tenterons de comprendre le système social de la communauté.

Nous ne retrouvons actuellement dans une ère gouvernée par les technologies cognitives, on découvre la réalité virtuelle ou augmentée, la reconnaissance visuelle et la reconnaissance vocale.

Nous vous invitons à en savoir plus sur la reconnaissance vocale en apprentissage profond à travers ce travail. Bien sûr, c'est tout ce qu'il faut pour comprendre le monde de la technologie vocale.

L'intelligence artificielle vise à imiter le fonctionnement du cerveau humain, ou du moins sa logique en matière de prise de décision. Jean-Claude Heudin, Yann Le Cun, " L'apprentissage profond, une révolution en intelligence artificielle » [1]. Le deeplearning fait l'objet d'importants investissements privés, notamment de la part des grands acteurs du web, mais aussi d'investissements publics. « De plus en plus d'entreprises ont des masses de données gigantesques à exploiter, trier, indexer, et cela demande des ressources considérables. L'intelligence artificielle et le Deeplearning peuvent aider à le faire de façon automatisée et plus efficace » [2], Yann Le Cun confirme qu'il est prudent quant aux fantasmes que suscitent ces évolutions. De nombreux progrès ont été réalisés, notamment dans la vision et la reconnaissance vocale.

Le concept d'apprentissage profond émerge au début des années 2010 avec la redécouverte des réseaux de neurones artificiels. L'apprentissage machine nécessite de grandes

bases de données et leur traitement bénéficie également des progrès technologiques autant des supports physiques que des logiciels. La recherche scientifique en apprentissage profond se multiplie. Tous les domaines des sciences sont concernés et on compare ces nouvelles méthodes avec les méthodes d'apprentissage automatique plus classiques déjà en usage.

La reconnaissance vocale est une utilisation qui n'a plus besoin de preuve. En effet, les interfaces vocales et les assistants vocaux sont aujourd'hui plus performants que jamais et se développent dans de nombreux domaines. Cette croissance exponentielle et continue a conduit à une diversification des applications de reconnaissance vocale et des technologies associées.

Dans ce PFE on va cibler la reconnaissance vocale qui à elle seule est un domaine très vaste en recherche, beaucoup de travaux y sont consacrés, d'où notre intérêt pour ce sujet, et donc mettre un programme d'un système de classification basé sur l'apprentissage en profondeur pour la reconnaissance du genre à l'aide de la parole. Notre PFE est composé en 3 chapitres, le premier présentera la parole soit la reconnaissance automatique de la voix, son déploiement, son fonctionnement et sa nécessité.

Le deuxième chapitre parlera de l'apprentissage profond, ses domaines et ses neurones d'applications, la classification.

Dans le troisième chapitre détaillera notre programme, les méthodes, le dataset à exploiter ainsi que les outils utilisés dans la réalisation de ce mémoire.

1.1. Introduction

La reconnaissance vocale est encore en chantier et attire beaucoup de développeurs amenés par la complexité de la technologie d'analyse de la voix (aussi appelée analyse du locuteur). Cette technologie s'applique avec succès là où les autres technologies sont difficiles à employer. Elle est utilisée dans des secteurs comme les centres d'appel, les opérations bancaires, l'accès à des comptes, sur PC domestiques, pour l'accès à un réseau ou encore pour des applications judiciaires. Le traitement vocal vise donc aussi un gain de productivité puisque c'est la machine qui s'adapte à l'homme pour communiquer, et non l'inverse et c'est pour ça que la reconnaissance vocale est quasiment imparfaite dans son domaine [3].

Pour cet objectif, ce chapitre présente essentiellement une introduction au domaine de la reconnaissance automatique de la parole RAP et ses principales composantes. Une perspective historique sur les inventions clés qui ont permis des progrès dans la reconnaissance vocale est présentée.

1.2. La parole humaine

1.2.1. Définition

La parole constitue le mode de communication le plus naturel dans toute société humaine du fait que son apprentissage s'effectue dès l'enfance. La parole se définit comme étant un signal réel, continu, d'énergie finie et non stationnaire, généré par l'appareil vocal humain [4]. Elle offre un moyen facile aux humains pour établir une communication bien claire.

1.2.2. La production de la parole

La production de la parole est l'une des activités humaines les plus complexes. Ceci, n'est peut-être pas tout à fait surprenant dans la mesure où bon nombre de processus neurologiques et physiologiques complexes sont impliqués dans la génération de la parole. La production de la parole commence dans le cerveau, où s'effectue la création du message et la structure lexico-grammaticale [4][5]. Une fois le message créé, une représentation de la séquence sonore et un certain nombre de commandes, à exécuter par les organes de l'appareil phonatoire, pour produire l'élocution sont nécessaires.

Le mécanisme lourd est plus particulièrement utilisé par les femmes et les enfants. Alors que le mécanisme léger est essentiellement utilisé par les hommes. Sons obtenus par la vibration des cordes vocales ne constituent pas encore des mots. A cet effet, une intervention du reste de l'appareil vocal s'effectue pour en devenir un son.

Le son laryngé est transformé en parole par modulation de différentes manières. Les sons de la parole se distinguent les uns des autres en fonction de l'endroit et de la manière dont ils sont articulés [6].

La Figure 1 illustre les différents constituants de l'appareil phonatoire.

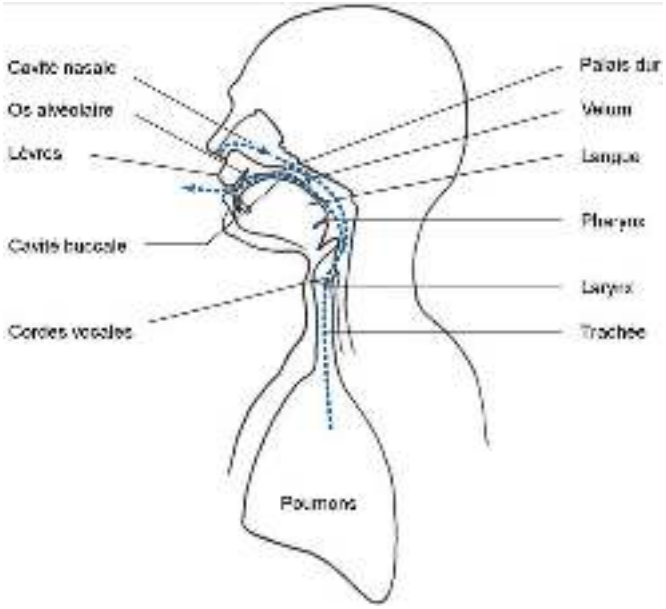

Figure 1–Les différents constituants de l'appareil phonatoire.

1.2.3. La perception de la parole

Le système auditif est divisé de manière anatomique et fonctionnelle en trois zones : oreille externe, oreille moyenne et oreille interne, comme indiqué sur la Figure 2.

Figure2–Structure de l'oreille humaine.

L'oreille externe est composée du pavillon auriculaire et du canal auditif externe. Le pavillon qui représente la partie la plus visible de l'oreille externe capte le son, participe à son amplification et le dirige vers le canal auditif externe. Quant à l'oreille moyenne, elle est composée du tympan et d'une cavité remplie d'air permettant également l'amplification du son. L'oreille interne agit comme un capteur, qui transforme les ondes sonores mécaniques en un signal électrique envoyé au cerveau [6].

La Figure 3 décrit les principales étapes de la perception humaine (partie droite) et illustre également la transcription de ces étapes dans le domaine du traitement du signal (partie gauche) [7].

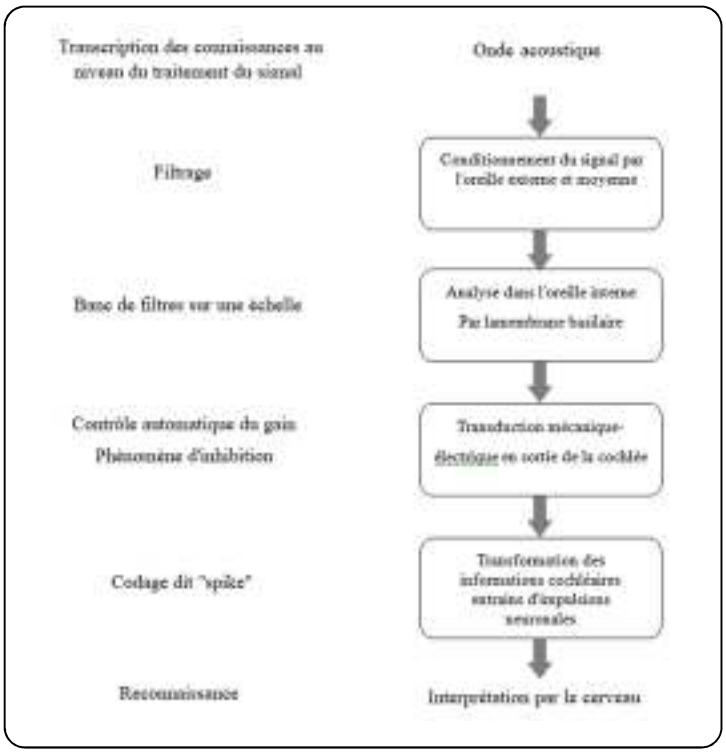

Figure 3 – Analogie entre perception humaine et machine.

1.2.4. Paramètres acoustiques 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 [8] [9]:

1.2.4.1. La fréquence fondamentale

D'un son est une caractéristique en acoustique propre à chaque personne. Elle fonctionne 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 [10]:

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

1.2.4.2. Le spectre fréquentiel

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.

1.2.4.3. L'énergie

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

La Figure 4illustre un exemple réel du signal de parole pour le mot " tash-ghil " dont la signification est "allumer".

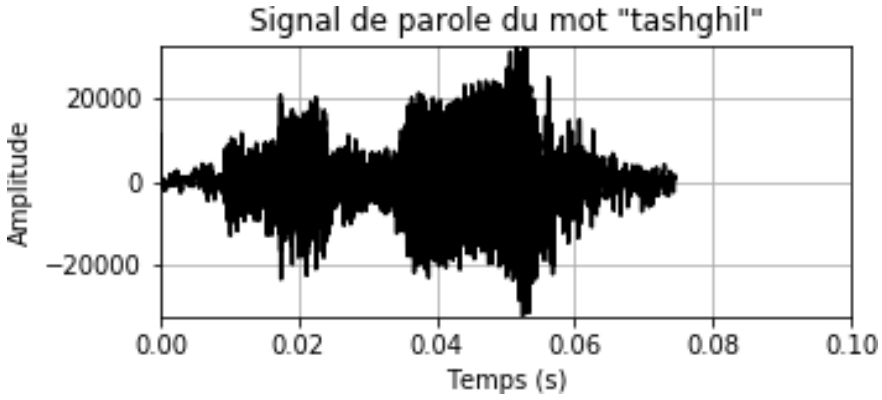

Figure 4–Signal enregistré du mot "tash-ghil " (allumer) [11].

Dans une perspective de reconnaissance, le signal de parole est considéré comme étant un signal très complexe, variable et souvent bruité. Une brève description de ces facteurs est donnée ci-après [10]:

- **Redondance :** L'intelligibilité de la parole est remarquablement robuste aux distorsions du signal acoustique. Il a été montré que même si on supprime ou on masque par du bruit des morceaux du signal de parole à intervalle régulier, le signal reste intelligible, ce qui montre une redondance phonétique dans le signal [11]. Cette redondance offre une certaine résistance au bruit, toutefois, elle rend l'extraction des informations pertinentes par un ordinateur plus délicate.

- **Continuité et coarticulation :** la production d'un son dépend fortement du son qui le précède et celui qui le suit en raison de l'anticipation du geste articulatoire. Cette forte articulation des mots rend la tâche de reconnaissance difficile.

- **Conditions d'enregistrement :** L'enregistrement de signaux vocaux dans des conditions difficiles rend difficile l'extraction des caractéristiques pertinentes nécessaires à la reconnaissance. En effet, les interférences (type, distance, direction) et l'environnement (bruit, réverbération) de la transmission du microphone rendent la reconnaissance vocale très compliquée.

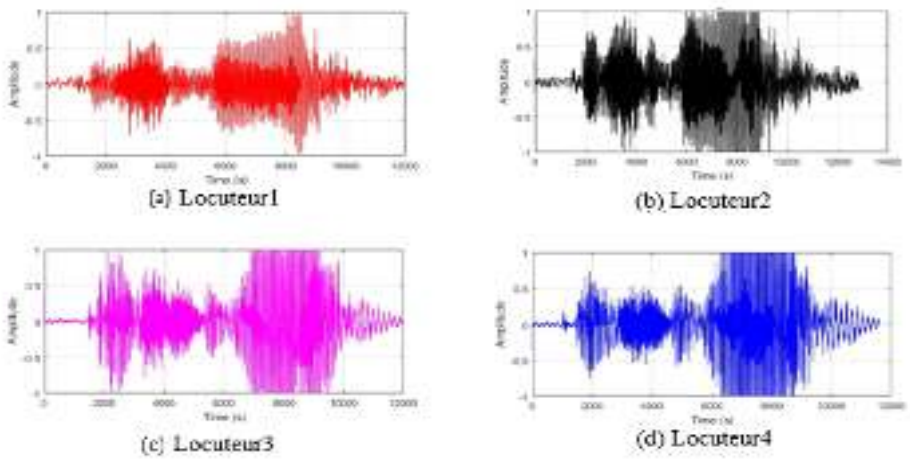

Figure 5–
Différentssignauxdumêmemot"tashghil"prononcépardifférentslocuteurs

En plus de ces difficultés, les signaux vocaux sont dégradés après leur génération en traversant un milieu contenant des interférences (d'abord l'air, puis les microphones et les câbles). En effet, on compte plusieurs interactions telles que:

- D'autres sons qui peuvent s'additionner au signal de parole.
- La forme du signal sonore peut être affectée par la géométrie de la pièce (effet d'écho).
- Le signal acoustique peut être modifié lors de sa conversion par le microphone.

Ces interactions amplifient d'autant la variabilité du signal de parole et augmentent les difficultés pour le reconnaître.

1.2.5. La reconnaissance automatique de la parole

La reconnaissance automatique de la parole est une branche de l'intelligence artificielle dont le but principal est de convertir automatiquement des signaux vocaux en séquences de mots grâce à des algorithmes mis en œuvre par des modules logiciels ou matériels. Par conséquent, l'objectif de la reconnaissance automatique de la parole est de développer des techniques et des systèmes qui reçoivent des signaux vocaux naturels en entrée et présentent leur signification (résultat de la reconnaissance) à la sortie [12].

Un système ASR est tout système permettant à la machine la compréhension et le traitement des informations fournies oralement par un utilisateur humain.

Les systèmes ASR peuvent être classés en plusieurs catégories différentes selon les types d'énoncés qu'ils sont capables de reconnaître. Ces catégories sont basées sur le fait que l'une des difficultés de l'ASR est la capacité de déterminer quand un locuteur commence et termine un énoncé [13].

- **Reconnaissance des mots isolés :** Les systèmes de reconnaissance de mots séparés n'acceptent qu'un seul mot à la fois. Ces systèmes ont généralement un état "n'écouter/ne pas écouter" qui oblige le locuteur à faire une pause entre les énoncés. La reconnaissance de mots convient lorsque le locuteur doit donner une seule réponse au système ASR ou lorsque les mots représentent des commandes.
- **Reconnaissance des mots connectés :** Le système de reconnaissance de mots chaînés permet le traitement de mots séparés par des pauses. Ils sont similaires aux mots simples, mais ils permettent à des énoncés séparés de s'exécuter ensemble avec des pauses minimales.
- **Reconnaissance de la parole continue :** Le système de reconnaissance de mots chaînés permet le traitement de mots séparés par des pauses. Ils sont similaires aux mots simples, mais ils permettent à des énoncés séparés de s'exécuter ensemble avec des pauses minimales.

- **Reconnaissance de la parole spontanée :** La parole spontanée peut être considérée comme une parole naturelle dont le contenu est inconnu à l'avance. Un système ASR qui gère la parole spontanée réelle doit être capable de gérer diverses caractéristiques de la parole naturelle telles que les mots prononcés simultanément (bégaiement léger et non-mots, par exemple : "um", "ah", etc.).

1.3. La reconnaissance vocale

1.3.1. Définition

La reconnaissance vocale ou reconnaissance automatique de la parole (Automatic Speech Recognition **ASR**) est une technique informatique qui permet d'analyser un mot ou une phrase captée au moyen d'un microphone pour la transcrire sous la forme d'un texte exploitable par une machine. La reconnaissance vocale, ainsi que la synthèse vocale, l'identification du locuteur ou la vérification du locuteur, font partie des techniques de traitement de la parole [13].

Ces techniques permettent notamment de réaliser des interfaces vocales c'est-à-dire des interfaces homme-machine (IHM) où une partie de l'interaction se fait à la voix. Parmi les nombreuses applications, on peut citer les applications de dictée vocale sur PC où la difficulté tient à la taille du vocabulaire et à la longueur des phrases, mais aussi les applications téléphoniques de type serveur vocal, où la difficulté tient plutôt à la nécessité de reconnaître n'importe quelle voix dans des conditions acoustiques variables et souvent bruyantes (téléphones mobiles dans des lieux publics).L'objectif de la reconnaissance vocale peut être l'identification du genre, l'émotion, authentification.

1.3.2. Principe de base

Une phrase enregistrée et numérisée est donnée au programme de reconnaissance vocale. Dans le formalisme ASR, le découpage fonctionnel est le suivant :

- Le traitement acoustique vise à numériser le signal de parole sous forme de vecteurs acoustiques qui constituent les données d'observation pour le système de reconnaissance. Le signal est alors numérisé et paramétré par une technique d'analyse fréquentielle utilisant les transformées de Fourier.

- L'apprentissage automatique qui réalise une association entre les segments élémentaires de paroles et les éléments lexicaux. Cette association fait appel à une modélisation statistique entre autres par modèles de Markov cachés (**HMM**, Hidden Markov Models) et/ou par réseaux de neurones artificiels (**ANN**, Artificial Neural Networks).

- La reconnaissance (back-end) qui en concaténant les segments élémentaires de paroles précédemment appris reconstitue le discours le plus probable. Il s'agit donc d'une correspondance de motif (pattern matching) temporelle, réalisée souvent par l'algorithme de déformation temporelle dynamique (en anglais **DTW**, dynamic time warping) [14].

1.3.2.1. On utilise la reconnaissance vocale dans différents domaines

- Une dictée vocale peut être associée à un traitement de texte : Un locuteur parle et le texte s'affiche ; ainsi, il n'a plus besoin de taper son texte au clavier.

- Les serveurs d'informations par téléphone.

- La messagerie.

- Elle permet l'autonomie : par exemple en médecine, lorsqu'un chirurgien a les deux mains occupées, il peut parler pour demander une information technique au lieu de taper sur un clavier (autonomie qui est aussi valable en industrie).

- La sécurité possible grâce à la signature vocale, La possibilité de commande et de contrôle d'appareils à distance.

- La classification : Un dispositif de reconnaissance vocale se classifie par un petit nombre de paramètres nommés modes de reconnaissance qui sont corrélés aux difficultés suivantes :

- Variabilité inter et intra-locuteur : Les dispositifs mono locuteurs effectuent un apprentissage in-situ des mots. Les dispositifs multilocuteurs sont capables de reconnaître un corpus fixe quel que soit le locuteur. Les dispositifs monolocuteurs sont les plus communs et tendent surtout à se généraliser grâce à la synthèse text to speech qui évite la phase d'apprentissage.
- Taille du vocabulaire
- Environnement

1.4. La classification du genre par le signal audio

L'information sur le genre est une propriété distinctive et la plus importante dans un discours. La détermination de ces informations à partir d'un signal de parole est un sujet important. Les systèmes de vérification des locuteurs utilisent également implicitement ou explicitement des informations sur le genre. En général, l'identification d'un genre de locuteur est importante pour des systèmes de dialogue de plus en plus naturels et personnalisés. Il existe un ensemble de fonctionnalités utilisées pour reconnaître le genre de la voix.

1.5. Réseaux neuronaux

Principalement mis à profit pour les algorithmes de l'apprentissage en profondeur, les réseaux neuronaux traitent les données d'entraînement en imitant l'inter-connectivité du cerveau humain par le biais de couches de nœuds. Si cette valeur de sortie dépasse un seuil donné, elle « déclenche » ou active le nœud, en transmettant les données à la couche suivante du réseau. [15].

1.6. Conclusion

Dans ce chapitre nous avons entamé le domaine de la reconnaissance automatique de la parole en présentant initialement la parole humaine comme acteur principal, ensuite les caractéristiques qui sont liées à la difficulté de sa reconnaissance sont présentées, la reconnaissance vocale. Dans le chapitre prochain nous allons parler du deeplearning et ses principes.

2.1. Introduction

Dans ce chapitre, nous présentons un aperçu détaillé des architectures de deeplearning. Nous proposons un aperçu de l'intelligence artificielle, de la machine learning et du deeplearning, en soulignant la différence entre eux, nous présentons une introduction aux modèles de deeplearning.

2.2. Intelligence artificielle IA

2.2.1. Définition

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

2.3. Le deep learning

2.3.1. Définition

Le deeplearning ou apprentissage profond est un sous-domaine de l'intelligence artificielle (IA). Ce terme désigne l'ensemble des techniques d'apprentissage automatique (machine learning), autrement dit une forme d'apprentissage fondée sur des approches mathématiques, utilisées pour modéliser des données. Pour mieux comprendre ces techniques, il faut remonter aux origines de l'intelligence artificielle en 1950, année pendant laquelle Alan Turning s'intéresse aux machines capables de penser.

Cette réflexion va donner naissance à la machine learning, une machine qui communique et se comporte en fonction des informations stockées. Ces neurones sont interconnectés pour traiter et mémoriser des informations, comparer des problèmes ou situations quelconques avec des situations similaires passées, analyser les solutions et résoudre le problème de la meilleure façon possible [16].

2.3.2. Domaine d'application du deeplearning

Le deeplearning est utilisé dans de nombreux domaines :

- Reconnaissance d'image;
- **Reconnaissance vocale;**
- Traduction automatique;
- Voiture autonome;

- Diagnostic médical;
- Recommandations personnalisées;
- Modération automatique des réseaux sociaux;
- Prédiction financière et trading automatisé;
- Identification de pieces défectueuses;
- Détection de malwares ou de fraudes;
- Chatbots (agents conversationnels);
- Exploration spatiale;
- Robots intelligents.

2.4. Réseau de neurones artificiels

Par le deeplearning, nous entendons un mode d'apprentissage automatique géré par un réseau de neurones artificiels. Comme les neurones du cerveau, les neurones artificiels arrivent donc à communiquer pour favoriser l'apprentissage et l'assimilation de divers éléments d'information [17].

2.4.1. Définition

Le réseau de neurones artificiels copie le cerveau humain pour favoriser l'apprentissage. Il s'agit donc d'un système qui se base sur le fonctionnement du cerveau humain pour l'adapter à des ordinateurs équipés de fonctions d'intelligence artificielle. Grâce au réseau de neurones artificiels, l'ordinateur arrive à résoudre des problèmes de manière autonome. Le réseau améliore aussi les capacités de l'ordinateur.

2.4.2. Le fonctionnement du réseau de neurones artificiels

Le réseau de neurones artificiels est basé sur plusieurs processeurs fonctionnant en parallèle. Ces processeurs sont organisés en tiers. Le premier tiers a pour fonction de recevoir les entrées de données brutes. Chacun des tiers reçoit ensuite les sorties d'informations transmises par le tiers précédent. Le dernier tiers est chargé de produire les résultats du système. Plus le problème est complexe, plus il faut de couches pour le traiter.

Comme pour le cerveau humain, les réseaux de neurones artificiels ne peuvent pas être programmés directement, mais doivent apprendre en étudiant et en analysant des exemples. Il existe trois méthodes d'apprentissage, soit :

- **L'apprentissage supervisé** : L'algorithme s'entraîne à partir de données étiquetées.

- **L'apprentissage non supervisé** : le réseau de neurones doit analyser un ensemble de données qui ne sont pas étiquetées.

- **L'apprentissage renforcé** : il s'agit d'une méthode par laquelle on procède par renforcements et sanctions selon que les résultats sont positifs ou négatifs.

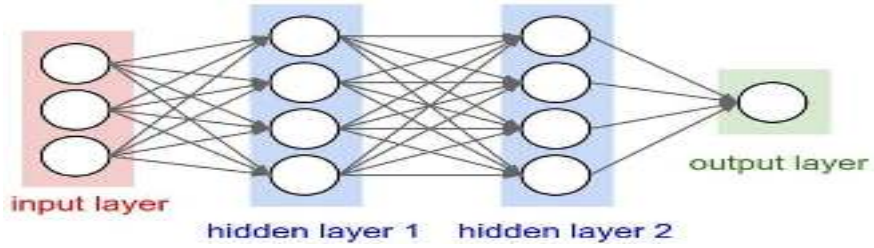

Figure 6 - Fonctionnement d'un réseau de neurones artificiels

2.4.3. Un perceptron

Commençons cependant par le commencement et le perceptron, ce réseau de neurone à une seule couche inventée par Rosenblatt en 1957.

Le perceptron est formée d'une première couche d'unités (ou neurones) qui permettent de « lire » les données : chaque unité correspond à une des variables d'entrée. On peut rajouter une unité de biais qui est toujours activée (elle transmet 1 quelles que soient les données). Ces unités sont reliées à une seule et unique unité de sortie, qui reçoit la somme des unités qui lui sont reliées, pondérée par des poids de connexion. Pour **p** variables **x1,x2,…,xp** , la sortie reçoit donc $\mathbf{w0+\sum_{j=1}^{p} w_j x_j}$.

L'unité de sortie applique alors une fonction d'activation a à cette sortie.

Un perceptron prédit donc grâce à une fonction de décision **f** définie par $\mathbf{f(x)=a(\sum_{j=1}^{p} w_j x_j+w0)}$. Cette fonction a une forme explicite, il s'agit bien d'un modèle paramétrique [18].

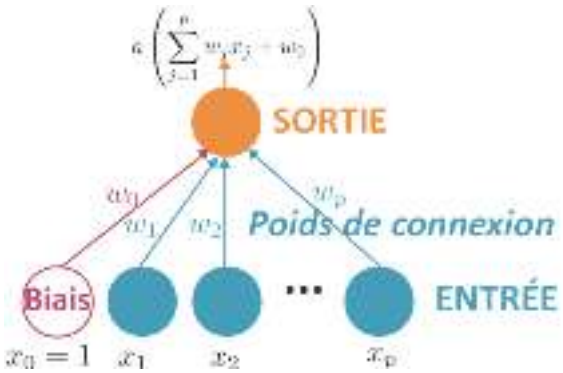

Figure 7 –Un perceptron

2.4.4. Architecture des ANN

Il est également connu sous le nom de réseau de neurones artificiels. Il s'agit d'un réseau neuronal à anticipation car les entrées sont envoyées dans le sens direct. Il peut également contenir des couches cachées qui peuvent rendre le modèle encore plus dense. Ils ont une longueur fixe spécifiée par le programmeur. Il est utilisé pour les données textuelles ou les données tabulaires. Une application réelle largement utilisée est la reconnaissance faciale. Il est comparativement moins puissant que CNN [19].

2.4.5. La propagation vers l'avant (Forward propagation)

Est le moyen de passer de la couche d'entrée (à gauche) à la couche de sortie (à droite) dans le réseau de neurones. Le processus de déplacement de droite à gauche, c'est-à-dire en arrière de la couche de sortie à la couche d'entrée, s'appelle la propagation vers l'arrière.

2.4.6. Erreurs dans le réseau neuronal

Jusqu'à présent, nous avons vu comment la propagation vers l'avant nous aide à calculer les sorties. Disons que pour une ligne particulière, la cible réelle est 0 et la cible prédite est 0,5. Nous pouvons utiliser cette valeur prédite pour calculer l'erreur pour une ligne particulière. Le type d'erreur que nous avons choisi ici est SquaredError. $\mathbf{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

2.4.7. Fonction de coût (Cost)

La sortie de la propagation vers l'avant est la probabilité d'événements binaires. Ensuite, la probabilité est comparée à la variable de réponse pour calculer le coût. L'entropie croisée est utilisée comme fonction de coût dans le problème de classification. L'erreur quadratique moyenne est utilisée comme fonction de coût dans le problème de régression. La formule pour l'entropie croisée est indiquée ci-dessous. $C(y, \hat{y}) = -\frac{1}{m}(y \cdot \log(\hat{y}) + (1 - y) \cdot \log(1 - \hat{y}))$

2.4.8. Fonction Perte (Loss)

Dans l'optimisation mathématique et la théorie de la décision, une fonction de perte ou de coût (parfois aussi appelée fonction d'erreur) est une fonction qui mappe un événement ou des valeurs d'une ou plusieurs variables sur un nombre réel représentant intuitivement un certain "coût" associé à l'événement.

En termes simples, la fonction de perte est une méthode d'évaluation de la qualité de la modélisation de votre ensemble de données par votre algorithme. C'est une fonction mathématique des paramètres de l'algorithme d'apprentissage automatique.

Dans la régression linéaire simple, la prédiction est calculée à l'aide de la pente (**m**) et de l'ordonnée à l'origine (**b**). La fonction de perte pour cela est le (Yi - Yihat) ^ 2, c'est-à-dire que la fonction de perte est la fonction de la pente et de l'interception [19].

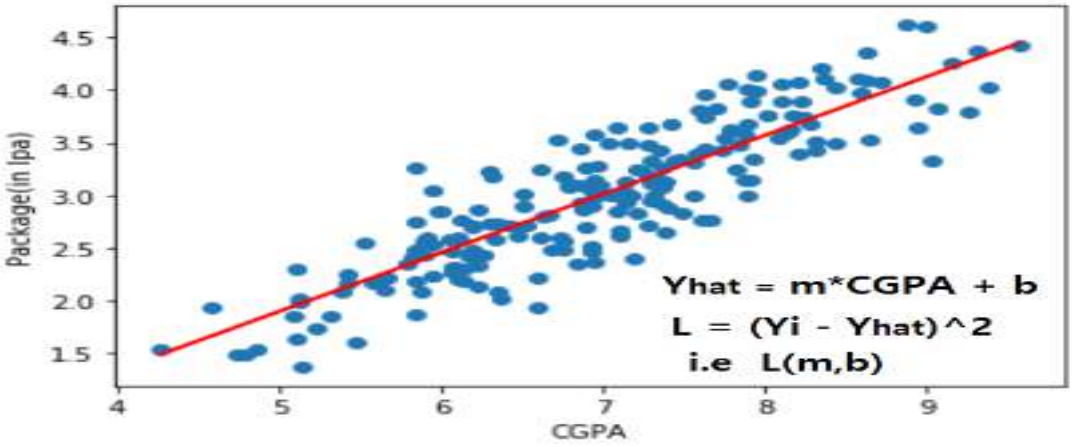

Figure 8 - La fonction de la pente et de l'interception

2.4.9. La propagation vers l'arrière

Est la méthode préférable pour ajuster ou corriger les poids pour atteindre la fonction de perte minimisée.

2.4.10. Algorithme d'optimisation de la descente de gradient

Gradient Descent est un solveur itératif. Le solveur itératif ne donne pas la solution exacte. Pas dans tous les cas, la fonction objectif est résoluble.

Dans de tels cas, les solveurs itératifs sont utilisés pour obtenir la solution approchée car le but est de minimiser la fonction objective.

Le principe de base de la descente de gradient est de choisir la taille du pas (également appelée taux d'apprentissage) de manière appropriée afin que nous puissions nous rapprocher de la solution exacte. Ainsi, le taux d'apprentissage contrôle essentiellement la taille d'un pas en descente [19].

La règle de mise à jour de descente de gradient est donnée comme suit :

$$W_j^{k+1} = W_j^k - \triangle W_j$$

Update on the j^{th} weight in the $k+1^{th}$ iteration
W_j^{k+1} is the next position
W_j^k is the current position
$\triangle W_j$ is the slope / derivative

Gradient Descent peut être résumé à l'aide de la formule

$$W_j^{k+1} = W_j^k - \left[\alpha \cdot \sum (\hat{Y} - Y) \cdot Xj \right]$$

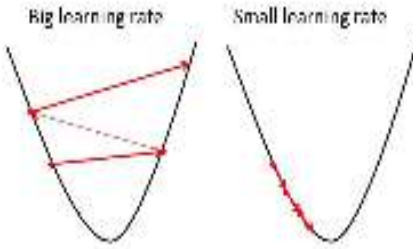

Figure 9- Gradient descent

2.4.11. Perceptron multicouche (Multilayer perceptron MLP)

Un type de réseau neuronal artificiel organisé en plusieurs couches. L'information circule de la couche d'entrée vers la couche de sortie uniquement : il s'agit donc d'un réseau à propagation directe (feedforward). Chaque couche est constituée d'un nombre variable de neurones, les neurones de la dernière couche dite « de sortie » étant les sorties du système global.

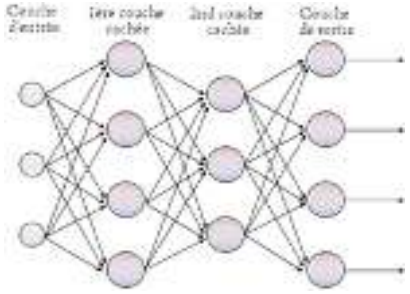

Figure 10-Perceptron multicouche

2.4.12. Learning Rate

Le taux d'apprentissage définit la rapidité avec laquelle un réseau met à jour ses paramètres.

Un faible taux d'apprentissage ralentit le processus d'apprentissage mais converge en douceur. Un taux d'apprentissage plus élevé accélère l'apprentissage mais peut ne pas converger.

Habituellement, un taux d'apprentissage décroissant est préféré.

2.4.13. Le sur-ajustement (Overfitting)

C'est un écueil courant dans les algorithmes d'apprentissage en profondeur dans lequel un modèle essaie de s'adapter entièrement aux données d'apprentissage et finit par mémoriser les modèles de données et le bruit et les fluctuations aléatoires.

Ces modèles ne parviennent pas à généraliser et à bien fonctionner dans le cas de scénarios de données invisibles, ce qui va à l'encontre de l'objectif du modèle.

2.4.13.1. Quand le sur-ajustement peut-il se produire ?

La variance élevée des performances du modèle est un indicateur d'un problème de sur-ajustement.

Le temps d'apprentissage du modèle ou sa complexité architecturale peuvent entraîner un sur-ajustement du modèle. Si le modèle s'entraîne trop longtemps sur les données d'entraînement ou est trop complexe, il apprend le bruit ou des informations non pertinentes dans l'ensemble de données.

2.4.14. Validation croisée (cross validation)

La validation croisée est une mesure robuste pour éviter le sur-ajustement. L'ensemble de données complet est divisé en parties. Dans la validation croisée K-folds standard, nous devons partitionner les données en k-folds. Ensuite, nous entraînons itérativement l'algorithme sur k-1 plis tout en utilisant le pli restant comme ensemble de test. Cette méthode nous permet d'ajuster les hyper paramètres du réseau de neurones ou du modèle d'apprentissage automatique et de le tester à l'aide de données totalement inédites [19].

2.4.15. La régularisation (Regularization)

Est un ensemble de techniques qui peuvent empêcher le sur-ajustement dans les réseaux de neurones et ainsi améliorer la précision d'un modèle d'apprentissage en profondeur face à des données complètement nouvelles du domaine problématique.

2.4.16. Performances du modèle de mesure (Measuring model performance)

Lorsque l'on cherche à prédire les valeurs d'une variable **Y** de nature quantitative, on parle de régression. Lorsque la variable **Y** à prédire est de nature qualitative, on parle alors de classification. XLSTAT possède plusieurs modèles d'apprentissage en régression et en classification.

Nous avons donc une variable d'intérêt à prédire et plus la prédiction de l'algorithme est proche de la variable cible, plus le modèle sera performant.

Il est important de pouvoir évaluer les performances d'un modèle pour mesurer les risques mais également pour comparer plusieurs algorithmes et/ou modèles.

Le module Indicateurs de performance a été développé principalement pour nous aider à répondre à la question suivante : À quel point je peux faire confiance à un modèle pour prédire des évènements futurs ?

2.4.16.1. Indicateurs de performance de modèles de classification

- **Notations :VP** (Vrais Positifs), **VN** (Vrais Négatifs), **FP** (Faux Positifs) et **FN** (Faux Négatifs).
- **Précision** : la précision est le rapport **VP/(VP+FP).** Elle correspond à la proportion de prédictions positives effectivement correcte.
- **Prévalence de l'événement** : fréquence de survenance de l'événement dans l'échantillon total **(VP+FN) /N**.
- **F-mesure** : la F-mesure aussi appelée F-score ou score-F1 peut être interprétée comme une moyenne pondérée de la précision et du rappel ou sensibilité. Sa valeur est comprise entre 0 et 1.
- **MCC** (coefficient de corrélation de Matthews) : le coefficient de corrélation de Matthews (MCC) ou coefficient phi est utilisé dans l'apprentissage automatique comme une mesure de la qualité des classifications binaires (à deux classes).

- **Courbe Roc** : la courbe ROC (Receiver Operating Characteristics) permet de visualiser la performance d'un modèle et de la comparer à celle d'autres modèles. Les termes utilisés viennent de la théorie de détection du signal. La courbe des points (1-spécificité, sensibilité) est la courbe ROC [18].

2.5. Convolutional neural networks CNN

Il est également connu sous le nom de réseaux de neurones convolutifs. Il est principalement utilisé pour les données d'image. Il est utilisé pour la vision par ordinateur. Certaines des applications réelles sont la détection d'objets dans les véhicules autonomes. Il contient une combinaison de couches convolutionnelles et de neurones.

2.5.1. Convolutional

Généralement, de nombreuses paires de couches de convolution et de mise en commun sont répétitives suivies d'une couche entièrement connectée et une couche de classification. Ces couches sont assemblées pour créer un modèle profond pour extraction automatique de caractéristiques à partir de spectrogrammes de parole [20]. La couche convolutive est la couche centrale du CNN. Tous les neurones de cette couche ont un petit champ récepteur dans le spectrogramme, image et calcule la sortie du champ récepteur avec un filtre linéaire.

Comme CNN est l'une des techniques DL les plus étudiées, plusieurs autres modèles CNN ont donc été mis en œuvre par des chercheurs tels qu'Alex Net, Caffe Net et GoogLe Net[21] pour la modélisation de phrases, la classification d'images et la reconnaissance de la parole. L'architecture de CNN est illustrée à la figure 11.

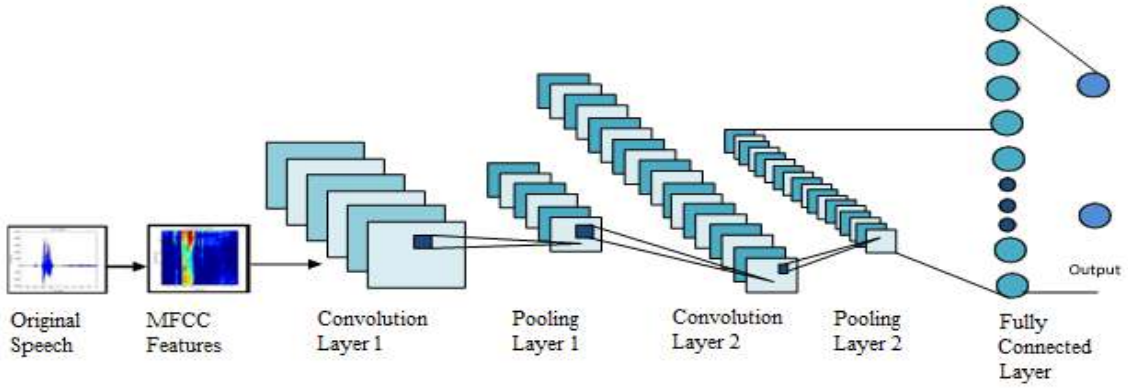

Figure 11- L'architecture de CNN[22]

2.6. Conclusion

Dans ce chapitre, nous avons parlé du deeplearning, ses architectures et ses algorithmes d’entraînement (convolutional neural networks CNN). Dans le chapitre suivant, nous présenterons les outils de conception qui nous ont permis d’accomplir ce travail, et par la suite, on présente les résultats de l’ensemble des tests effectués sur les échantillons de notre dataset.

3.1. Introduction

La phase la plus importante du projet est la phase de développement afin d'atteindre notre objectif de réaliser un programme robuste se rapprochant d'une application concrète avec le concept de reconnaissance vocale du genre basée sur le deeplearning.

Dans ce chapitre, nous mentionnerons les différentes étapes suivies pour implémenter notre programme, depuis l'analyse et la conception jusqu'au développement final.

3.2. Les réseaux de neurone convolutif CNN

Les réseaux de neurone convolutif (CNN) ont été créés à partir d'un réseau de neurones formé de plusieurs couches. Chaque couche est formée de neurones qui sont connectés aux neurones de la couche suivante. Ils sont entraînés et en demandant au réseau de les classer.

Un CNN est simplement un empilement de plusieurs couches de convolution, pooling, flatten et fully-connected. Chaque son reçu en entrée va donc être filtré, réduit et corrigé plusieurs fois, pour finalement former un vecteur.

- **Couche de convolution** : La couche de convolution est la composante clé des réseaux de neurones, convolutifs, elle constitue toujours au moins leur première couche. Les couches de convolution sont formées de ce qu'on appelle des filtres. Les filtres sont des tableaux de valeurs appelées feature maps. Chaque couche de convolution prend en entrée un son et produit une feature map.
- **Couche de pooling** : Ce type de couche est souvent placé entre deux couches de convolution : elle reçoit en entrée plusieurs feature maps, et applique à chacune d'entre elles l'opération de pooling. Une couche de pooling, agit comme une couche de réduction. Le max-pooling prend la valeur maximale de chaque « morceau d'un son ».
- **Le flattening**: Cela consiste tout simplement à prendre la totalité des valeurs de nos matrices précédemment calculées, et à les empiler, en vue de les exploiter dans la couche d'entrée d'un réseau de neurones.
- **ReLU**: il signifie Rectified Linear Unit (ReLu) et est utilisé pour un entraînement plus efficace et plus rapide. Il mappe les valeurs négatives sur 0 et conserve les valeurs positives. C'est aussi appelé activation.
- **Couche fully-connected** : Les CNNs sont généralement formés de plusieurs couches de convolution et de pooling, suivies par une couche fully-connected qui combine les features extraites par les couches précédentes pour classifier l'image, elle renvoie un

vecteur de taille N, où N est le nombre de classes dans notre problème de classification du son. Chaque élément du vecteur indique la probabilité pour l'image en entrée d'appartenir à une classe.

3.3. Paramètres d'un signal audio

La conception de la parole et son traitement est la première étape de la reconnaissance de la parole. Il existe de nombreuses techniques, certaines plus puissantes et efficaces que d'autres. Le traitement automatique de la parole repose sur des données analogiques en fonction du temps. L'extraction des paramètres optimaux aide certainement dans ce traitement.

La numérisation consiste à transformer un signal analogique qui contient une quantité infinie d'amplitudes en un signal numérique contenant, lui, une quantité finie de valeurs.

Le passage de l'analogique au numérique repose sur trois étapes successives : l'échantillonnage, la quantification, et le codage.

Amplitude : l'amplitude fait référence au déplacement maximal des molécules d'air à partir de la position de repos.

Crest and Trough : la crête est le point le plus haut de la vague alors que le creux est le point le plus bas.

Wavelength : la distance entre 2 crêtes ou creux successifs s'appelle une longueur d'onde.

Cycle : Chaque signal audio traverse sous forme de cycles. Un mouvement complet vers le haut et un mouvement vers le bas du signal forment un cycle.

Frequency : La fréquence fait référence à la vitesse à laquelle un signal change sur une période de temps.

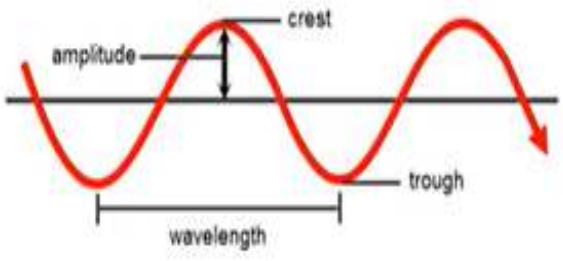

Figure 12 – Paramètre d'un signal audio

3.4. Base de données (DATASET)

Cette base de données [23] a été créée pour identifier une voix comme masculine ou féminine, sur la base des propriétés acoustiques de la voix et de la parole. L'ensemble de données se compose de 30 000 échantillons de voix enregistrés, collectés auprès de locuteurs masculins et féminins, de chiffres parlés (0-9) de 60 dossiers et 500 fichiers audio chacun.

3.4.1. L'ensemble de données

Nous avons utilisé le fichier training_data.csv, il contient un échantillon caractérisé par 21 attributs (columns) :

- **Meanfreq** : fréquence moyenne (en kHz)
- **Sd** : écart type de fréquence
- **Median** : fréquence médiane (en kHz)
- **Q25** : premier quantile (en kHz)
- **Q75** : troisième quantile (en kHz)
- **IQR** : plage interquantile (en kHz)
- **Skew** : skewness (voir note dans la description de la spécification)
- **Sp.Ent** : entropie spectrale
- **Centroïde** : fréquence centroïde
- **Kurt** : kurtosis (voir note dans la description de la spécification)
- **Sfm** : planéité spectrale
- **Mode** : fréquence de mode
- **Peakf** : fréquence de crête (fréquence avec la plus haute énergie)
- **Meanfun** : moyenne de la fréquence fondamentale mesurée à travers le signal acoustique
- **Minfun** : fréquence fondamentale minimale mesurée à travers le signal acoustique

- **Maxfun** : fréquence fondamentale maximale mesurée à travers le signal acoustique

- **Meandom (Moyenne)** : moyenne de la fréquence dominante mesurée sur le signal acoustique

- **Mindom** : minimum de la fréquence dominante mesurée à travers le signal acoustique

- **Maxdom** : maximum de la fréquence dominante mesurée à travers le signal acoustique

- **Dfrange** : gamme de fréquences dominantes mesurée à travers le signal acoustique

- **Modindx** : indice de modulation. Calculé comme la différence absolue accumulée entre les mesures adjacentes des fréquences fondamentales divisée par la plage de fréquences

- **Label** : mâle ou femelle

On va travailler avec la classe **Label (Male/Female)** pour définir le genre dans notre programme

3.5. Environnement matériel et logiciel

3.5.1. Langage de programmation

Nous avons choisi **python** comme langage de programmation. Python est un langage de programmation interprété, polyvalent et facile à apprendre. Il se distingue par sa syntaxe claire et lisible, ce qui en fait un langage accessible même aux débutants. Python est largement utilisé dans différents domaines tels que le développement web, l'analyse de données, l'intelligence artificielle et l'automatisation des tâches. Il offre une vaste bibliothèque standard et bénéficie d'une communauté active qui contribue à son développement.

3.5.2. Environnement de programmation

- PyCharm est un environnement de développement intégré (IDE) spécialement conçu pour le langage de programmation Python. Il est développé par JetBrains et offre une large gamme de fonctionnalités pour faciliter le développement en Python.

- Spyder est un environnement de développement intégré (IDE) conçu spécialement pour les scientifiques et les analystes de données travaillant avec le langage de programmation Python. Il offre un ensemble d'outils et de fonctionnalités adaptés à l'analyse de données, à la visualisation et à la manipulation des données.

3.5.3. Configuration matérielle et logicielle

- Intel(R) Core(TM) i7-4712HQ CPU @ 2.30GHz 2.30 GHz
- Une mémoire vive d'une capacité de 16 GO.
- Système d'exploitation: Windows 10.
- Langage de programmation: Python (IDE : Pycharm, Spyder).

3.6. Réalisation

Ci-dessous des screenshots de notre programme ainsi de quelque explication

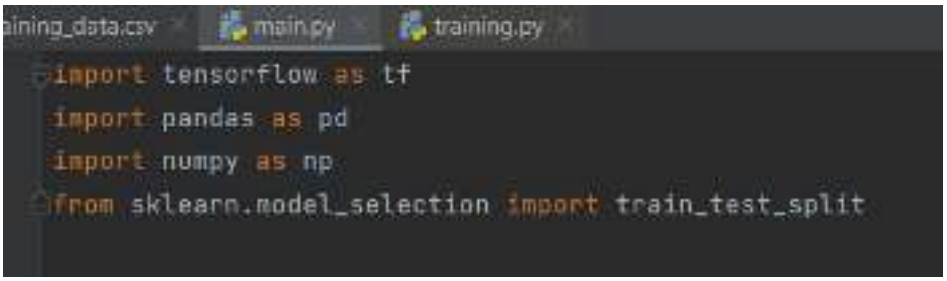

Figure 13 - Les bibliothèques nécessaires

Importer les bibliothèques necessaries

- **`import tensorflow as tf` :** Importe la bibliothèque TensorFlow, qui est une bibliothèque populaire pour l'apprentissage automatique et la construction de modèles de réseaux neuronaux.
- **`import pandas as pd` :** Importe la bibliothèque Pandas, qui fournit des structures de données et des outils pour l'analyse de données.
- **`import numpy as np` :** Importe la bibliothèque NumPy, qui fournit un support pour les tableaux et les opérations mathématiques en Python.
- **`fromsklearn.model_selection import train_test_split` :** Importe la fonction `train_test_split` de la bibliothèque scikit-learn, qui est utilisée pour diviser des ensembles de données en ensembles d'entraînement et de test.

```python
def import_and_clean():
    df = pd.read_csv('training_data.csv', header=0)
    df['label'] = df['label'].map({'female': 0, 'male': 1}).astype(int)
    return df
```

Figure 14–Import and clean

Définir la fonction `import_and_clean()`

- La fonction commence par lire un fichier CSV appelé **'training_data.csv'** et le stocke dans un objet DataFrame appelé **`df`** en utilisant la fonction **`pd.read_csv()`**.
- Ensuite, elle effectue une opération de nettoyage des données. Elle modifie la colonne appelée **'label'** dans **`df`** en utilisant la méthode **`map()`**. Elle fait correspondre les valeurs **'female'** à 0 et **'male'** à 1.
- Enfin, elle convertit les valeurs de la colonne 'label' en entiers en utilisant la méthode **`astype(int)`**.
- La fonction renvoie l'objet DataFrame `df` avec les données importées et nettoyées.

```python
def training_cnn(input_df):
    y = input_df['label'].copy()
    x = input_df.drop('label', axis=1).copy()
    x_train, x_test, y_train, y_test = train_test_split(x, y, train_size=0.8, random_state=42)


    # Using CNNs 2D
    model = tf.keras.models.Sequential()
    # pad the input with zeros to make sure that input vectors is of length 2
    # Reshape the data into a 5x4 image
    model.add(tf.keras.layers.Reshape((5, 4, 1), input_shape=(20,)))

    # Add convolutional layers
    model.add(tf.keras.layers.Conv2D(32, (3, 3), activation='relu'))

    model.add(tf.keras.layers.Conv2D(57, (3, 2), activation='relu'))

    model.add(tf.keras.layers.Conv2D(52, (2, 2), activation='relu', padding='same'))
```

Figure 15.1–Training CNN

Copier les étiquettes de la colonne 'label' :

- La première ligne **`y = input_df['label'].copy()`** crée une copie de la colonne 'label' du DataFrame d'entrée et l'assigne à la variable `y`. Cela permet de conserver les étiquettes séparément pour une utilisation ultérieure.

Copier les données d'entrée et les diviser en ensembles d'entraînement et de test :

- La ligne suivante **`x = input_df.drop('label', axis=1).copy()`** crée une copie du DataFrame d'entrée en excluant la colonne 'label' et l'assigne à la variable `x`. Cela permet de conserver les caractéristiques d'entrée séparément.
- Ensuite, la ligne **`x_train, x_test, y_train, y_test = train_test_split(x, y, train_size=0.8, random_state=42)`** divise les données d'entrée et les étiquettes en ensembles d'entraînement et de test à l'aide de la fonction **`train_test_split`** importée précédemment. Les données d'entrée sont divisées en **`x_train`** (ensemble d'entraînement) et **`x_test`** (ensemble de test), tandis que les étiquettes sont divisées en **`y_train`** et **`y_test`**. L'argument **`train_size=0.8`** spécifie que 80% des données seront utilisées pour l'entraînement et 20% seront utilisées pour les tests. L'argument **`random_state=42`** fixe la graine aléatoire pour assurer la reproductibilité des résultats.

Configuration d'un modèle de réseau neuronal convolutionnel (CNN) :

- La ligne **`model = tf.keras.models.Sequential()`** crée un modèle séquentiel à partir de Keras, qui est une pile linéaire de couches.
- La ligne suivante **`model.add(tf.keras.layers.Reshape((5, 4, 1), input_shape=(20,)))`**ajoute une couche de remodelage (Reshape) au modèle. Elle reformate les données d'entrée en une image de taille 5x4 avec une seule chaîne de profondeur. La taille d'entrée d'origine est spécifiée comme `(20,)`.
- Ensuite, trois couches de convolution (Conv2D) sont ajoutées au modèle à l'aide des lignes suivantes :

- `model.add(tf.keras.layers.Conv2D(32, (3, 3), activation='relu'))`

- `model.add(tf.keras.layers.Conv2D(32, (2, 2), activation='relu'))`

- `model.add(tf.keras.layers.Conv2D(32, (2, 2), activation='relu', padding='same'))`

Chaque couche de convolution a un certain nombre de filtres (32 dans ce cas) et une taille de noyau spécifiée (3x3, 2x2, 2x2 respectivement). La fonction d'activation 'relu' est utilisée pour introduire de la non-linéar**ité**. La dernière couche a également l'argument **`padding='same'`**

```
# Flatten the output from convolutional layers
model.add(tf.keras.layers.Flatten())

# Add a dense layer
model.add(tf.keras.layers.Dense(64, activation='relu'))

# Output layer
model.add(tf.keras.layers.Dense(1, activation='sigmoid'))

# Compile the model
model.compile(
    optimizer=tf.keras.optimizers.Adam(),
    loss=tf.keras.losses.BinaryCrossentropy(),
    metrics=[
        tf.keras.metrics.Accuracy(name='accuracy'),
        tf.keras.metrics.AUC(name='auc')
    ]
)
```

Figure 15.2–Training CNN

Aplatir la sortie des couches de convolution :

- La ligne **`model.add(tf.keras.layers.Flatten())`** ajoute une couche d'aplatissement (Flatten) au modèle. Cette couche transforme la sortie des couches de convolution en un vecteur unidimensionnel, préparant ainsi les données pour la couche dense suivante.

Ajouter une couche dense :

- La ligne `**model.add(tf.keras.layers.Dense(64, activation='relu'))**` ajoute une couche dense au modèle. Cette couche contient 64 neurones et utilise la fonction d'activation 'relu' pour introduire de la non-linéarité.

Couche de sortie :

- La ligne `**model.add(tf.keras.layers.Dense(1, activation='sigmoid'))**` ajoute une couche de sortie au modèle. Cette couche a un seul neurone et utilise la fonction d'activation 'sigmoid', ce qui est approprié pour un problème de classification binaire. Elle génère une probabilité de sortie entre 0 et 1, indiquant la probabilité d'appartenance à la classe positive.

Compiler le modèle :

- La ligne `**model.compile(...)**` compile le modèle en spécifiant l'optimiseur, la fonction de perte et les métriques à utiliser lors de l'entraînement.
- L'optimiseur utilisé ici est Adam, défini par `**optimizer=tf.keras.optimizers.Adam()`.**
- La fonction de perte utilisée est la BinaryCrossentropy, définie par **`loss=tf.keras.losses.BinaryCrossentropy()`.** Elle est couramment utilisée pour les problèmes de classification binaire.
- Deux métriques sont spécifiées pour le suivi de l'entraînement : l'exactitude (accuracy) et l'aire sous la courbe (AUC). Elles sont définies respectivement par `**tf.keras.metrics.Accuracy(name='accuracy')**` et `**tf.keras.metrics.AUC(name='auc')**`.

```
model.fit(
    x_train,
    y_train,
    validation_split=0.2,
    batch_size=32,
    epochs=25,
    callbacks=[
        tf.keras.callbacks.EarlyStopping(
            monitor='loss',
            patience=3,
            restore_best_weights=True
        )
    ]
)
return model
```

Figure 15.3–Training CNN

Entraîner le modèle :

- La ligne **`model.fit(...)`** entraîne le modèle en utilisant les données d'entraînement et les étiquettes correspondantes.
- **`x_train`** correspond aux données d'entraînement.
- **`y_train`** correspond aux étiquettes d'entraînement.
- L'argument **`validation_split=0.2`** spécifie que 20% des données d'entraînement seront utilisées comme ensemble de validation pour évaluer les performances du modèle pendant l'entraînement.
- L'argument **`batch_size=32`** indique que les données d'entraînement seront divisées en lots (batches) de taille 32 lors de l'entraînement.
- L'argument **`epochs=25`** définit le nombre d'itérations complètes à effectuer lors de l'entraînement.
- Le callback **`tf.keras.callbacks.EarlyStopping`** est utilisé pour arrêter l'entraînement prématurément si la perte (loss) ne s'améliore pas pendant 3 epochs consécutives.
- L'argument **`monitor='loss'`** spécifie que la perte est surveillée pour décider si l'entraînement doit être arrêté. L'argument **`patience=3`** indique le nombre d'epochs à attendre avant d'arrêter l'entraînement. L'argument **`restore_best_weights=True`** permet de restaurer les poids du modèle correspondant à la meilleure performance sur l'ensemble de validation.

Renvoyer le modèle entraîné :

- La ligne **`return model`** renvoie le modèle entraîné une fois que l'entraînement est terminé.

Figure 16 – main function

Prédiction d'exemple :

- Les lignes suivantes définissent un exemple de données `**x_sample**` et une étiquette correspondante `**y_sample**`. Les valeurs des caractéristiques `x_sample` sont fournies sous forme de liste numérique.

- Ensuite, la fonction `**import_and_clean()**` est appelée pour importer et nettoyer les données.

- La fonction `**training_cnn()**` est appelée en utilisant les données importées `Data` pour entraîner un modèle CNN.

- La prédiction est effectuée sur `**x_sample**` en utilisant le modèle entraîné `**tuned_cnn**`**.** La méthode `**predict()**` est utilisée pour obtenir la prédiction pour `**x_sample**`.

- En fonction de la valeur de la prédiction, une condition `**if**` est utilisée pour afficher le résultat correspondant.

Prédiction d'exemple supplémentaire :

- Le code pour une autre prédiction d'exemple peut être ajouté ici pour obtenir la prédiction correspondante à d'autres données.

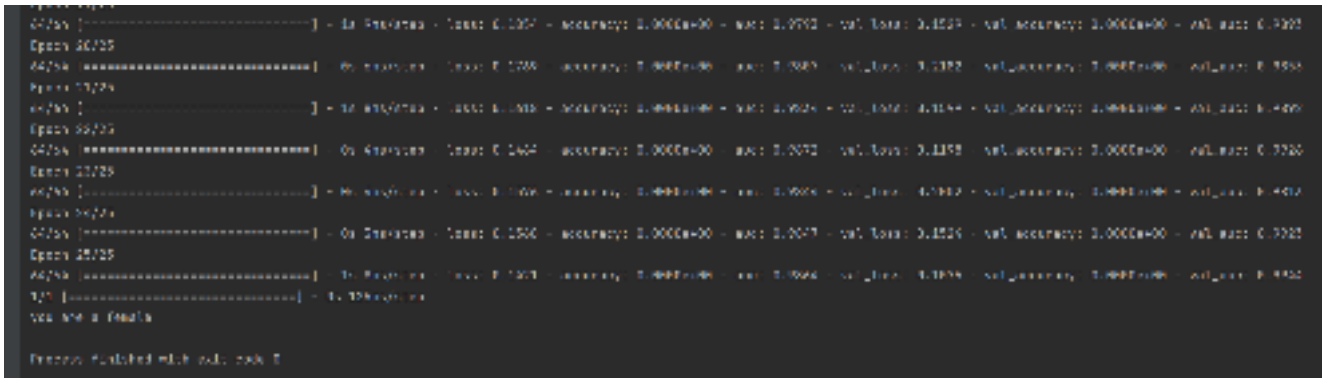

Figure 17 – Partie d'exécution

Les lignes suivantes représentent les résultats de l'entraînement du modèle CNN :

- Chaque ligne correspond à une époque de l'entraînement.
- Pour chaque époque, les informations suivantes sont affichées : la perte **(loss)**, l'exactitude **(accuracy)** et l'aire sous la courbe **(AUC)** pour l'ensemble d'entraînement **(train)** et l'ensemble de validation **(val)**.

- Le message **"Processfinishedwith exit code 0"** indique que le programme s'est terminé avec succès, sans erreur.

Ces résultats montrent que le modèle a été entraîné avec succès et qu'il est capable de prédire le genre sur de nouvelles données d'entrée.

3.7. Conclusion

Dans ce chapitre nous avons énuméré les différentes phases de développement et de mise en œuvre d'un programme, ce chapitre est le fruit de notre travail, nous avons mené une recherche bibliographique sur plusieurs mois, sélectionné la démarche à suivre et en fin les phases de mise en œuvre et les résultats.

Nous voulons un produit satisfaisant, encore imparfait, mais nous pouvons être sûrs que ce PFE nous permettra de mettre en pratique toutes nos connaissances en informatique.

Conclusion générale

Grâce au deeplearning, l'avenir de l'intelligence artificielle s'annonce prometteur. Les réseaux de neurones convolutifs sont l'un des domaines émergents de l'intelligence artificielle, machine learning, et l'apprentissage en profondeur. Il a diverses applications dans le monde actuel dans presque tous les secteurs. Compte tenu de son utilisation croissante, on s'attend à ce qu'il se développe davantage et soit plus utile pour résoudre les problèmes du monde réel.

Ce PFE traitant le domaine de la classification par genre à l'aide de la parole, nous avons conçu un programme qui classifie le genre à partir d'un fichier audio basée sur le deeplearning, motivé par la fusion de différentes sources d'informations pour améliorer la précision de la reconnaissance vocale, ce PFE se concentre sur l'exploitation des informations dans les sources vocales. Les paramètres de la source vocale sont généralement considérés comme moins discriminants mais difficiles à extraire. Cependant, les progrès de la technologie, du stockage et des ressources informatiques dans le domaine de la compréhension des phénomènes de production et de perception de la parole ont incité les chercheurs à reconsidérer ces biais et à tenter de maximiser l'utilisation de ces informations supplémentaires pour améliorer les performances du système, système de reconnaissance vocale. Par conséquent, le principal défi de la technologie de reconnaissance vocale est d'améliorer la robustesse du système dans des conditions incompatibles.

Notre système phonétique fournit principalement des indices acoustiques pour la classification des phonèmes et également la personnalité pour caractériser et reconnaître la parole. La parole humaine est considérée comme une émission sonore structurée, qui est essentiellement porteuse de communication. Par conséquent, les signaux vocaux transportent généralement un message à envoyer à une autre personne. Les variations de la nature du signal acoustique rendent très difficile le traitement des données brutes de celui-ci. En effet, ces données contiennent des informations complexes, souvent redondantes et mêlées de bruit. Nous avons fait des recherches bibliographiques sur le sujet, ce qui nous a amené à choisir les technologies et bibliothèques utilisées, comme base de notre implémentation, nous avons ensuite procédé à la conception, à la réalisation, et conclu aux tests.

Pour arriver à ces résultats, nous avons passé beaucoup de temps à lire et à rechercher des publications et des articles pour voir qu'elle était la meilleure classification de cela nous pouvons concevoir nos propres modèles. Au final, ce travail nous permettra de mettre en pratique nos connaissances sur les réseaux de neurones et d'acquérir des connaissances supplémentaires, et le temps passé à lire des articles peut être une bonne introduction à la recherche.

Bibliographie

Livre, monographie

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

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

Articles de revue

[1][2]Yann LeCun, « *L'apprentissage profond, une révolution en intelligence artificielle* », La lettre du Collège de France [En ligne], 41 | 2015-2016, mis en ligne le 01 novembre 2016, consulté le 28 décembre 2022. URL : http://journals.openedition.org/lettre-cdf/3227 ; DOI : https://doi.org/10.4000/lettre-cdf.32

[5]WilliamJHardcastleandAlainMarchal.*Speechproductionandspeechmodelling*,volume55.SpringerScience&BusinessMedia,2012.8

[6] Nacereddine Hammami. *Contribution to the automatic speech recognition of arabic language an its applications*. PhD thesis, University of Annaba, 2014.8,10

[7] Chetouani Mohamed. *Codage neuro-prédictif pour l'extraction de caractéristiques de signaux de parole*. PhDthesis, Université Pierre & Marie Curie,2004.10,24,33

[8] Asmaa Amehraye. *Débruitage perceptuel de la parole*. PhDthesis, Télécom Bretagne,2009.11,28,29

[9] Xuedong Huang, Alex Acero, and Hsiao-WuenHon.*Spoken language processing : guide to algorithms and system development*. Prentice Hall, 2001. 11,25

[10] René Boite. *Traitement de la parole*. PPUR presses polytechniques,2000.11

[11]Abdenour Hacine-Gharbi. *Sélection de paramètres acoustiques pertinents pour la reconnaissance de la parole*. PhD thesis, Université d'Orléans, France et Université Ferhat Abbas-Sétif, Algérie,2012.12,17

[12] George.A Millerand Joseph CRLicklider. *Theint eligibility of interrupted speech*. The Journal of the Acoustical Society of America, 22(2) :167–173, 1950.12

Article d'un ouvrage collectif

[18] Rashid Jahangir, Ying Wah Teh, Faiqa Hanif & Ghulam Mujtaba. Deep *learning approaches for speech emotion recognition: state of the art and research challenges*, Published in 2 January 2021

[20] Fukushima, K. (1988). Neocognitron: *A hierarchical neural network capable of visual pattern recognition*. Neural Networks, 1, 119–130.

[21] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). *Gradient-based learning applied to document recognition*. Proceedings of the IEEE, 86, 2278–2324.

[22] Hubel, D. H., & Wiesel, T. N. (1968). *Receptive fields and functional architecture of monkey striate cortex.* The Journal of Physiology., 195(1), 215–243

Documents web

[13] https://www.techno-science.net/glossaire-definition/Reconnaissance-vocale.html

[14] http://deptinfo.unice.fr/twiki/pub/Linfo/PlanningDesSoutenances20032004/Benguigui-Ismais-Hamdan.pdf

[15] https://www.ibm.com/fr-fr/cloud/learn/speech-recognition

[16] https://www.futura-sciences.com/tech/definitions/intelligence-artificielle-deep-learning-17262/

[17] https://ryax.tech/fr/deep-learning-comprendre-les-reseaux-de-neurones-artificiels-artificial-neural-networks/

[19] https://www.analyticsvidhya.com/

[23] https://www.kaggle.com/datasets/sripaadsrinivasan/audio-mnis