RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE Dr. TAHAR MOULAY SAIDA

FACULTE : TECHNOLOGIE

DEPARTEMENT : INFORMATIQUE

MÉMOIRE DE MASTER

Option : Sécurité Informatique et Cryptographie

Reconnaissance vocale par les techniques de deep learning : détection des coups de feu dans un bruit d'une ville.

Présenté par :

Bouazza Mustapha

Benhaddad Fatiha

Encadré par :

D. Mohamed El Hadi Rahmani

Année Universitaire 2020-2021

Voice recognition by deep learning techniques : detection of gunshots in the noise of a city.

Abstract —

Audio signals are all around us. As such, there is an increasing interest in audio classification for various scenarios, from fire alarm detection for hearing impaired people, through engine sound analysis for maintenance purposes, to baby monitoring. Though audio signals are temporal in nature, in many cases it is possible to leverage recent advancements in the field of image classification and use popular high performing convolutional neural networks for audio classification. In this blog post we will demonstrate such an example by using the popular method of converting the audio signal into the frequency domain.

In order to improve the accuracy of speech recognition by various sources of information, this memo focuses on the exploitation of information about a specific speech source. Information in audio is generally less distinct but more difficult to extract but, with the development of technology, storage and resources And studies in the field of understanding the phenomenon of sound production and perception have led researchers to reconsider these prejudices and attempt to obtain most of this additional information to improve the performance of the speech recognition system. The main challenge for speech recognition technology was to improve the robustness of systems under incompatible conditions. Our system provides acoustic indicators as a basis for the classification of sounds, as well as for the discrimination of sounds. The rapid global development and growth of telecommunications, both in terms of size and diversity (physical travel, financial transactions, access to services, etc.) implies the need to verify the identity of individuals. The importance of these issues motivates fraudsters to overcome existing security systems. The voice conveys different information, the word Human seen as a set of structured sounds, is fundamentally a means of communication. As such, a voice signal is usually the carrier of the message to another person. A change in the nature of the sound signal makes it very difficult to process the raw data of the latter. This is because this data contains complex information, and is often redundant and mixed with noise.

Keywords :voice recognition, single audio, audio classification, deep learning, Recurrent Neural Networks, Convolutional Neural Network.

Reconnaissance vocale par les techniques de deep learning :détection des coup de feu dans un bruit d'une ville.

Résumé —

Les signaux audio sont partout autour de nous. En tant que tel, il existe un intérêt croissant pour la classification audio pour divers scénarios, de la détection d'alarme incendie pour les personnes malentendantes à l'analyse du son du moteur à des fins de maintenance, en passant par la surveillance des bébés. Bien que les signaux audio soient de nature temporelle, dans de nombreux cas, il est possible de tirer parti des avancées récentes dans le domaine de la classification des images et de l'utilisation des réseaux de neurones convolutifs à haute performance pour la classification audio. Dans cet article de blog, nous allons démontrer un tel exemple en utilisant la méthode populaire de conversion du signal audio dans le domaine fréquentiel.

Dans le but d'améliorer la précision de la reconnaissance vocale par diverses sources d'informations, cette note se concentre sur l'exploitation des informations sur une source vocale spécifique. L'information en audio est généralement moins distincte mais plus difficile à extraire mais, avec le développement de la technologie, du stockage et des ressources Et des études dans le domaine de la compréhension du phénomène de production et de perception du son, ont conduit les chercheurs à reconsidérer ces préjugés et à tenter d'obtenir la plupart de ces informations supplémentaires pour améliorer les performances du système de reconnaissance vocale. Le principal défi de la technologie de reconnaissance vocale était d'améliorer la robustesse des systèmes dans des conditions incompatibles. Notre système fournit des indicateurs acoustiques comme base pour la classification des sons, ainsi que pour la discrimination des sons. Le développement mondial rapide et la croissance des télécommunications, tant en termes de taille que de diversité (déplacements physiques, transactions financières, accès aux services...) implique la nécessité de vérifier l'identité des individus. L'importance de ces enjeux, motive les fraudeurs à vaincre les systèmes de sécurité existants. La voix véhicule des informations différentes, la parole L'humain vu comme un ensemble de sons structurés, est fondamentalement un moyen de communication. En tant que tel, un signal vocal est généralement le porteur du message à une autre personne. Un changement dans la nature de le signal sonore rend le traitement des données brutes de ces derniers très difficile. En effet, ces données contiennent des informations complexes, et sont souvent redondant et mélangé avec du bruit

Mots clés : reconnaissance vocal ,single audio, classification audio , l'apprentissage profond ,Réseaux de neurones récurrents, Convolutional Neural Network .

التعرف على الصوت من خلال تقنيات التعلم العميق: الكشف عن الطلقات النارية في ضوضاء المدينة

الملخص ---

االإشارات الصوتية في كل مكان حولنا. على هذا النحو ، هناك اهتمام متزايد بتصنيف الصوت لمختلف السيناريوهات ، من اكتشاف إنذار الحريق للأشخاص ضعاف السمع ، من خلال تحليل صوت المحرك لأغراض الصيانة ، إلى مراقبة الأطفال. على الرغم من أن الإشارات الصوتية مؤقتة بطبيعتها ، فمن الممكن في كثير من الحالات الاستفادة من التطورات الحديثة في مجال تصنيف الصور واستخدام الشبكات العصبية التلافيفية ذات الأداء العالي لتصنيف الصوت. في منشور المدونة هذا سوف نوضح مثل هذا المثال باستخدام الطريقة الشائعة لتحويل الإشارة الصوتية إلى مجال التردد.

من أجل تحسين دقة التعرف على الكلام من خلال مصادر المعلومات المختلفة ، تركز هذه المذكرة على استغلال المعلومات حول مصدر كلام معين. المعلومات في الصوت أقل تميزًا بشكل عام ولكن استخراجها أكثر صعوبة ولكن مع تطور التكنولوجيا والتخزين والموارد وقد دفعت الدراسات في مجال فهم ظاهرة إنتاج الصوت والإدراك الباحثين إلى إعادة النظر في هذه التحيزات ومحاولة الحصول على معظم هذه المعلومات الإضافية لتحسين أداء نظام التعرف على الكلام. كان التحدي الرئيسي لتكنولوجيا التعرف على الكلام هو تحسين متانة الأنظمة في ظل ظروف غير متوافقة. يوفر نظامنا مؤشرات صوتية كأساس لتصنيف الأصوات ، وكذلك لتمييز الأصوات. إن التطور العالمي السريع ونمو الاتصالات السلكية واللاسلكية ، من حيث الحجم والتنوع (السفر المادي ، والمعاملات المالية ، والوصول إلى الخدمات ، وما إلى ذلك) يعني ضمناً الحاجة إلى التحقق من هوية الأفراد. تحفز أهمية هذه المشكلات المحتالين على التغلب على أنظمة الأمان الحالية. ينقل الصوت معلومات مختلفة ، فالكلمة التي ينظر إليها الإنسان على أنها مجموعة من الأصوات المنظمة ، هي في الأساس وسيلة اتصال. على هذا النحو ، عادة ما تكون الإشارة الصوتية هي الناقل للرسالة إلى شخص آخر. يؤدي التغيير في طبيعة الإشارة الصوتية إلى صعوبة معالجة البيانات الأولية للأخيرة. هذا لأن هذه البيانات تحتوي على معلومات معقدة ، وغالبًا ما تكون زائدة عن الحاجة ومختلطة بالضوضاء
.

الكلمات المفتاحية :التعرف على الصوت ،التصنيف الصوتي ، التعلم العميق ،الشبكيات العصبية التكررة ،االشبكة العصبية التلافقية :

Remerciement

Avant toute chose, nous tenons à remercier Allah pour cette grâce d'être en vie et en bonne santé, et pour avoir terminé ce travail dans les meilleures conditions et ce malgré toutes les contraintes et les obstacles que nous avons rencontré.

Il est souvent difficile de remercier les gens qui vous aident à accomplir les tâches qui vous sont données, et pourtant nous nous devons exprimer l'entière gratitude que nous ressentons envers eux.

Nous tenons donc à présenter un remerciement bien distingué à notre encadrant D.Mohamed Elhadi Rahmani pour son soutien, son aide, et ses conseils qui nous ont guidés durant l'élaboration de ce travail.

Nous portons toutes nos gratitudes aux enseignants de l'université de Moulay Tahar, pour leurs dévouements et leurs Assistantes tout au long de notre études universitaires.

Nous tenons aussi à exprimer nos gratitudes reconnaissance à tous les membres de jury d'avoir accepté et d'évaluer notre travail.

Enfin on remercie toutes personnes qui ont contribuées de près ou de loin à la réalisation de ce travail, ainsi qu'au bon déroulement du stage, et dont les noms ne figurent pas dans ce document

DEDICACE

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

À mon cher oncle **« Mohamed »,** qui était comme un père pour moi, tu as toujours été à mes côtés pour me soutenir et m'encourager, Que ce travail traduit ma gratitude et mon affection, et sa belle femme **« beddiar .R ».**

À la femme qui a souffert sans me laisser souffrir, qui n'a jamais dit non âmes exigences et qui n'a épargné aucun effort pour me rendre heureuse: mon adorable mère **« Fatima ».**

À l'homme, mon précieux offre du dieu, qui doit ma vie, ma réussite et tout mon respect : mon cher père **« Mohamed ».**

À ma cher frère **« Noureddine »** qui n'ont pas cessé de me conseiller, encourager et soutenir tout au long de mes études. Que Dieu le protège et l'offre la chance et le bonheur.

À ma grand-mère. Dieu lui a donné Une vie longue et heureuse.

A mon binôme **« Bouazza Mustapha »** pour son soutien moral, sa patience et sa compréhension tout au long de ce projet

Benhaddad fatiha

DEDICACE

À mon cher père **« Mohamed »**

Qui n'ont jamais cessé de formuler des prières à mon égard de me soutenir et de m'épauler pour que je puisse atteindre mes objectifs.

À ma chère mère **« Youssra »**

Quoi que je fasse ou que je dise, je ne saurai point te remercier comme il se doit. Ta bienveillance me guide et ta présence à mes côtés a toujours été ma source de force pour affronter les différents obstacles.

À mes belles sœurs « **Soumia » « khadidja » « Nedjat »** et leurs enfants
« Mohamed » « Retadj » « Oussama » « Fatima »

À tous mes oncles, cousins. Merci pour leur amour et leurs encouragements.

À mes amis **« Isalm » « Bouazza » « Touati » « Djamal » « Bachir »**

À mon binôme **« Benhaddad fatiha »** pour son soutien moral, sa patience et sa compréhension tout au long de ce projet

Bouazza mustapha

TABLE DES MATIÈRES

Table des matières **11**

Page

Table des figures **14**

1 Introduction générale **1**

2 Traitement de signal auditif **7**

2.1 Introduction 7
2.2 Le signal 8
2.3 Le son 9
2.4 signal audio 9
2.5 Les modes de transmission d'un signal 9
2.6 Numérisation du signal : 9
2.7 Conversion analogique-numérique (CAN) : 10
2.8 Les modes de représntation du son 10
2.8.1 Représentation temporelle 11
2.8.2 Représentation fréquentielle (ou spectrale) 11
2.8.3 Représentation tridimensionnelle : le sonagramme ou spectrogramme . . . 12
2.9 Analyse du signal de parole 13
2.10 Le traitement du signal à la prise etla restitution du son 13
2.11 Que pouvons-nous attendre de la Reconnaissance vocale ? 13
2.12 La reconnaissance vocale définition 14
2.13 Domaine d'application de la reconnaissance vocale 14
2.14 Reconnaissance vocale et la biométrie 15
2.15 La reconnaissance automatique du locuteur (RAL) 16
2.16 Différentes Tâches en RAL 16
2.16.1 Identification automatique du locuteur (IAL) 16
2.16.2 La Vérification Automatique du Locuteur (VAL) 17
2.17 Comment ça marche ? 17
2.18 État de l'art 18

2.18.1 Origin-STT 18
2.18.2 Amazon Transcribe Medical 19
2.18.3 GRIFOS 19
2.19 Conclusion 20

3 deep learning **21**
3.1 Introduction 21
3.2 Définition de l'apprentissage profond (deep learning) 22
3.3 Les applications du Deep Learning [40] 22
3.3.1 La reconnaissance faciale 22
3.3.2 La détection d'objets 23
3.3.3 Le Natural Language Processing 23
3.4 Domaines d'application de l'apprentissage profonde : [45] 23
3.4.1 Conduite automatisée : 23
3.4.2 Recherche médicale : 23
3.4.3 Électronique 23
3.5 Histoire de Deep Learning [21] 24
3.6 Forces et faiblesses 24
3.6.1 Les points forts de l'apprentissage en profondeur 24
3.6.2 Les points faibles de l'apprentissage profond 24
3.7 Comment ça marche ? 25
3.8 Réseaux de neurones 25
3.9 Principe de fonctionnement 26
3.10 L'apprentissage 27
3.10.1 Apprentissage supervisé 28
3.10.2 Apprentissage non supervisé 28
3.10.3 Apprentissage semi-supervisé 29
3.11 L'écueil du surapprentissage 29
3.12 Réseaux de neurones récurrents 30
3.13 Convolutional Neural Network 30
3.14 les différentes couches d'un CNN 30
3.14.1 La couche de convolution 31
3.14.2 La couche de pooling 31
3.14.3 La couche de correction ReLU 31
3.14.4 La couche fully-connected 32
3.15 RNN et LSTM 32
3.16 RNN VS CNN VS ANN 33
3.17 Conclusion 33

4 La Réalisation **35**

4.1 Introduction 35

4.2 Les outils de réalisation 35

4.2.1 Python 35

4.2.2 Jupyter Notebook 36

4.3 Base de données 37

4.4 Présentation du fichier audio 38

4.5 Exploration des données 38

4.6 Pré-traitement des données 41

4.7 Extraire des fonctionnalités 42

4.8 Conversion des données et des étiquettes, puis fractionnement de l'ensemble de données 44

4.9 Construire notre modèle 45

4.10 Résultats 48

4.11 Conclusion 49

Bibliographie **51**

*

TABLE DES FIGURES

FIGURE **Page**

2.1 Signal analogique et signal numérique. 10
2.2 Dispositif d'enregistrement numérique d'un son. 10
2.3 (signal analogique , signal échantillonné , puis quantifié. 11
2.4 Représentation temporelle. 11
2.5 Représentation fréquentielle. 12
2.6 Représentation tridimensionnelle. 12
2.7 Principe de base de l'identification du locuteur 16
2.8 Principe de base de Automatique du locuteur 17

3.1 la relation entre l'IA,ML,Deep learning. 22
3.2 Topologie de réseau de neurones avec une seule sortie. 26
3.3 Topologie de réseau de neurones profond. 27
3.4 allure de la foction ReLU 32

4.1 logo Python 36
4.2 logo jupyter 37
4.3 Présentation du fichier audio. 38
4.4 les formes des sons répétitifs 39
4.5 code de l'extraction des propriétés de chacun des fichiers audio. 39
4.6 code de l'extraction des propriétés de chacun des fichiers audio. 40
4.7 Canaux audio. 40
4.8 Fréquence d'échantillonnage. 41
4.9 Profondeur de bits. 41
4.10 comparison entre trois représentations visuelles différentes d'une onde sonore. 43
4.11 code pour extrairons un MFCC. 43
4.12 code pour extrairons un MFCC. 44
4.13 Conversion des données et des étiquettes puis fractionnement de l'ensemble de données. 45
4.14 Construire notre modèle 46
4.15 Construire notre modèle 46
4.16 compiler notre modèle 47

4.17 entraîner le modèle 47
4.18 examinera la précision du modèle 48
4.19 Nombre de paramètres par couche 48

CHAPITRE 1

INTRODUCTION GÉNÉRALE

Avec l'explosion récente de l'utilisation de l'informatique, et ce dans tous les domaines de la société, la sécurité des données et des équipements est devenu un enjeu majeur. Que nous soyons professionnels de l'informatique, simple utilisateur ou utilisateur avancé, nous nous servons d'outils informatiques quasi quotidiennement. Qu'il s'agisse de stocker des données personnelles, de consulter ou d'administrer des sites internet, de travailler, d'échanger des données ou de se divertir, nous sommes absolument dépendants du bon fonctionnement et de la fiabilité de l'informatique que nous utilisons. Même quelqu'un n'ayant jamais utilisé un ordinateur devient un utilisateur de l'informatique lorsqu'il retire de l'argent au guichet automatique de sa banque, lorsqu'il prend l'ascenseur, qu'il consulte les horaires de trains à la gare ou qu'il utilise sa voiture. En effet aujourd'hui l'informatique se cache dans énormément de choses courantes et nous ne nous en rendons très peu compte. Mais l'informatique est vulnérable, surtout si l'on ne prend pas les précautions nécessaires et que l'on n'est pas au courant des menaces qui pèsent sur elle

[27]

Contexte de la thèse

L'identification de la voix est considérée par les utilisateurs comme une des formes les plus normales de la technologie biométrique, car elle n'est pas intrusive et n'exige aucun contact physique avec le lecteur du système.

[32]

Les sons sont partout autour de nous. Que ce soit directement ou indirectement, nous sommes toujours en contact avec des données audio. Les sons définissent le contexte de nos activités

quotidiennes, depuis les conversations que nous avons lorsque nous interagissons avec des personnes, la musique que nous écoutons et tous les autres sons de l'environnement que nous entendons quotidiennement, tels que le bruit d'une voiture ou d'un camion. que nous entendons quotidiennement, comme le passage d'une voiture, le bruit de la pluie ou tout autre type de bruit de fond. bruit de fond. Le cerveau humain traite et comprend en permanence ces données audio. [34]

Les imitateurs essayent habituellement de reproduire les caractéristiques vocales qui sont les plus évidentes au système auditif humain et ne recréent pas les caractéristiques moins accessibles qu'un système automatisé d'identification de voix analyse. Il n'est donc pas possible d'imiter la voix d'une personne inscrite dans la base de données.

La variabilité d'une personne à une autre démontre les différences du signal de parole en fonction du locuteur. Cette variabilité, utile pour différencier les locuteurs, est également mélangée à d'autres types de variabilité - variabilité due au contenu linguistique, variabilité intra-locuteur (qui fait que la voix dépend aussi de l'état physique et émotionnel d'un individu), variabilité due aux conditions d'enregistrement du signal de parole (bruit ambiant, microphone utilisé, lignes de transmission) - qui peuvent rendre l'identification du locuteur plus difficile.

Malgré toutes ces difficultés apparentes, la voix reste un moyen biométrique intéressant à exploiter
La technologie d'analyse de la voixs'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.

[41]

Procédure d'authentification

Une fois le modèle créé, le test d'authentification mesure la ressemblance d'un enregistrement de la voix avec toutes les signatures connues par le système. Le résultat du test est un score de vraisemblance proportionnel à la ressemblance entre l'enregistrement et le modèle testé. Si la personne est déjà connue du système, on peut alors lui attribuer l'identité du modèle qui obtient le meilleur score. Si elle n'est pas connue du système, on mesure alors simplement la ressemblance de sa voix avec les voix du système.

[11]

Motivation

La reconnaissance vocale et le contrôle vocal via la biométrie vocale vont devenir complémentaires pour sécuriser les opérations des utilisateurs [17]
Les points positifs de la biométrie pour la sécurité Si la biométrie a tant de succès, il y a une raison : elle est très difficile à falsifier. L'authentification a évolué. Elle a commencé par un nom d'utilisateur et un mot de passe, par exemple. Mais il est facile de s'en emparer ou de tromper les personnes pour les amener à divulguer les informations qu'elles seules connaissent. Les techniques d'authentification se sont déplacées vers des objets que l'on a tous .
Mais ce n'est pas assez. Les cyber-escrocs peuvent toujours se procurer ou falsifier les appareils des utilisateurs. La prochaine étape de l'authentification est l'utilisation des caractéristiques uniques de l'utilisateur, révélées grâce à la biométrie. Car il est bien plus difficile de falsifier une voix, des empreintes digitales, l'iris, etc

[53]

Problématique

Vivant dans un monde entouré de différentes formes de sons provenant de différentes sources, notre cerveau et notre système auditif identifient constamment chaque son qu'il entend, à sa manière.la reconnaissance vocale ou du son est un domaine de recherche majeur depuis de nombreuses années et il existe de nombreuses méthodes éprouvées avec différents modèles et fonctionnalités qui se sont avérées utiles et précises.la reconnaissance vocale peut aller de domaines tels que le multimédia, la surveillance bioacoustique, la détection d'intrus dans les zones fauniques à la surveillance audio et aux sons environnementaux. [16]

existe également plusieurs problèmes dans ce domaine, dont les plus importants sont :

La parole

La parole est le résultat de l'air faisant vibrer les cordes vocales et passant dans le conduit vocal constitué par la bouche et le nez. Si ces éléments anatomiques influencent la personnalité d'une voix, ils n'en fixent pas pour autant toutes les caractéristiques. Ainsi, une même personne ne parle pas tout le temps de la même façon. La voix change avec l'âge, l'humeur ou encore un rhume. En jargon scientifique, les variations de la voix d'une même personne sont appelées variabilité intra-locuteur. En raison de ces aspects comportementaux, on parle de signature vocale, plutôt que d'empreinte.
Outre la variabilité de la voix d'une même personne, une autre difficulté pour que l'ordinateur puisse reconnaître une voix vient du fait que les conditions et la qualité d'enregistrement d'une

même voix peuvent être très différentes.

Une voix passant à travers un microphone, transmise par exemple par radio ou téléphone portable, subit des déformations. C'est le problème de la variabilité du canal. Un environnement calme ou bruité rend aussi plus ou moins facile la détection de la voix. Cette variabilité due au bruit environnant est difficilement prévisible et nécessite des traitements spécifiques pour être neutralisée

[10]

Variabilité due au locuteur

Une dégradation croissante des performances a été observée au fur et à mesure que le temps qui sépare la session d'apprentissage de la session de test augmente . De plus, le comportement des locuteurs se modifie lorsque ceux-ci s'habituent au système. Les modèles des locuteurs doivent donc être régulièrement mis à jour avec les nouvelles données d'exploitation du système Les altérations de la voix dues à l'état physique (fatigue, rhume) ou émotionnel (stress) mettent aussi en échec l'efficacité des systèmes. [52]

Objectif

Notre objectif est d'améliorer l'identification du vocale en environnement réel
Ce domaine en plein essor peut grandement bénéficier de la riche expérience et des divers outils développés pour les tâches de vision par ordinateur
Et d'autres objectifs, notamment :

Les Technologies vocales pour les personnes âgée

[33]
Face à l'accroissement démographique important de la population âgée dans les années à venir, la restructuration de l'environnement adapté aux seniors est au cœur des projets. Le maintien à domicile grâce à des professionnels et des moyens adaptés semble être une alternative intéressante face au manque de disponibilité et au coût des infrastructures spécialisées.

La perte d'autonomie se présente suivant des degrés variés, il est donc possible d'intervenir jusqu"a un certain stade en assistant la personne dans son quotidien. De nombreux projets ont émergeé pour répondre à cette demande. En effet, d'une part du point de vue social, on constate depuis les années 90 une forte hausse des services d'aide à la personne (âgée ou non) à domicile par des infirmières ou des aides ménagères. Par ailleurs, du point de vue technologique, on remarque que les TIC sont de plus en plus appliqués pour faciliter le quotidien des personnes fragiles. Détournées de leur but commercial premier, de nombreuses technologies sont aujourd'hui utilisées

et adaptées à la population handicapée et âgée de plus en plus sollicitant à l'égard des technologies. Trés vite, la manipulation classique des différentes technologies (ordinateurs, téléphones, internet, systèmes domotiques) peut cependant s'avérer complexe pour des personnes non initiées ou handicapées. Bien que les performances ne soient pas encore tout à fait satisfaisantes, la RAP peut se montrer trées pertinente par son aspect naturel et intuitif pour l'utilisation de ces outils .

Réinitialisation de mots de passes

Un dispositif de reconnaissance vocale peut servir à sécuriser ou faciliter la gestion et la réinitialisation des mots de passe utilisés pour accéder au système d'information d'une société.

Ce procédé permet alors de générer et de réinitialiser automatiquement des mots de passe et repose sur la reconnaissance du gabarit de l'empreinte de la voix des employés : la voix est numérisée puis segmentée par unités échantillonnées.
[42]

Sécurisation des transactions à risque par carte de crédit

La biométrie vocale constitue aussi une solution sûre et pratique pour vérifier les transactions à risque par carte de crédit (par exemple celles en dehors des habitudes de consommation du client ou de son emplacement géographique habituel). Quand une opération à risque est détectée, une demande de vérification de la transaction peut être envoyée au titulaire de la carte de crédit, via un appel sortant automatique, sur son téléphone portable. Le détenteur est alors invité à prononce une phrase clé : "J'autorise cette transaction par ma signature vocale." A l'inverse, si la transaction est suspecte, il peut tout aussi facilement rejeter celle-ci, ce qui permet alors à l'institution financière d'investiguer sur les transactions marquées comme suspectes.
[29]

Paiement en ligne

La biométrie vocale consiste à utiliser la voix comme mot de passe pour accéder à un compte ou pour sécuriser un paiement par internet. Au niveau mondial, les banques se sont lancées avec engouement dans la biométrie vocale . Cette technologie est stratégique pour les banques qui souhaitent rester innovantes et ne pas laisser aux GAFA la gestion unique de l'identité client. [18]

CHAPITRE 2

TRAITEMENT DE SIGNAL AUDITIF

2.1 Introduction

Le numérique a envahi notre existence, et l'univers du son ne fait pas exception. Peu importe que le projet soit modeste ou des plus ambitieux, on travaille aujourd'hui majoritairement sur des flux de données numériques [38]

La théorie du signal fournit la description mathématique (ou modélisation) des signaux. Le traitement des signaux est la discipline technique qui, s'appuyant sur la théorie du signal et de l'information, les ressources de l'électronique, de l'informatique et de la physique appliquée, a pour objet l'élaboration ou l'interprétation des signaux porteurs d'information. Elle trouve son application dans tous les domaines concernés par la perception, la transmission ou l'exploitation de ces informations. [14]

Le traitement du signal audio est au cœur de l'enregistrement, de l'amélioration, du stockage et de la transmission de contenu audio. Le traitement du signal audio est utilisé pour convertir entre les formats analogiques et numériques, pour réduire ou augmenter les plages de fréquences sélectionnées, pour supprimer les bruits indésirables, pour ajouter des effets et pour obtenir de nombreux autres résultats souhaités. Aujourd'hui, ce processus peut être effectué sur un PC ou un ordinateur portable ordinaire, ainsi que sur un équipement d'enregistrement spécialisé.

[31]

2.2 Le signal

Le signal est une grandeur physique, dotée d'une unité et donc mesurable ; l'information est un message.

Pour qu'un signal soit porteur d'une information, il est nécessaire d'établir une convention. Par exemple, une tension électrique en Volt peut représenter la présence ou l'absence d'un objet. Le message « un objet est présent » est associé à une valeur de tension spécifiée au préalable de façon conventionnelle.
Pour qu'il y ait communication, trois éléments sont indispensables :

- un émetteur, qui délivre un signal porteur de l'information.
- un récepteur qui reçoit le signal et décode l'information que ce signal contient.
- une transmission du signal.

Toute grandeur physique peut devenir un signal dès lors que l'on associe à sa valeur un message. La grandeur physique peut être l'amplitude ou la fréquence d'une onde électromagnétique (lumière visible, infrarouge, radio. . .) ou d'une onde acoustique, une différence de potentiel (tension électrique), une intensité d'un courant, une concentration, etc. Dans le monde du vivant, par exemple, le taux de glucose dans le sang (glycémie) est un signal qui envoie au cerveau une information associée à la faim.

Dans la vie courante, un signal sonore dans une cour d'école est associé à l'information « récréation terminée », une sirène d'alarme domestique adresse une information d'intrusion etc. Des récepteurs sensoriels et des capteurs transforment un signal chimique, lumineux ou sonore en un signal électrique transmis au cerveau. Le cerveau est capable de décoder l'information portée par ce signal électrique. Dans la vie quotidienne, de très nombreuses informations sont échangées. Exemples :

- sécurité maritime : les signaux transmettent des informations liées à la sécurité de la navigation ou à des situations particulières de navires (non manœuvrant, panne machine, maladie contagieuse à bord). Ces signaux sont émis par des phares, sémaphores, pavillons, balises, sirènes, panneaux, feux, etc. ;

- sécurité routière et ferroviaire : les signaux transmettent des informations liées à la sécurité de la circulation, comme les limitations de vitesse ou l'obligation d'arrêt (panneau stop, feux tricolores, etc.)

[23]

2.3 Le son

Le son est une vibration de l'air qui se propage avec des caractéristiques variables d'intensité, de fréquence, de portée, d'écho, ... L'oreille humaine est sensible aux sons dans certaines limites d'intensité et de fréquence, c'est le processus de l'audition. Quand les cordes vocales créent des sons, c'est la voix et le processus de la phonation. [49]

2.4 signal audio

Un signal habituellement est une fonction du temps créée par un capteur pour mesurer une grandeur physique. Le signal audio est un cas particulier de signal qui traduit la mesure d'un son. Présenté à l'entrée du CAN, ce signal issu du micro est en réalité une tension électrique qui reproduit les vibrations de l'air. Cette tension est proportionnelle à tout instant à la pression de l'air mesure donc l'intensité instantanée du son. On la représente aisément dans un chronogramme. [49]

2.5 Les modes de transmission d'un signal

Toute grandeur physique peut supporter un signal et les modes de transmission sont propres à leur nature. Ces modes de transmission sont donc très nombreux. Les grandeurs électriques sont souvent utilisées comme signaux transportant les informations. Les matériaux conducteurs sont alors utilisés (câbles, pistes en cuivre sur les cartes électroniques). Les ondes électromagnétiques se propagent par variation locale des champs électriques et magnétiques ; elles peuvent se propager avec ou sans support matériel, dans le vide, l'atmosphère ou un milieu matériel ne perturbant pas la propagation (par exemple une paroi en bois ou en brique). Dans le cas des ondes acoustiques, un milieu matériel est nécessaire à la propagation du signal (air, eau, matériau solide). Ci contre : exemple de capteur de présence utilisant une onde électromagnétique,ici infra-rouge. [20]

2.6 Numérisation du signal :

[35]

Un signal analogique est un signal continu qui peut prendre une infinité de valeurs, alors que le signal numérique est un signal discret (discontinu), qui se résume en une succession de « 0 » et de « 1 ».

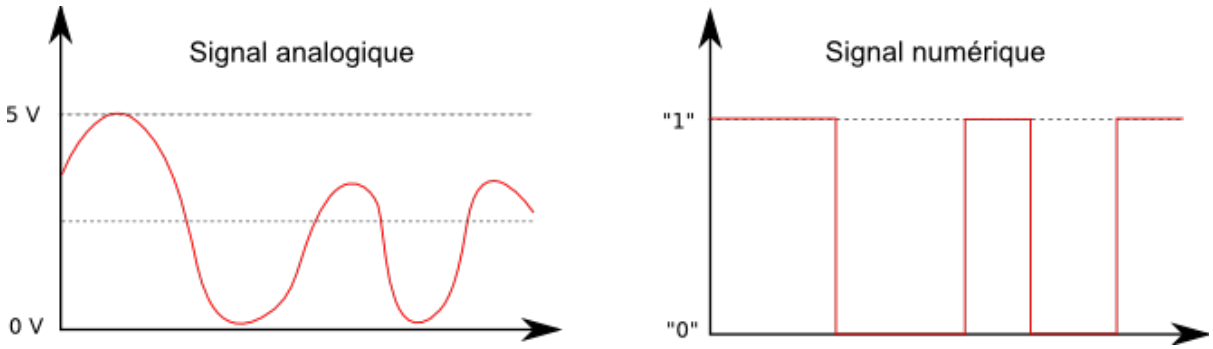

FIGURE 2.1: Signal analogique et signal numérique.

L'objectif de la numérisation est de transformer le 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 consiste en 2 étapes successives : l'échantillonnage et la conversion analogique-numérique (CAN).

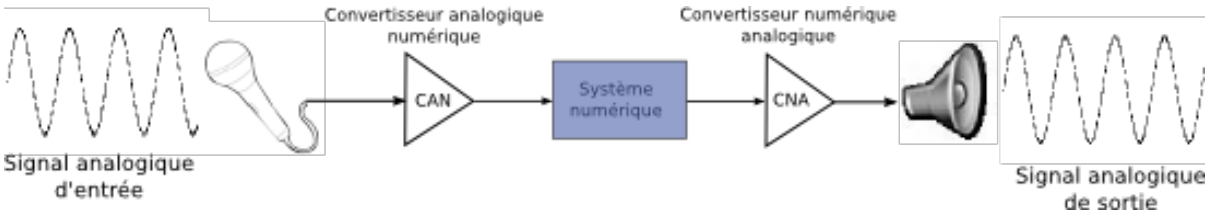

FIGURE 2.2: Dispositif d'enregistrement numérique d'un son.

2.7 Conversion analogique-numérique (CAN) :

[22]
Un convertisseur analogique – numérique (CAN) est un dispositif électronique permettant la conversion d'un signal analogique en un signal numérique. Conceptuellement, la conversion analogique – numérique peut être divisée en trois étapes : l'échantillonnage temporel (on prélève la valeur du signal à une fréquence définie), la quantification (on affecte une valeur numérique à chaque échantillon prélevé) et le codage.

2.8 Les modes de représntation du son

[13]

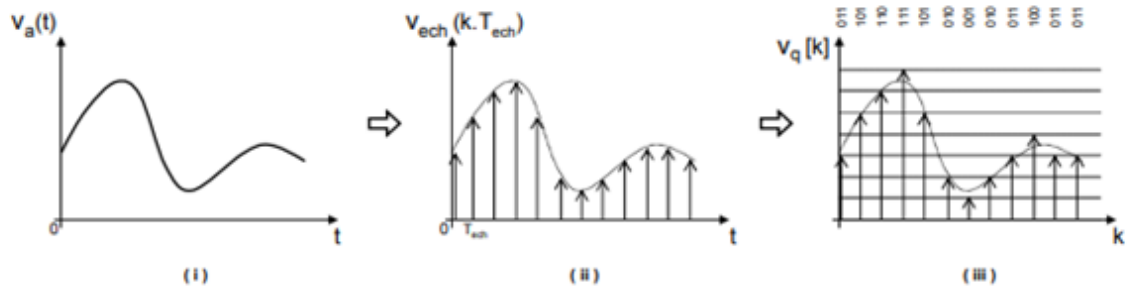

FIGURE 2.3: (i) signal analogique (ii) signal échantillonné (iii) puis quantifié.

2.8.1 Représentation temporelle

Cette représentation montre l'évolution de l'intensité du signal sonore dans le temps.

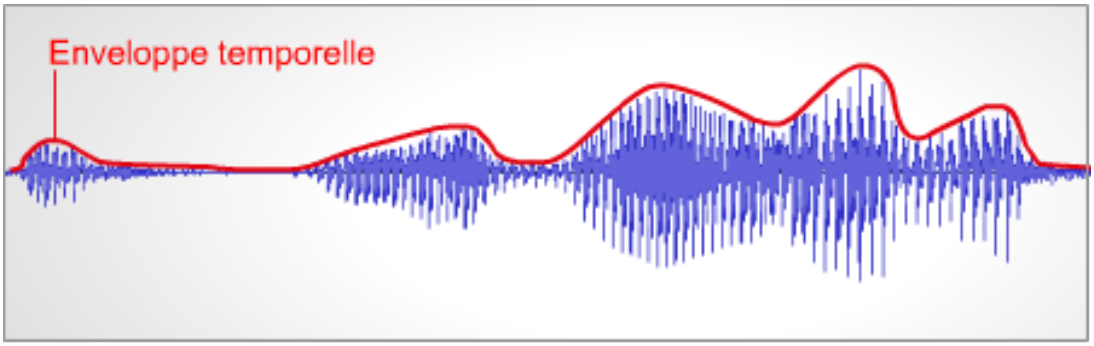

FIGURE 2.4: Représentation temporelle.

La partie en bleu montre l'évolution de l'intensité d'un son de parole dans le temps. Cette vue temporelle permet notamment d'apprécier l'évolution de l'enveloppe temporelle (ligne rouge) qui, par exemple, joue un rôle important dans la perception de la parole.

2.8.2 Représentation fréquentielle (ou spectrale)

Ce mode permet de visualiser la composition fréquentielle d'un son mais également l'intensité de chaque fréquence.

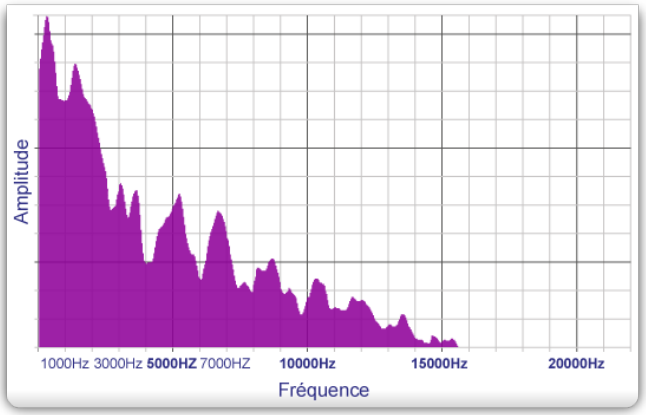

FIGURE 2.5: Représentation fréquentielle.

Sur ce graphe, on peut voir la composition spectrale de l'échantillon sonore précédent. On peut ainsi voir dans cet exemple que les fréquences du son choisi s'étendent de 80 Hz à 15500 Hz.

2.8.3 Représentation tridimensionnelle : le sonagramme ou spectrogramme

Il s'agit de la représentation temps-fréquence du son. On trace la répartition énergétique du son en fonction du temps et des fréquences. Le sonagramme est très utilisé pour étudier le signal de parole.

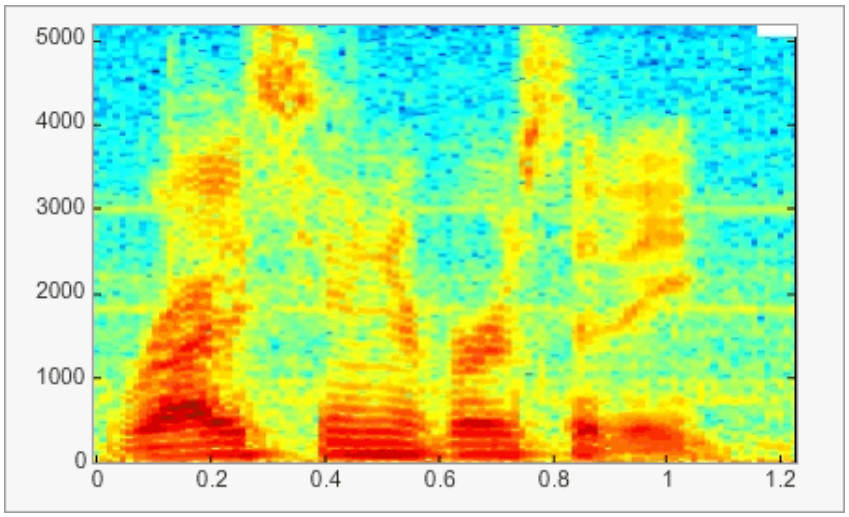

FIGURE 2.6: Représentation tridimensionnelle.

L'exemple ci-dessus montre bien l'évolution de la fréquence et de l'intensité dans le temps. L'intensité est définie par la couleur : plus la couleur évolue vers le rouge plus l'intensité est importante. Les trait noirs soulignent les formants des voyelles, les transitoires formantiques, ...

2.9 Analyse du signal de parole

On ne conserve comme signal de parole que les zones où l'énergie est supérieure au seuil choisi. Par des transformations mathématiques, on extrait ensuite de ce signal segmenté certaines caractéristiques propres à la voix. Seules les fréquences propres à la voix humaine, c'est-à-dire comprises entre 200 Hz et 3400 Hz, sont analysées. Les caractéristiques extraites sont en relation avec le contenu fréquentiel de la parole, la forme du conduit vocal, l'intonation ou encore la prosodie. Elles concernent les fréquences les plus présentes dans la voix, ainsi qu'une information d'intonation ou de transition entre les fréquences à chaque instant. Pour chaque trame de parole, on extrait ainsi un vecteur de 20 à 30 caractéristiques qui sont les coefficients « cepstraux », leurs dérivées et l'énergie du signal. [11]

2.10 Le traitement du signal à la prise etla restitution du son

[15]

Le traitement de la parole fait l'objet derecherches dans tous les laboratoires desgrands opérateurs de télécommunications,souvent depuis leurs premières annéesd'existence. Les travaux se sont intensifiésavec l'apparition du traitement numériquedu signal. Ce vaste domaine est classiquementdécoupé en quatre grandes spécialités :

-le traitement du signal à la prise et la restitution du son
-le codage de la parole
-la synthèse de la parole
-la reconnaissance de la parole.

2.11 Que pouvons-nous attendre de la Reconnaissance vocale ?

[50]

À notre avis, et de l'avis de nombreux experts, Reconnaissance vocale ou la biométrie vocale devrait, pour l'instant, être utilisée en plus d'autres méthodes d'authentification plus éprouvées. Ce faisant, les avantages respectifs des différentes méthodes peuvent devenir complémentaires. Par exemple, la combinaison de l'identification vocale et faciale est déjà une piste explorée par de nombreux acteurs.
Un exemple, que vous pouvez déjà utiliser à la maison si vous avez un locuteur intelligent, est le Voice Match, la capacité des assistants à reconnaître les personnes d'une même famille. D'où la personnalisation avancée de l'expérience, en termes de préférences, d'accessibilité ou d'autorisations par exemple.

2.12 La reconnaissance vocale définition

[30]

La reconnaissance vocale est un domaine scientifique ayant toujours eu un grand attrait aussi bien auprès des chercheurs qu'auprès du grand public. Elle consiste à employer des techniques d'appariement afin de comparer une onde sonore à un ensemble d'échantillons, composés généralement de mots mais aussi, plus récemment, de phonèmes (unité sonore minimale). Le secteur de la reconnaissance vocale est en pleine croissance et cette technologie bien que très avancée, n'est pas encore aboutie, pouvant commencer à répondre aux attentes de l'homme. Bien que des progrès soient encore à faire sur les systèmes complexes de traitement et reconnaissance, il est annoter que la reconnaissance Vocale est quasiment parfaite. Sans compter le coût de ces systèmes qui a considérablement chuté ces dernières années mais aussi le gain qu'ils peuvent apporter à un particulier et surtout à une entreprise. 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.

2.13 Domaine d'application de la reconnaissance vocale

[9]

traduction automatique : de conversations téléphoniques avec un interlocuteur de langue étrangère

télématique et services vocaux : composeur vocal, serveurs vocaux interactifs, service PCV, consultation de messagerie vocale, majordome d'accueil vocal téléphonique, etc.

bornes interactives : renseignements sur les horaires (train, avion, bateau) et prise de réservations

bureautique : services télématiques vocaux et commandes vocales d'éditeur

contrôle de qualité et saisie de données : l'interface vocale libère la vue et les mouvements, l'utilisateur peut donc se déplacer librement pour manipuler des objets ou entrer des données

aide à la conception graphique : système d'interaction multimodale, incluant parole, geste et vision

avionique : permet aux pilotes une meilleure attention visuelle

aide à la navigation en voiture : permet le positionnement du véhicule, la planification de l'itinéraire et notamment le guidage du conducteur par des messages vocaux

aide à la formation : apprentissage des langues, de la lecture, formation des contrôleurs aériens (meilleure connaissance de la phraséologie spécialisée du domaine)

aide au handicap : aide à la rééducation de la voix, contrôle d'objets de l'environnement pour les tétraplégiques, consultation de documents pour les aveugles (tâches d'édition et de consultation)

dictée automatique ou entrée vocale de textes : contrôle d'un microscope, interrogation vocale d'une base de données, constitution automatique de rapports médicaux par dictée vocale

identification / vérification du locuteur : pour assurer une meilleure sécurité pour l'accès en direct à des bases de données confidentielles

2.14 Reconnaissance vocale et la biométrie

dans le domaine de la biométrie, on rivalise des moyens pour identifier tout un chacun. Depuis quelques années déjà, nos doigts sont un sésame pour déverrouiller nos téléphones. on peut également régler une transaction désormais avec son visage. La prochaine rivale de l'empreinte digitale considérée comme infaillible est peut-être la voix, selon le Wall Street Journal. « On sait depuis des siècles que la voix porte en elle quantité d'informations. Grâce à l'intelligence artificielle, on peut soutirer ces informations », déclare au journal Rita Singh, chercheuse spécialisée en apprentissage machine appliqué à la voix, à la Carnegie Mellon University. Il y a quelques mois, toujours au même journal, l'universitaire avait précisé que « la voix humaine contient des informations, liées à nos caractéristiques physiques, physiologiques, démographiques, médicales, et environnementales ». Utile pour le profilage en tout genre.

Les entreprises auraient déjà recours à la voix pour prévenir les fraudes. Les banques, notamment. HSBC est ainsi la première banque à avoir mis en place un système de reconnaissance vocale pour ses clients. Quelque 15 millions de clients peut accéder en Grande-Bretagne à leurs comptes d'un simple « bonjour », même avec un rhume, s'amusait en 2016 le Guardian. En réalité, le client doit dire « my voice is my password » (« ma voix est mon mot de passe ») pour s'identifier. Quelques mois plus tard, un journaliste de la BBC pointe une première défaillance du système : il demande alors à son jumeau de tenter d'accéder à son compte, ce qu'il réussit à faire... après 7 tentatives. La banque déclare avoir depuis rectifié le tir et a annoncé en avril dernier avoir évité pour 300 millions de livres sterling (325 millions d'euros) de fraudes. Au WSJ, Daniel Capozzi,

un des porte-paroles de Discover Financial Services – une entreprise américaine spécialisée dans les cartes de crédit – explique avoir réduit les fraudes de 10 depuis que la société a recours à un système d'analyse vocale.

[37]

2.15 La reconnaissance automatique du locuteur (RAL)

La reconnaissance automatique du locuteur : est interprété comme une tâche particulière de reconnaissance de formes .Ce domaine regroupe les problème relatifs à l'identification ou à la vérification du locuteur sur base de l'information contenue dans le signe dans le signal acoustique : il s'agit de reconnaître une personne à partir de sa voix . [55]

2.16 Différentes Tâches en RAL

2.16.1 Identification automatique du locuteur (IAL)

Le principe de l'identification automatique du locuteur l'IAL consiste à retrouver l'identité du locuteur associé parmi une population de locuteurs connus. D'un point de vue schématique, une séquence de parole est donnée en entrée du système d'IAL. Pour chaque locuteur connu du système, la séquence de parole est " comparée " à une référence caractéristique du locuteur. L'identité du locuteur dont la référence est la plus "proche" de la séquence de parole est donnée en sortie du système d'IAL [43]

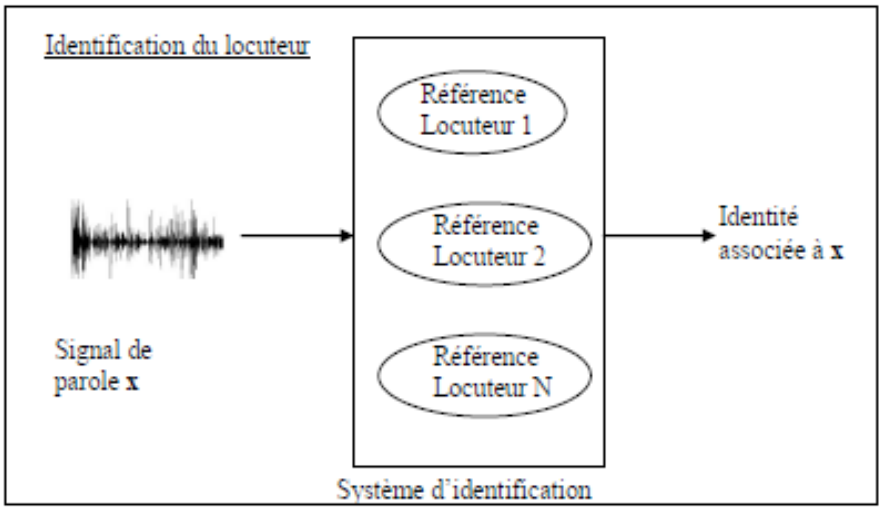

FIGURE 2.7: Principe de base de l'identification du locuteur .

2.16.2 La Vérification Automatique du Locuteur (VAL)

[12]

La Vérification Automatique du Locuteur (VAL) est le processus décisionnel permettant de déterminer, au moyen d'un message vocal, la véracité de l'identité revendiquée par un individu dont la FIGURE 2.2 représente le principe de VAL

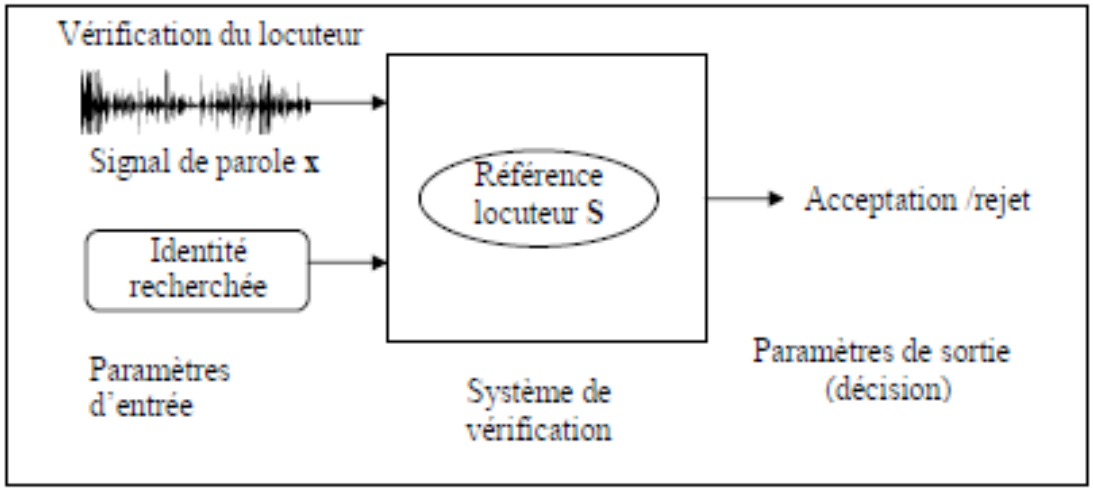

FIGURE 2.8: Principe de base de Automatique du locuteur .

2.17 Comment ça marche ?

[24]

Du jour où le téléphone a été inventé, des esprits ingénieux ont relevé que le micro constituait un détecteur de son susceptible d'envoyer un signal électrique au plus bas bruit. D'où l'idée d'envoyer ce signal à une machine pour qu'elle à la voix [. . .]. Pour que la reconnaissance vocale devienne une réalité, il fallait donc un outil capable d'analyser les sons pour séparer le bruit de fond d'un atelier et les commandes du machiniste, le progrès de la reconnaissance vocale était étroitement lié à l'évolution des ordinateurs.

En 1971, apparait le premier dispositif de commande vocale, le voice commande system, fondé sur une calculatrice capable, après un cycle d'apprentissage reconnaître 24 ordres. Le procédé va ensuite se scinder en deux grandes sections familiales.

Le premier est une simple commande vocale l'application de la chose commande. L'appareil doit être capable de reconnaître l'énoncé de nom et pour cela, il faut répéter plusieurs le nom choisir qui sera alors mémorisé sous forme d'une séquence sonore. Une fois le nom aura été de l'appareil analyse l'évolution de l'intensité sur plusieurs plages de fréquences (en général huit) pour chacun de ces échantillons vocaux.

Après avoir réalisé une moyenne des huit échantillons, l'appareil en déduit un profil acoustique numérisé du nom et le met en mémoire en usage normal, pour reconnaître un nom prononcé par l'utilisateur, il compare le nouveau profil à ceux qu'il possède en mémoire .il attribue alors des notes statistiques de ressemblance et décrète que le nom prononcé est celui qui la meilleure note
La seconde famille, celle des systèmes de dictée, nécessite un ordinateur puissant car leur principe de reconnaissance vocale est infiniment plus complexe : ils doivent être capables de reconnaître des mots prononcés en langage dit naturel,
C'est –à-dire au sein de phrases ou – notamment en français – les mots s'enchaînent.
Qui plus est , ils doivent tenir compte des liaison , et aussi des homophones qui se créent spontanément par la juxtaposition de mots . [. . .]
Après amplification et tri par fréquences , grâce à un jeu de filtres électroniques rappelant les " égaliseurs " des chaînes haute fidélité , un spectrogramme de la phrase est obtenu .
Pour l'ordinateur , la première tâche consiste à séparer chaque phonème . Il le transforme alors en un fichier numérique . un tableau de données .
Ensuite , il compare ces tableaux obtenus à ceux qeu contient un dictionnaire , stocké sur le disque dur , ou sont associes phonèmes ou groupes de phonèmes et mots réels . Mais , par la présence des liaisons ou par homophonie , de très nombreux mots peuvent correspondre au message parlé .Interviennent alors des logiciels d'analyse contextuelle et sémantique souvent , l'ordinateur devra attendre qu'une grande partie de la phrase ait été prononcée pour que ces logiciels puissent commencer leur travail . L'apparition d'un nouveau mot peut bouleverser son sens et donc le choix de coupe des phonèmes . Par déduction statistique les logiciels retiennent la solution la plus probable qui se révèle le plus souvent , être la bonne .
Beaucoup de systèmes actuels sont encore monolocuteurs , c'est -à-dire qu'ils ne peuvent " comprendre " qu'une seule personne , et encore après une phase d'apprentissage. On cherche donc maintenant à faire des systèmes multilocuteurs qui " entendront " aussi bien le langage d'un homme à l'accent parisien que celui d'une femme au par les méditerranéen , et cela sans phase d'apprentissage .

2.18 État de l'art

2.18.1 Origin-STT

À l'heure où la recherche vocale se répand sur les réseaux sociaux, la prise de notes peut également se faire à l'oral. C'est sur cette base qu'a été créé l'application Origin-STT, initiée et développée par un groupe de jeunes techniciens de l'entreprise VAIS.
Origin-STT a décroché, en 2019, le premier prix en termes de technologies et d'information du prix "Talent du Vietnam".
Origin-STT est une application made in Vietnam qui permet à ses utilisateurs de retranscrire automatiquement toutes sortes d'enregistrements sonores. Grâce à l'appui de l'intelligence artifi-

cielle, Origin-STT peut donc retranscrire, sous forme de texte, toutes sortes d'enregistrements de voix parlée, mais en vietnamien uniquement. Cette application est capable de reconnaître les différents accents des trois régions du pays. Une fois installée sur votre ordinateur ou votre tablette, Origin-STT vous propose des transcriptions fidèles à 94 %, les noms propres étant automatiquement mis à part, en majuscules.
C'est en 2018 qu'Origin-STT a été lancée. Première application de reconnaissance vocale en langue vietnamienne, elle a été utilisée dès 2019 lors des séances de question-réponse de l'Assemblée nationale.
Grâce à ses performances, en 2018 et 2019, Origin-STT a décroché le premier prix d'un concours de traitement de la voix en langue vietnamienne

[19]

2.18.2 Amazon Transcribe Medical

en 2019. Aujourd'hui, après une approche sur la reconnaissance vocale en faveur du handicap, nous nous intéressons aux applications de cette technologie dans le monde médical.
La première application dont nous allons parler ici est la transcription et retranscription automatique dans le domaine médical et de la santé. Notre premier exemple concerne Amazon – le géant de l'e-commerce en ligne – qui s'est à juste titre lancé sur le secteur de la retranscription médicale. En effet, selon le média américain CNBC, Amazon laisse les docteurs enregistrer leurs conversations et les mettre dans leurs dossiers médicaux.
Ainsi, la reconnaissance vocale d'Amazon s'utilise pour permettre aux médecins de passer plus de temps avec leurs patients et moins de temps à leur ordinateur. En effet, l'entreprise lance un service nommé « Amazon Transcribe Medical ». Cette application a pour objectif de transcrire les échanges entre docteur et patient puis la transcription écrite – le texte – est directement sauvegardé dans le dossier médical.
« le but principal est de libérer les docteurs, afin qu'ils puissent se concentrer davantage sur leur patient qui est directement concerné »

[4]

2.18.3 GRIFOS

Il est indispensable d'aider les personnes ayant des problèmes d'élocution à communiquer plus efficacement, d'une part pour améliorer leur qualité de vie et d'autre part, pour leur offrir une plus grande indépendance. Comme le montrent les résultats du projet de RDT intitulé OLP, la technologie peut constituer un important vecteur dans ce domaine.
L'un des principaux résultats du projet a été le développement de l'outil de reconnaissance vocale OLP, appelé GRIFOS. Il a été créé à l'Universidad Politecnica de Madrid (Espagne) en 2020. GRIFOS est un système automatique de reconnaissance vocale destiné aux personnes souffrant de troubles importants de la parole tels que la dysarthrie.

Pendant une phase de formation, GRIFOS «apprend » à reconnaître l'élocution particulière du patient. Une base de données existante contenant des informations d'autres utilisateurs souffrant du même problème d'élocution vient compléter ce système.
GRIFOS est utilisé initialement pour aider les patients à maintenir leurs modèles d'élocution dans des seuils quantitatifs particuliers définis par le système. À terme, les exercices sont enrichis pour s'attaquer à la question du discours continu. Dans le cas de la dysarthrie, GRIFOS vise à encourager la cohérence en termes de modèles d'élocution.
Comme les autres outils logiciels d'OLP, GRIFOS représente une approche innovante à l'orthophonie. Il s'agit d'un outil très performant mis à disposition des patients, mais également de leurs thérapeutes. Vous pouvez obtenir un prototype auprès de l'Universidad Politecnica de Madrid et de ses partenaires du projet OLP.

[3]

2.19 Conclusion

La reconnaissance vocale est un usage qui n'est plus à prouver. En effet, les interfaces vocales et assistants vocaux sont aujourd'hui plus performants que jamais et se développent dans de nombreux domaines. Cette croissance exponentielle et continue donne lieu à une diversification des applications de la reconnaissance vocale et des technologies liées. Donc Dans ce chapitre on a fait une étude, qui suit la reconnaissance vocale depuis sa production jusqu'à son analyse, ainsi que le son et single et et sa transformation en représentation temporelle à la représentation fréquentielle,

CHAPITRE 3

DEEP LEARNING

3.1 Introduction

L' intelligence artificielle est une discipline scientifique et un ensemble de théories et des techniques recherchant des solutions des problèmes à forte complexité logique ou algorithmique, elle est basée sur une démarche d'apprentissage afin de reproduire une partie de l'intelligence humaine à travers une application, un système ou un processus. La reconnaissance faciale, la perception visuelle et autre sont des exemples de systèmes d'intelligence artificielle.

- La machine Learning (ML) est un sous-domaine de l'IA qui utilise les réseaux neuronaux artificiels (ANN) pour imiter la façon dont les êtres humains prennent des décisions. Le machine Learning permet aux ordinateurs de développer des modèles d'apprentissage par eux-mêmes, sans aucune programmation, à partir de gros ensembles de données.

Par conséquent, L'apprentissage profond (en anglais deep learning, deep structured learning, hierarchical learning) est une forme d'intelligence artificielle, dérivée de la machine learning basé sur un type particulier de mécanisme d'apprentissage.

Pour illustrer la relation entre ces termes, nous pouvons utiliser des cercles concentriques :

• Intelligence artificielle IA (intelligence artificielle) : le cercle plus large est l'idée qui a émergé en premier dans ce domaine

• Apprentissage automatique (machine Learning) : au milieu, il a prospéré plus tard après l'IA

• Apprentissage approfondie (Deep Learning) : le plus petit cercle est une expansion de l'IA actuellement [2]

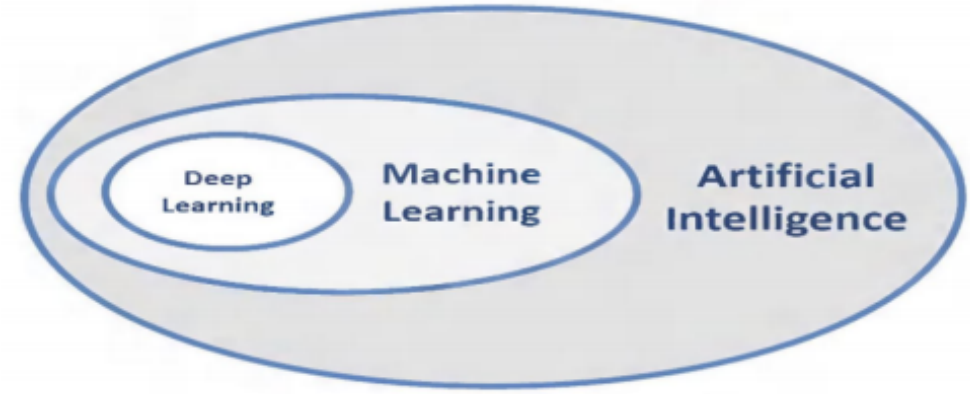

FIGURE 3.1: la relation entre l'IA,ML,Deep learning.

3.2 Définition de l'apprentissage profond (deep learning)

L'apprentissage profond (« deep learning ») est un domaine de recherche sur l'apprentissage automatique qui a permis des avancées importantes en intelligence artificielle dans les dernières années. [1]
L'apprentissage profond a été introduit dans l'étude des réseaux de neurones, inspirée à l'origine par une étude du fonctionnement physiologique du cerveau, L'apprentissage profond capable d'apprendre et de traiter des données complexes, et tente également de résoudre des tâches complexes

3.3 Les applications du Deep Learning [40]

C'est une branche du Machine Learning très prometteuse. Que ce soit pour reconnaître des visages sur des images, analyser des textes et les interpréter automatiquement ou encore avoir des voitures qui conduisent toutes seules, les applications du Deep Learning sont nombreuses. Aujourd'hui, nous vous proposons un workshop qui va vous permettre de comprendre ce domaine et voir en quoi nous pourrions concretement l'utiliser.

3.3.1 La reconnaissance faciale

Les yeux, le nez, la bouche, tout autant de caractéristiques qu'un algorithme de Deep Learning va apprendre à détecter sur une photo. Il va s'agir en premier lieu de donner un certain nombre d'images à l'algorithme, puis à force d'entraînement, l'algorithme va être en mesure de détecter un visage sur une image.

3.3.2 La détection d'objets

Sur une image complexe où il y a plusieurs éléments, les algorithmes de détection d'objets vont être maintenant capables d'identifier et de localiser au pixel près un élément ou une personne. 800 millions d'images sont uploadées chaque jour sur Facebook : son algorithme Deep Learning est effectivement capable d'identifier telle ou telle personne sur une photo dès lors qu'elle est uploadée

3.3.3 Le Natural Language Processing

Le Natural LanguageProcessing est une autre application du Deep Learning. Son but étant d'extraire le sens des mots, voire des phrases pour faire de l'analyse de sentiments. L'algorithme va par exemple comprendre ce qui est dit dans un avis Google, ou va communiquer avec des personnes via des chatbots. La lecture et l'analyse automatique de textes est aussi un des champs d'application du Deep Learning avec le TopicModeling : tel texte aborde tel sujet.

3.4 Domaines d'application de l'apprentissage profonde : [45]

L'apprentissage profond à de nombreuses applications en informatique, on cite quelques domaines :

3.4.1 Conduite automatisée :

Les chercheurs du secteur automobile ont recours au Deep Learning pour détecter automatiquement des objets tels que les panneaux stop et les feux de circulation. Le Deep Learning est également utilisé pour détecter les piétons, évitant ainsi nombre d'accidents.

3.4.2 Recherche médicale :

À l'aide du Deep Learning, les chercheurs en cancérologie peuvent dépister automatiquement les cellules cancéreuses. Des équipes de l'Université de Californie à Los Angeles (UCLA) ont conçu un microscope qui génère un ensemble de données de grandes dimensions afin d'entraîner une application de Deep Learning à identifier avec précision des cellules cancéreuses.

3.4.3 Électronique

Le Deep Learning est utilisé pour la reconnaissance audio et vocale. Par exemple, les appareils d'assistance à domicile qui répondent à votre voix et connaissent vos préférences fonctionnent grâce à des applications de Deep Learning.

3.5 Histoire de Deep Learning [21]

Depuis 2012, les algorithmes à base de deep learning (apprentissage profond) semblent prêts à résoudre bien des problèmes : reconnaitre des visages comme le propose DeepFace, vaincre des joueurs de go ou de poker ou bientôt permettre la conduite de voitures autonomes ou encore la recherche de cellules cancéreuses.

Pourtant, les fondements de ces méthodes ne sont pas si récents : le deep learning a été formalisé en 2007 à partir des nouvelles architectures de réseaux de neurones dont les précurseurs sont McCulloch et Pitts en 1943. Suivront de nombreux développements comme le perceptron, les réseaux de neurones convolutifs de Yann Le Cun et Yoshua Bengio en 1998 et les réseaux de neurones profonds qui en découlent en 2012 et ouvrent la voie à de nombreux champs d'application comme la vision, le traitement du langage ou la reconnaissance de la parole. Pourquoi maintenant ? parce que ces nouvelles techniques de machine learning profitent de données massives (big data) que l'on est désormais capables d'analyser ainsi que de capacités de calcul phénoménales notamment grâce aux processeurs graphiques. Preuve que chaque domaine irrigue les autres, c'est pour pouvoir utiliser les immenses promesses du deep learning que Google a mis au point les accélérateurs TPU.

3.6 Forces et faiblesses

[48]

L'apprentissage profond a permis d'augmenter considérablement la puissance des intelligences artificielles. L'attention que lui portent les médias et le grand public est donc tout à fait justifiée. Mais pour que cette technologie réalise son plein potentiel, il convient de corriger certaines de ses faiblesses.

3.6.1 Les points forts de l'apprentissage en profondeur

. Meilleurs résultats qu'avec d'autres méthodes d'apprentissage machine
.Aucun développement manuel des fonctionnalités nécessaire, aucun étiquetage des données nécessaire
.Exécution efficace des tâches de routine, sans écarts de qualité Traitement des données non structurées
.Diversification des services visant à simplifier l'utilisation des réseaux de neurones artificiels

3.6.2 Les points faibles de l'apprentissage profond

.Nécessite une grande puissance de calcul
.Le développement d'algorithmes d'apprentissage prend un temps relativement long

.Nécessite une vaste base de données
.Davantage de données d'instruction initiales requises qu'avec d'autres méthodes d'apprentissage machine
.Décisions difficilement ou pas du tout compréhensibles (boîte noire)

3.7 Comment ça marche ?

Pour comprendre comment fonctionne le Deep Learning, nous allons utiliser un exemple concret de reconnaissance faciale. Imaginons que notre objectif soit de lui faire reconnaître les photos qui comportent une voiture.
Pour pouvoir reconnaître une voiture, l'algorithme doit d'une part savoir distinguer tous les types de voitures existantes, mais aussi savoir identifier une voiture de manière précise et autonome, quel que soit l'angle sous lequel elle se trouve.
Pour y arriver c'est assez simple : le réseau de neurones artificiels est entraîné en analysant des milliers d'images de voitures et apprend à les reconnaître au milieu de photos d'autres objets.
Ces données vont ensuite être assignées à différentes informations permettant à l'algorithme intelligent de déduire si oui ou non se trouve une voiture sur l'image qu'il est en train d'analyser.
Le réseau artificiel va également comparer cette réponse aux bonnes réponses indiquées par les humains. Si il a vu juste, l'algorithme de reconnaissance garde cette réussite en mémoire et s'en resservira plus tard pour reconnaître des voitures. Au contraire, s'il s'est trompé, il en prend note et corrige son erreur de lui-même la fois suivante.
C'est en répétant ce système d'entraînement des milliers de fois que le réseau de neurones finit par être capable de reconnaître une voiture dans toutes circonstances (avec un degré de réussite proportionnel à la durée d'entraînement du réseau et au nombre de couches qu'il possède).
Cette technique d'apprentissage est appelée apprentissage supervisé ou “supervised learning”. [5]

3.8 Réseaux de neurones

Avec l'augmentation rapide des capacités de calcul des ordinateurs et l'évolution des techniques d'apprentissage, l'utilisation des réseaux de neurones connaît un vif essor en particulier au sein des communautés du traitement d'image, de la traduction automatique et du traitement de la parole.
Au cours des dernières années, deux types de réseaux de neurones ont marqué une rupture technologique dans le domaine du traitement de la parole : les réseaux dits "profonds" et les réseaux récurrents. Nous détaillons dans cette partie beaucoup plus le fonctionnement de réseaux de neurones et les travaux que nous avons réalisés sur les modèles eux-mêmes et sur les méthodes d'apprentissage. [1]

3.9 Principe de fonctionnement

Le réseau comporte 3 composants : couche d'entrée couche cachée et couche de sortie. Le terme « profond » se rapporte généralement au nombre de couches cachées du réseau de neurones. Les réseaux de neurones classiques ne comportent que 2 à 3 couches cachées, tandis que les réseaux profonds peuvent en compter jusqu'à 150 [46]
L'idée est d'utiliser la structure de couche de réseau neuronal en empilant plusieurs couches les unes sur les autres, de manière à faciliter le mécanisme de décomposition. Par conséquent, chaque couche d'un réseau de neurones profonds (Deep Neural Networks DNN) fonctionne comme une seule transformation pour extraire davantage les données [51]

Le réseau de neurones le plus connu et le plus simple à comprendre est le réseau de neurones multicouches à anticipation. Il contient un calque d'entrée, un ou plusieurs calques masqués et un seul calque de sortie. Chaque couche peut avoir un nombre différent de neurones et chaque couche est entièrement connectée à la couche adjacente

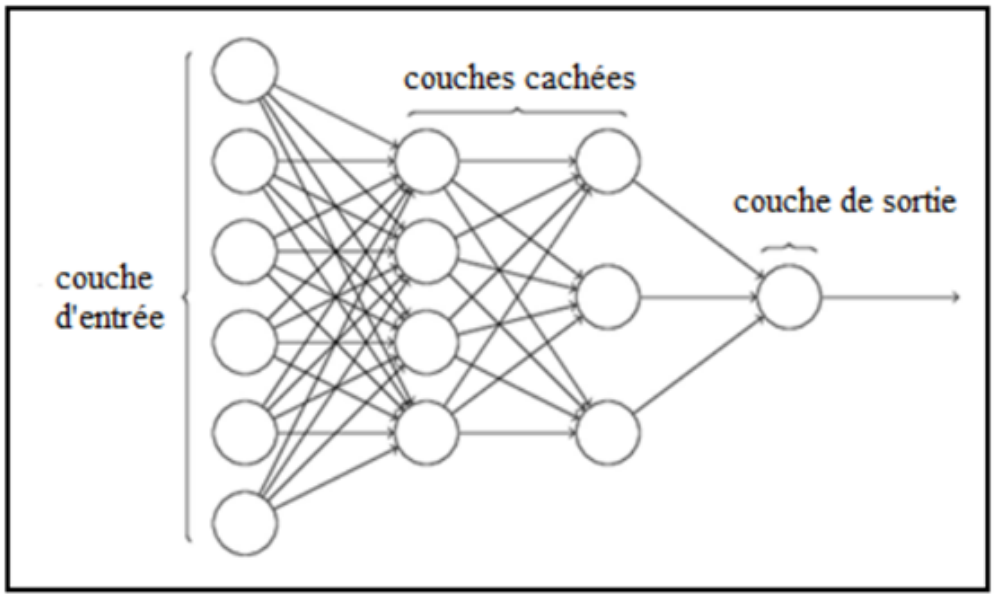

FIGURE 3.2: Topologie de réseau de neurones avec une seule sortie.

Un réseau de neurones est défini comme un ensemble de nœuds (appelés neurones) connectés via des liaisons dirigées (flèche), chaque flèche représente une connexion entre la sortie d'un neurone et l'entrée d'un autre(les flèches entrantes étant les entrées du neurone et les flèches sortantes étant les sorties du neurone), Chaque flèche porte un poids, reflétant son importance, chaque nœud étant une unité de traitement qui exécute une fonction de nœud statique sur son signal entrant pour générer une sortie de nœud unique [47]

. Les valeurs d'entrée, ou en d'autres termes, nos données sous-jacentes, sont transmises

via ce «réseau» de couches masquées jusqu'à ce qu'elles convergent vers la couche de sortie. La couche en sortie correspond à notre prédiction : il peut s'agir d'un nœud si le modèle ne génère qu'un nombre ou de quelques nœuds s'il s'agit d'un problème de classification multi-classe. La forme à l'intérieur des neurones dans les couches centrales représente une fonction d'activation (typiquement un 1 = (1 + e x)) qui est appliquée à la valeur du neurone avant de le transmettre à la sortie [26]

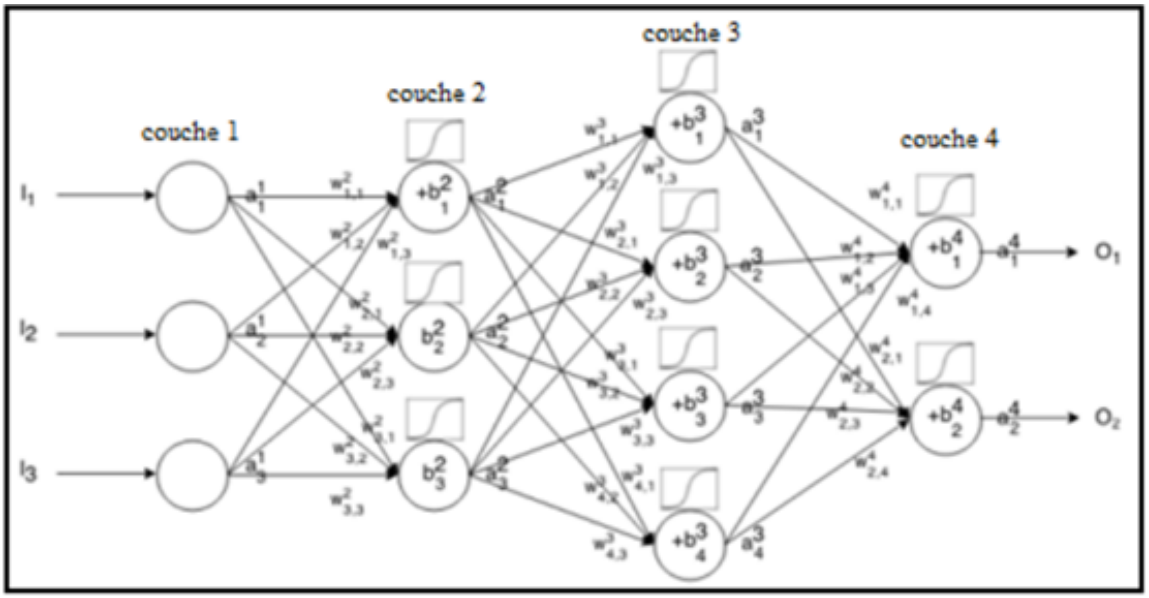

FIGURE 3.3: Topologie de réseau de neurones profond.

Les couches cachées d'un réseau de neurones apportent des modifications aux données pour éventuellement déterminer quelle est sa relation avec la variable cible. Chaque nœud a un poids et multiplie sa valeur d'entrée par ce poids. Pour déterminer ce que devraient être ces petits poids, nous utilisons généralement un algorithme appelé Back propagation.

3.10 L'apprentissage

En effet, les réseaux de neurones peuvent trouver le lien qui unit des valeurs de sortie à celles en entrée, et ce, même lorsqu'on ne connaît pas cette fonction a priori. Afin de parvenir à ce résultat, il faut « entraîner » le modèle, à l'aide d'un jeu de données initial.
Celui-ci est alors divisé en deux parties : une pour l'apprentissage, l'autre pour tester le réseau. Dans un premier temps, on va donc soumettre des données d'entraînement au programme, qui comprennent des valeurs d'entrée, ainsi que les valeurs de sortie attendues. Au début, le réseau de neurones va tenter de calculer les résultats, mais avec peu d'informations, et va donc

commettre des erreurs. On va ensuite ajuster ses paramètres, de sorte à réduire ces écarts à chaque itération.
Par exemple, un réseau de neurones peut être utilisé pour prévoir les risques d'apparition d'une maladie chez certains individus. En lui soumettant les caractéristiques de patients, le modèle va d'abord effectuer des prédictions aléatoires. Puis, en apprenant de ses erreurs, il sera de plus en plus pertinent dans son analyse [6]

3.10.1 Apprentissage supervisé

La majorité des apprentissages automatiques utilisent un apprentissage supervisé (supervised learning).

L'apprentissage supervisé consiste en des variables d'entrée (x) et une variable de sortie (Y). Vous utilisez un algorithme pour apprendre la fonction de mapping de l'entrée à la sortie.

Y = f (X)

Le but est d'appréhender si bien la fonction de mapping que, lorsque vous avez de nouvelles données d'entrée (x), vous pouvez prédire les variables de sortie (Y) pour ces données.

C'est ce qu'on appelle l'apprentissage supervisé, car le processus d'un algorithme tiré de l'ensemble de données de formation (training set) peut être considéré comme un enseignant supervisant le processus d'apprentissage. Nous connaissons les réponses correctes, l'algorithme effectue des prédictions itératives sur les données d'apprentissage et est corrigé par l'enseignant. L'apprentissage s'arrête lorsque l'algorithme atteint un niveau de performance acceptable.

3.10.2 Apprentissage non supervisé

L'apprentissage non supervisé (Unsupervised Learning) consiste à ne disposer que de données d'entrée (X) et pas de variables de sortie correspondantes.

L'objectif de l'apprentissage non supervisé est de modéliser la structure ou la distribution sous-jacente dans les données afin d'en apprendre davantage sur les données.

On l'appelle apprentissage non supervisé car, contrairement à l'apprentissage supervisé ci-dessus, il n'y a pas de réponse correcte ni d'enseignant. Les algorithmes sont laissés à leurs propres mécanismes pour découvrir et présenter la structure intéressante des données.

L'apprentissage non supervisé comprend deux catégories d'algorithmes : Algorithmes de regroupement et d'association.

3.10.3 Apprentissage semi-supervisé

Les problèmes pour lesquels vous avez une grande quantité de données d'entrée (X) et que seules certaines données sont étiquetées (Y) sont appelés problèmes d'apprentissage semi-supervisés. Par conséquent, ces problèmes se situent entre l'apprentissage supervisé et l'apprentissage non supervisé.
Exemple : une archive de photos dans laquelle seules certaines images sont étiquetées (chien, chat, personne, par exemple) et la plupart ne le sont pas.
De nombreux problèmes de machine learning du monde réel tombent dans ce domaine. En effet, il peut être coûteux en temps ou en argent d'étiqueter des données car cela peut nécessiter un accès à des experts de domaine. Considérant que les données sans étiquette sont peu coûteuses et faciles à collecter et à stocker.
Vous pouvez utiliser des techniques d'apprentissage non supervisées pour découvrir et apprendre la structure dans les variables d'entrée.

Vous pouvez également utiliser des techniques d'apprentissage supervisé pour établir des prévisions optimales pour les données non étiquetées, les transférer dans l'algorithme d'apprentissage supervisé en tant que données d'apprentissage et utiliser le modèle pour effectuer des prédictions sur de nouvelles données invisibles.

[28]

3.11 L'écueil du surapprentissage

Mais à quel moment faut-il arrêter d'entraîner le réseau de neurones ? En réalité, il ne suffit pas d'obtenir les meilleurs résultats sur le jeu d'entraînement. Car le programme pourrait alors être si précis sur ces données qu'il deviendrait incapable de généraliser. On parle alors de « surapprentissage ».
Pour schématiser, reprenons le cas de la prévention de maladies. Imaginons que, dans le jeu de données initial, le programme constate que plusieurs patients à lunettes sont atteints d'un cancer. Avec du surapprentissage, le programme pourrait déduire que le port de lunettes est symptomatique d'un risque élevé d'apparition d'une telle maladie.
Afin d'éviter cet écueil, lors de la phase d'apprentissage, on divise le jeu d'entraînement en deux sous-ensembles, qu'on soumet en parallèle au réseau de neurones. Et tandis que le modèle apprend avec le premier, le deuxième sert d'échantillon de validation, pour vérifier la capacité

d'extrapolation du programme. La meilleure configuration intervient alors quand l'erreur est minimale sur le jeu de validation.Les réseaux de neurones artificiels se distinguent donc par leur capacité à apprendre et à généraliser. Parmi les structures possibles, une forme est de plus en plus utilisée [6]

3.12 Réseaux de neurones récurrents

Un réseau de neurones récurrent (RNN, recurrent neural network) est un type de réseau de neurones artificiels principalement utilisé dans la reconnaissance vocale et le traitement automatique du langage naturel. Les RNN sont conçus de manière à reconnaître les caractéristiques séquentielles et les modèles d'utilisation des données requis pour prédire le scénario suivant le plus probable.
Les Réseaux de Neurones récurrents traitent l'information en cycle. Ces cycles permettent au réseau de traiter l'information plusieurs fois en la renvoyant à chaque fois au sein du réseau.
La force des Réseaux de neurones récurrents réside dans leur capacité de prendre en compte des informations contextuelles suite à la récurrence du traitement de la même information. Cette dynamique auto-entretient le réseau. Les Réseaux de neurones récurrents se composent d'une ou plusieurs couches. Le modèle de Hopfield (réseau temporel) est le réseau de neurones récurrent d'une seule couche le plus connu.
Les Réseaux de neurones récurrents à couches multiples revendiquent quant à eux la particularité de posséder des couples (entrée/sortie) comme les perceptrons entre lesquels la donnée véhicule à la fois en propagation en avant et en rétro propagation.

[8]

3.13 Convolutional Neural Network

Un réseau neuronal convolutif (CNN) est une architecture réseau pour le Deep Learning qui apprend directement à partir des données, ce qui évite d'extraire manuellement les caractéristiques.
Les réseaux neuronaux convolutifs sont particulièrement utiles pour trouver des patterns dans des images afin de reconnaître des objets, des visages et des scènes. Ils peuvent également être très efficaces pour classer des données autres que des images, telles que des contenus audio, des séries temporelles et des signaux. [44]

3.14 les différentes couches d'un CNN

[39]
Il existe quatre types de couches pour un réseau de neurones convolutif : la couche de convolution,

la couche de pooling, la couche de correction ReLU et la couche fully-connected. Dans ce chapitre, je vais vous expliquer le fonctionnement de ces différentes couches.

3.14.1 La couche de convolution

La couche de convolution est la composante clé des réseaux de neurones convolutifs, et constitue toujours au moins leur première couche.

Son but est de repérer la présence d'un ensemble de features dans les images reçues en entrée. Pour cela, on réalise un filtrage par convolution : le principe est de faire "glisser" une fenêtre représentant la feature sur l'image, et de calculer le produit de convolution entre la feature et chaque portion de l'image balayée. Une feature est alors vue comme un filtre : les deux termes sont équivalents dans ce contexte.

3.14.2 La 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.

L'opération de pooling consiste à réduire la taille des images, tout en préservant leurs caractéristiques importantes.

Pour cela, on découpe l'image en cellules régulières, puis on garde au sein de chaque cellule la valeur maximale. En pratique, on utilise souvent des cellules carrées de petite taille pour ne pas perdre trop d'informations. Les choix les plus communs sont des cellules adjacentes de taille 2×2 pixels qui ne se chevauchent pas, ou des cellules de taille 3×3 pixels, distantes les unes des autres d'un pas de 2 pixels (qui se chevauchent donc).

On obtient en sortie le même nombre de feature maps qu'en entrée, mais celles-ci sont bien plus petites.

La couche de pooling permet de réduire le nombre de paramètres et de calculs dans le réseau. On améliore ainsi l'efficacité du réseau et on évite le sur-apprentissage.

Les valeurs maximales sont repérées de manière moins exacte dans les feature maps obtenues après pooling que dans celles reçues en entrée – c'est en fait un grand avantage ! En effet, lorsqu'on veut reconnaître un chien par exemple, ses oreilles n'ont pas besoin d'être localisées le plus précisément possible : savoir qu'elles se situent à peu près à côté de la tête suffit !

3.14.3 La couche de correction ReLU

ReLU (Rectified Linear Units) désigne la fonction réelle non-linéaire définie par ReLU(x)=max(0,x).

La couche de correction ReLU remplace donc toutes les valeurs négatives reçues en entrées par des zéros. Elle joue le rôle de fonction d'activation.

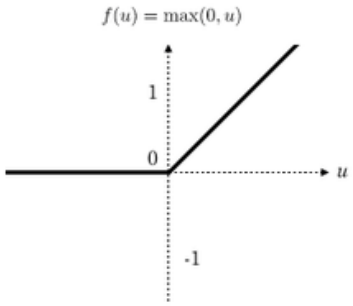

FIGURE 3.4: allure de la foction ReLU.

3.14.4 La couche fully-connected

La couche fully-connected constitue toujours la dernière couche d'un réseau de neurones, convolutif ou non – elle n'est donc pas caractéristique d'un CNN.

Ce type de couche reçoit un vecteur en entrée et produit un nouveau vecteur en sortie. Pour cela, elle applique une combinaison linéaire puis éventuellement une fonction d'activation aux valeurs reçues en entrée.

La dernière couche fully-connected permet de classifier l'image en entrée du réseau : elle renvoie un vecteur de taille N, où N est le nombre de classes dans notre problème de classification d'images. Chaque élément du vecteur indique la probabilité pour l'image en entrée d'appartenir à une classe.

Par exemple, si le problème consiste à distinguer les chats des chiens, le vecteur final sera de taille 2 : le premier élément (respectivement, le deuxième) donne la probabilité d'appartenir à la classe "chat" (respectivement "chien"). Ainsi, le vecteur [0.90.1] signifie que l'image a 90

Chaque valeur du tableau en entrée "vote" en faveur d'une classe. Les votes n'ont pas tous la même importance : la couche leur accorde des poids qui dépendent de l'élément du tableau et de la classe.

Pour calculer les probabilités, la couche fully-connected multiplie donc chaque élément en entrée par un poids, fait la somme, puis applique une fonction d'activation (logistique si N=2, softmax si N>2) :

Ce traitement revient à multiplier le vecteur en entrée par la matrice contenant les poids. Le fait que chaque valeur en entrée soit connectée avec toutes les valeurs en sortie explique le terme fully-connected.

3.15 RNN et LSTM

Les réseaux LSTM sont un peu plus complexe que les RNN « standards » en cela qu'ils ont plusieurs portes (gate), c'est à dire des couches qui permettent de modifier ou mémoriser le signal.

L'objectif des LSTM est de résoudre le problème du vanishing gradient c'est à dire quand les gradients diminuent trop vite au court du temps, rendant impossible la prise en compte de relation entre des instants lointains (aka. long termrelationships).
Ces nouvelles portes, étant des couches à part entière, ajoutent un certain nombre de paramètres à entraîner : 4x plus pour un LSTM, On pourrait penser qu'il s'agit donc d'un inconvenient. En pratique, cette augmentation est vraiment négligeable, par contre l'interêt des LSTM est bien réel. Par conséquent, dans le cas général il est vraiment plus interessant d'utiliser un LSTM.
Le seul contexte où je pourrais utiliser un RNN est éventuellement pour le comparer avec les performances sur LSTM. En effet, si les performances sont similaires, alors on pourrait déduire que la tâche apprise ne repose pas tant que ça sur des relations longue dans le temps. C'est une info qui peut être intéressante, mais c'est assez marginal . [25]

3.16 RNN VS CNN VS ANN

CNN est un réseau de neurones à feed forward généralement utilisé pour la reconnaissance d'images et la classification d'objets. Tandis que RNN fonctionne sur le principe de sauvegarder la sortie d'une couche et de la restituer à l'entrée afin de prédire la sortie de la couche.
CNN considère uniquement l'entrée actuelle, tandis que RNN considère l'entrée actuelle ainsi que les entrées précédemment reçues. Il peut mémoriser les entrées précédentes en raison de sa mémoire interne.
CNN a 4 couches, à savoir : couche de convolution, couche Relu, pooling et couche entièrement connectée. Chaque couche a sa propre fonctionnalité et effectue des extractions de caractéristiques et découvre des modèles cachés.
Il existe 4 types de RNN, à savoir : un à un, un à plusieurs, plusieurs à un et plusieurs à plusieurs.
RNN peut gérer des données séquentielles alors que CNN ne le peut pas.
Dans RNN, les états précédents sont alimentés en entrée de l'état actuel du réseau. RNN peut être utilisé en PNL,
. . [54]

3.17 Conclusion

L'apprentissage profond est le domaine le plus émergeant de l'apprentissage automatique et a apporté une contribution importante dans divers domaines de recherche. Cela a permis de surmonter les inconvénients des méthodes traditionnelles en rendant les systèmes moins complexes et plus rapides. L'apprentissage profond a été utilisé avec le traitement automatique du langage dans plusieurs domaines de recherche, ce qui est très prometteur et constitue un succès. Dans ce chapitre nous avons exposé c'est quoi le deep learning , ainsi que ses avantages, et les domaines d'application, et le réseau de neurones , enfin le réseau de neurones récurrents et Convolutional Neural Network . ..

CHAPITRE 4

LA RÉALISATION

4.1 Introduction

La Reconnaissance vocale des bruits environnementaux est un domaine de recherche en pleine croissance avec de nombreuses applications dans le monde réel. Bien qu'il existe un grand nombre de recherches dans des domaines audio connexes tels que la parole et la musique, les travaux sur la classification des sons environnementaux sont relativement rares. De même, en observant les récents progrès dans le domaine de la classification d'images où les réseaux de neurones convolutifs sont utilisés pour classer des images avec une grande précision et à grande échelle, cela soulève la question de l'applicabilité de ces techniques dans d'autres domaines, tels que la classification sonore.
Il existe une pléthore d'applications dans le monde réel pour cette recherche, telles que :
• Indexation et récupération multimédia basées sur le contenu
• Aide aux personnes sourdes dans leurs activités quotidiennes
• Cas d'utilisation de la maison intelligente tels que les capacités de sûreté et de sécurité à 360 degrés
• Utilisations industrielles telles que la maintenance prédictive

4.2 Les outils de réalisation

4.2.1 Python

Python est un langage de programmation de haut niveau interprété pour la programmation à usage général. Créé par Guido van Rossum et publié pour la première fois en 1991, Python

repose sur une philosophie de conception qui met l'accent sur la lisibilité du code, notamment en utilisant des espaces significatifs. Il fournit des constructions permettant une programmation claire à petite et grande échelle. En juillet 2018, Van Rossum a démissionné en tant que leader de la communauté après 30 ans.

Python propose un système de typage dynamique et une gestion automatique de la mémoire. Il prend en charge plusieurs paradigmes de programmation, notamment orienté objet, impératif, fonctionnel et procédural, et dispose d'une bibliothèque standard étendue et complète.

Les interpréteurs Python sont disponibles pour de nombreux systèmes d'exploitation. CPython, l'implémentation de référence de Python, est un logiciel open source et dispose d'un modèle de développement basé sur la communauté, comme le font presque toutes les autres implémentations de Python. Python et CPython sont gérés par l'association à but non lucratif Python Software Foundation.

[57]

FIGURE 4.1: logo Python

4.2.2 Jupyter Notebook

Jupyter Notebook est un outil open source permettant d'écrire du code informatique et de le partager pour collaborer. Grâce à ses nombreux avantages, ce " bloc-note " de calcul est devenu une référence incontournable pour les Data Scientists [7]

Jupyter a été créé pour faciliter la présentation du travail de programmation d'un développeur et permettre à d'autres d'y participer. Il permet de mélanger du code, des commentaires et des visualisations dans un document interactif appelé notebook qui peut être partagé, réutilisé et retravaillé. Et comme Jupyter Notebook s'exécute dans un navigateur web, le « cahier » lui-même peut être hébergé au choix sur l'ordinateur du développeur ou sur un serveur distant.

[56]

FIGURE 4.2: logo jupyter

4.3 Base de données

Pour cela, nous utiliserons un ensemble de données appelé Urbansound8K. L'ensemble de données contient 8732 extraits sonores (<=4s) de sons urbains de 10 classes, qui sont :

- Climatiseur
- Klaxon de voiture
- Enfants jouant
- Aboiement de chien
- Forage
- Moteur au ralenti
- Coup de feu
- Marteau-piqueur
- Sirène
- Musique de rue

4.4 Présentation du fichier audio

Ces extraits sonores sont des fichiers audio numériques au format .wav. Les ondes sonores sont numérisées en les échantillonnant à des intervalles discrets appelés taux d'échantillonnage (généralement 44,1 kHz pour un son de qualité CD, ce qui signifie que les échantillons sont prélevés 44 100 fois par seconde). Chaque échantillon est l'amplitude de l'onde à un intervalle de temps particulier, où la profondeur de bits détermine le degré de détail de l'échantillon sera également connu sous le nom de plage dynamique du signal (généralement 16 bits, ce qui signifie qu'un échantillon peut aller de 65 536 valeurs d'amplitude).

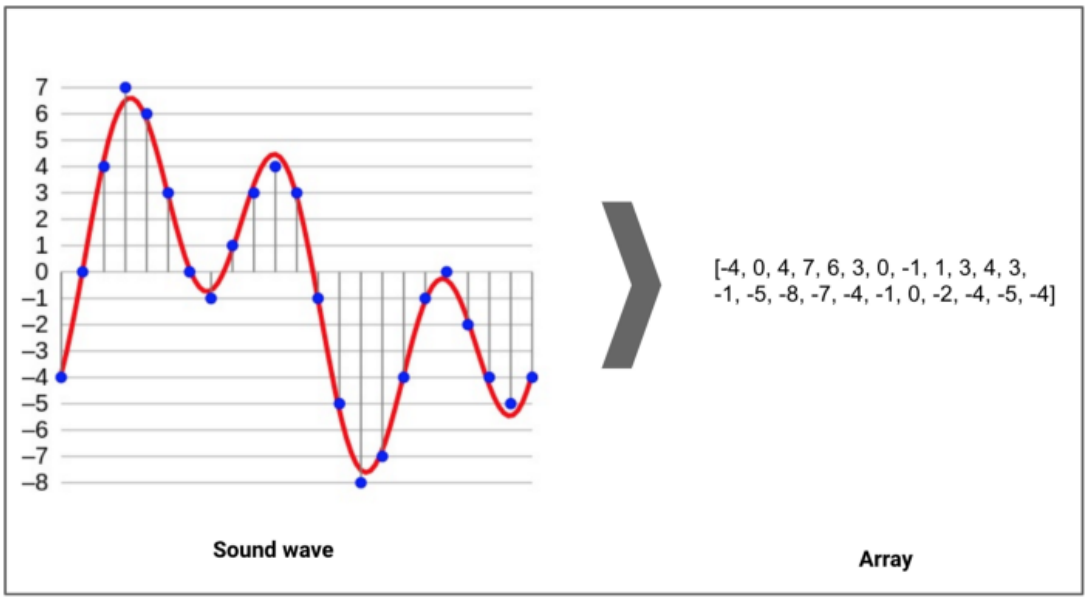

FIGURE 4.3: Présentation du fichier audio

Une onde sonore, en rouge, représentée numériquement, en bleu (après échantillonnage et quantification 4 bits), avec le tableau résultant affiché à droite

L'image ci-dessus montre comment un extrait sonore est extrait d'une forme d'onde et transformé en un tableau ou un vecteur unidimensionnel de valeurs d'amplitude.

4.5 Exploration des données

A partir d'une inspection visuelle, nous pouvons voir qu'il est difficile de visualiser la différence entre certaines des classes. En particulier, les formes d'onde des sons répétitifs pour le climatiseur, le forage, le moteur au ralenti et le marteau-piqueur sont de forme similaire.

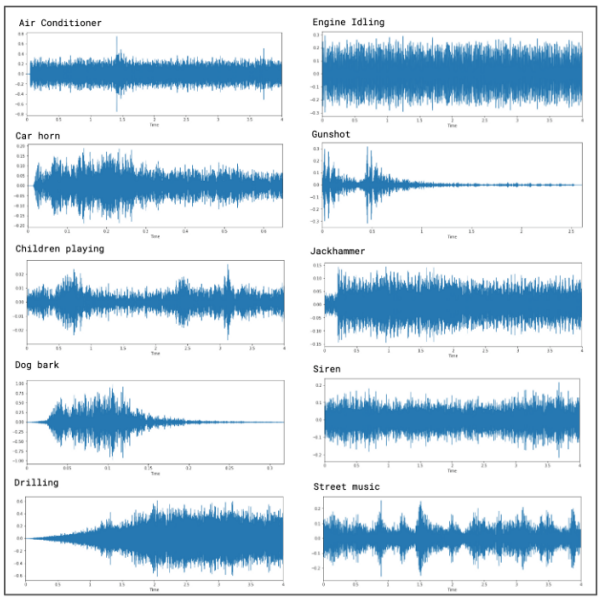

FIGURE 4.4: les formes des sons répétitifs

Ensuite, nous allons approfondir l'extraction des propriétés de chacun des fichiers audio, le nombre de canaux audio, la fréquence d'échantillonnage et la profondeur de bits à l'aide du code suivant.

```
# Load various imports
import pandas as pd
import os
import librosa
import librosa.display

from helpers.wavfilehelper import WavFileHelper
wavfilehelper = WavFileHelper()

audiodata = []
for index, row in metadata.iterrows():

    file_name = os.path.join(os.path.abspath('/UrbanSound8K/audio/'),'fold'+str(row["fold"])
    data = wavfilehelper.read_file_properties(file_name)
    audiodata.append(data)

# Convert into a Panda dataframe
audiodf = pd.DataFrame(audiodata, columns=['num_channels','sample_rate','bit_depth'])
```

FIGURE 4.5: code de l'extraction des propriétés de chacun des fichiers audio

```python
import struct

class WavFileHelper():

    def read_file_properties(self, filename):

        wave_file = open(filename,"rb")

        riff = wave_file.read(12)
        fmt = wave_file.read(36)

        num_channels_string = fmt[10:12]
        num_channels = struct.unpack('<H', num_channels_string)[0]

        sample_rate_string = fmt[12:16]
        sample_rate = struct.unpack("<I",sample_rate_string)[0]

        bit_depth_string = fmt[22:24]
        bit_depth = struct.unpack("<H",bit_depth_string)[0]

        return (num_channels, sample_rate, bit_depth)
```

FIGURE 4.6: code de l'extraction des propriétés de chacun des fichiers audio

Ici, nous pouvons voir que l'ensemble de données a une gamme de propriétés audio variables qui devront être standardisées avant de pouvoir l'utiliser pour entraîner notre modèle.

Canaux audio la plupart des échantillons ont deux canaux audio (c'est-à-dire stéréo) avec quelques-uns avec un seul canal (mono).

```
# num of channels
print(audiodf.num_channels.value_counts(normali

2    0.915369
1    0.084631
```

FIGURE 4.7: Canaux audio

Fréquence d'échantillonnage Il existe une large gamme de fréquences d'échantillonnage qui ont été utilisées pour tous les échantillons, ce qui est préoccupant (allant de 96 kHz à 8 kHz).

```
# sample rates

print(audiodf.sample_rate.value_counts(normalize=True))

44100     0.614979
48000     0.286532
96000     0.069858
24000     0.009391
16000     0.005153
22050     0.005039
11025     0.004466
192000    0.001947
8000      0.001374
11024     0.000802
32000     0.000458
```

FIGURE 4.8: Fréquence d'échantillonnage

Profondeur de bits, il existe également une gamme de profondeurs de bits (allant de 4 bits à 32 bits).

```
# bit depth
print(audiodf.bit_depth.value_counts(normalize=True))

16    0.659414
24    0.315277
32    0.019354
8     0.004924
4     0.001031
```

FIGURE 4.9: Profondeur de bits

4.6 Pré-traitement des données

Dans la section précédente, nous avons identifié les propriétés audio suivantes qui nécessitent un prétraitement pour assurer la cohérence dans l'ensemble de l'ensemble de données :

- Canaux audio
- Taux d'échantillonnage

• Profondeur de bits

Pandas est une bibliothèque écrite pour le langage de programmation Python permettant la manipulation et l'analyse des données. Elle propose en particulier des structures de données et des opérations de manipulation de tableaux numériques et de séries temporelles. [36]

Le module os est un module fournit par Python dont le but d'interagir avec le système d'exploitation, il permet ainsi de gérer l'arborescence des fichiers, de fournir des informations sur le système d'exploitation processus, variables systèmes, ainsi que de nombreuses fonctionnalités du systèmes [58]

Le module struct permet de manipuler des données agrégées sous forme binaire dans des fichiers ou à travers des connecteurs réseau. Il utilise Chaînes de spécification du format comme description de l'agencement des structures afin de réaliser les conversions depuis et vers les valeurs Python.

Librosa est un package Python pour le traitement de la musique et de l'audio de Brian McFee et nous permettra de charger l'audio dans notre ordinateur portable sous forme de tableau numpy pour l'analyse et la manipulation.

Pour une grande partie du prétraitement, nous pourrons utiliser la fonction load() de Librosa , qui convertit par défaut le taux d'échantillonnage à 22,05 KHz, normaliser les données de sorte que les valeurs de profondeur de bits se situent entre -1 et 1 et aplatir les canaux audio en mono

4.7 Extraire des fonctionnalités

L'étape suivante consiste à extraire les fonctionnalités dont nous aurons besoin pour entraîner notre modèle. Pour ce faire, nous allons créer une représentation visuelle de chacun des échantillons audio qui nous permettra d'identifier les caractéristiques pour la classification, en utilisant les mêmes techniques que celles utilisées pour classer les images avec une grande précision.

Les spectrogrammes sont une technique utile pour visualiser le spectre des fréquences d'un son et leur variation sur une très courte période de temps. Nous utiliserons une technique similaire connue sous le nom de coefficients cepstraux Mel-Frequency (MFCC) .

La principale différence est qu'un spectrogramme utilise une échelle de fréquence espacée linéaire (de sorte que chaque case de fréquence est espacée d'un nombre égal de Hertz), alors qu'un MFCC utilise une échelle de fréquence espacée quasi-logarithmique , qui est plus similaire à la façon dont le système auditif humain traite les sons.

L'image ci-dessous compare trois représentations visuelles différentes d'une onde sonore, la première étant la représentation dans le domaine temporel, comparant l'amplitude au fil du temps. Le suivant est un spectrogramme montrant l'énergie dans différentes bandes de fréquences changeant au fil du temps, puis enfin un MFCC que nous pouvons voir est très similaire à un spectrogramme mais avec des détails plus distinguables.

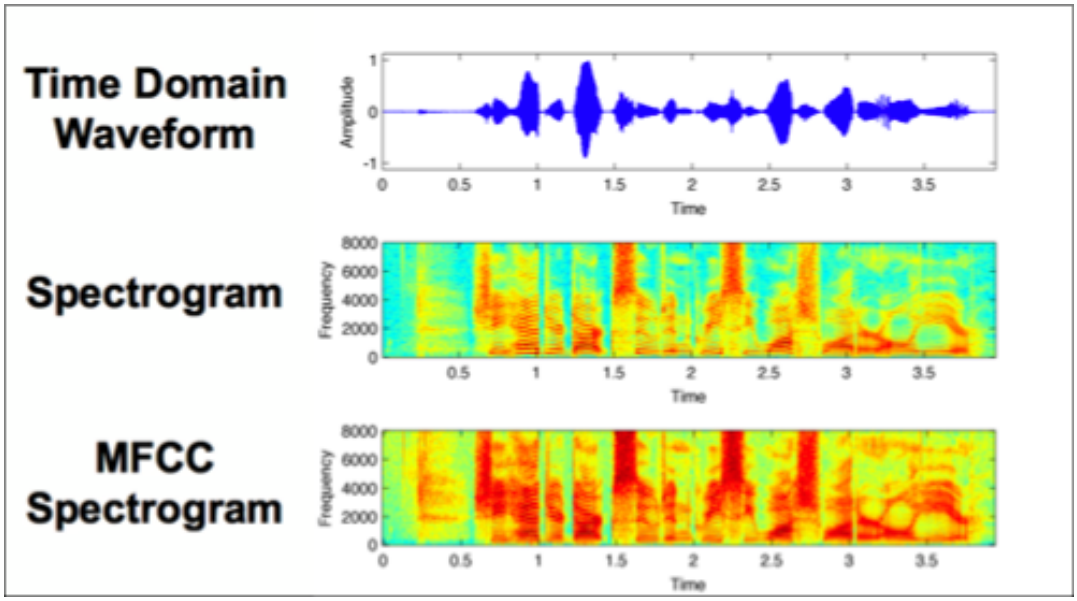

FIGURE 4.10: comparison entre trois représentations visuelles différentes d'une onde sonore

Pour chaque fichier audio de l'ensemble de données, nous extrairons un MFCC (ce qui signifie que nous avons une représentation d'image pour chaque échantillon audio) et le stockerons dans un Panda Dataframe avec son étiquette de classification. Pour cela, nous utiliserons la fonction mfcc() de Librosa qui génère un MFCC à partir de données audio de séries temporelles.

```
def extract_features(file_name):

    try:
        audio, sample_rate = librosa.load(file_name, res_type='kaiser_fast')
        mfccs = librosa.feature.mfcc(y=audio, sr=sample_rate, n_mfcc=40)
        mfccsscaled = np.mean(mfccs.T,axis=0)

    except Exception as e:
        print("Error encountered while parsing file: ", file)
        return None

    return mfccsscaled


# Load various imports
import pandas as pd
import os
import librosa

# Set the path to the full UrbanSound dataset
fulldatasetpath = '/Urban Sound/UrbanSound8K/audio/'

metadata = pd.read_csv(fulldatasetpath + '../metadata/UrbanSound8K.csv')
```

FIGURE 4.11: code pour extrairons un MFCC

```
features = []

# Iterate through each sound file and extract the features
for index, row in metadata.iterrows():

    file_name = os.path.join(os.path.abspath(fulldatasetpath),'fold'+str(row["fold"])+'/'

    class_label = row["class_name"]
    data = extract_features(file_name)

    features.append([data, class_label])

# Convert into a Panda dataframe
featuresdf = pd.DataFrame(features, columns=['feature','class_label'])

print('Finished feature extraction from ', len(featuresdf), ' files')
```

FIGURE 4.12: code pour extrairons un MFCC

4.8 Conversion des données et des étiquettes, puis fractionnement de l'ensemble de données

Nous utiliserons "sklearn.preprocessing.LabelEncoder " pour coder les données textuelles catégorielles en données numériques compréhensibles par le modèle.
Ensuite, nous utiliserons "sklearn.model_selection.train_test_split"pour diviser l'ensemble de données en ensembles d'apprentissage et de test. La taille de l'ensemble de test sera de 20 %. et nous définirons un état aléatoire.

```
from sklearn.preprocessing import LabelEncoder
from keras.utils import to_categorical

# Convert features and corresponding classification labels into numpy arrays
X = np.array(featuresdf.feature.tolist())
y = np.array(featuresdf.class_label.tolist())

# Encode the classification labels
le = LabelEncoder()
yy = to_categorical(le.fit_transform(y))

# split the dataset
from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test = train_test_split(X, yy, test_size=0.2, random_state = 42)
```

FIGURE 4.13: Conversion des données et des étiquettes puis fractionnement de l'ensemble de données

4.9 Construire notre modèle

La prochaine étape consistera à créer et à entraîner un réseau de neurones profonds avec ces ensembles de données et à faire des prédictions.
Ici, nous utiliserons un réseau de neurones convolutifs (CNN). Les CNN font généralement de bons classificateurs et fonctionnent particulièrement bien avec les tâches de classification d'images en raison de leurs parties d'extraction de caractéristiques et de classification. Je pense que cela sera très efficace pour trouver des modèles dans les MFCC, tout comme ils sont efficaces pour trouver des modèles dans les images.
Nous utiliserons un modèle séquentiel, en commençant par une architecture de modèle simple, composée de quatre couches de convolution Conv2D, notre couche de sortie finale étant une couche dense. Notre couche de sortie aura 10 nœuds (num_labels) qui correspondent au nombre de classifications possibles.

```
import numpy as np
from keras.models import Sequential
from keras.layers import Dense, Dropout, Activation, Flatten
from keras.layers import Convolution2D, Conv2D, MaxPooling2D, GlobalAveragePooling2D
from keras.optimizers import Adam
from keras.utils import np_utils
from sklearn import metrics

num_rows = 40
num_columns = 174
num_channels = 1

x_train = x_train.reshape(x_train.shape[0], num_rows, num_columns, num_channels)
x_test = x_test.reshape(x_test.shape[0], num_rows, num_columns, num_channels)

num_labels = yy.shape[1]
filter_size = 2
```

FIGURE 4.14: Construire notre modèle

```
# Construct model
model = Sequential()
model.add(Conv2D(filters=16, kernel_size=2, input_shape=(num_rows, num_columns, num_channels),
model.add(MaxPooling2D(pool_size=2))
model.add(Dropout(0.2))

model.add(Conv2D(filters=32, kernel_size=2, activation='relu'))
model.add(MaxPooling2D(pool_size=2))
model.add(Dropout(0.2))

model.add(Conv2D(filters=64, kernel_size=2, activation='relu'))
model.add(MaxPooling2D(pool_size=2))
model.add(Dropout(0.2))

model.add(Conv2D(filters=128, kernel_size=2, activation='relu'))
model.add(MaxPooling2D(pool_size=2))
model.add(Dropout(0.2))
model.add(GlobalAveragePooling2D())

model.add(Dense(num_labels, activation='softmax'))
```

FIGURE 4.15: Construire notre modèle

Pour compiler notre modèle, nous utiliserons les trois paramètres suivants :

```
# Compile the model
model.compile(loss='categorical_crossentropy', metrics=['accuracy'], optimizer='adam')

# Display model architecture summary
model.summary()

# Calculate pre-training accuracy
score = model.evaluate(x_test, y_test, verbose=1)
accuracy = 100*score[1]

print("Pre-training accuracy: %.4f%%" % accuracy)
```

FIGURE 4.16: compiler notre modèle

Ici, nous allons entraîner le modèle. Comme la formation d'un CNN peut prendre beaucoup de temps, nous allons commencer avec un faible nombre d'époques et une faible taille de lot. Si nous pouvons voir à partir de la sortie que le modèle converge, nous augmenterons les deux nombres.

```
from keras.callbacks import ModelCheckpoint
from datetime import datetime

num_epochs = 72
num_batch_size = 256

checkpointer = ModelCheckpoint(filepath='saved_models/weights.best.basic_cnn.hdf5',
                               verbose=1, save_best_only=True)
start = datetime.now()

model.fit(x_train, y_train, batch_size=num_batch_size, epochs=num_epochs, validation_data=


duration = datetime.now() - start
print("Training completed in time: ", duration)
```

FIGURE 4.17: entraîner le modèle

Ce qui suit examinera la précision du modèle sur les ensembles de données d'apprentissage et de test.

```
# Evaluating the model on the training and testing set
score = model.evaluate(x_train, y_train, verbose=0)
print("Training Accuracy: ", score[1])

score = model.evaluate(x_test, y_test, verbose=0)
print("Testing Accuracy: ", score[1])
```

FIGURE 4.18: examinera la précision du modèle

4.10 Résultats

Notre modèle entraîné a obtenu une précision d'entraînement de 98,19 % et une précision de test de 91,92%.
Les performances sont très bonnes et le modèle s'est bien généralisé, semblant bien prédire lorsqu'il est testé par rapport à de nouvelles données audio.

```
Training completed in time:  0:02:15.519198


_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_1 (Dense)              (None, 256)               10496
_________________________________________________________________
activation_1 (Activation)    (None, 256)               0
_________________________________________________________________
dropout_1 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_2 (Dense)              (None, 256)               65792
_________________________________________________________________
activation_2 (Activation)    (None, 256)               0
_________________________________________________________________
dropout_2 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_3 (Dense)              (None, 10)                2570
_________________________________________________________________
activation_3 (Activation)    (None, 10)                0
=================================================================
Total params: 78,858
Trainable params: 78,858
Non-trainable params: 0
_________________________________________________________________
Pre-training accuracy: 13.3371%
```

FIGURE 4.19: Nombre de paramètres par couche

Figure 4.19 présente le nombre de paramètres entrainés dans chaque couche d'où on voix

que la deuxième couche dense entraine la majorité des paramètres (65792/78858) alors que la première couche dense entraine 10436 paramètres.

4.11 Conclusion

Le processus utilisé pour ce projet peut être résumé avec les étapes suivantes :

1. Le problème initial a été défini et l'ensemble de données public pertinent a été localisé.
2. Les données ont été explorées et analysées.
3. Les données ont été prétraitées et les caractéristiques extraites.
4. Un premier modèle a été formé et évalué.
5. Le modèle final a été évalué.

Dès l'exploration initiale des données à l'étape 2, nous avons envisagé que le travail de prétraitement à l'étape 3 prendrait énormément de temps. Cependant, cela était en fait relativement facile grâce à la Outil Python Librosa. Nous pensons aussi que l'extraction de caractéristiques serait beaucoup plus délicate mais encore une fois Librosa a raccourci énormément l'effort requis. Les MFCC que nous avons extraits à l'étape 3 fonctionnent bien mieux que ce à quoi je m'attendais. Cependant, nous avons dû revisiter le processus d'extraction lorsque nous sommes passés à l'utilisation d'un CNN comme modèle.

Conclusion générale

Ce mémoire s'inscrit dans la volonté de reconnaître le son et, plus précisément, de détecter les coups de feu dans le bruit de la ville.

L'approche proposée est basée, dans un premier temps, sur le traitement de signal auditif à partir de différents endroits Pour améliorer les performances du système reconnaissance vocale .

Les chapitres précédents ont exposé le domaine de deep learning. La première partie de ce traité est consacrée à introduire peu à peu le lecteur à La base de nos recherches,les modele cnn et rnn , pour produire et analyser un signal de son qui donne Les fondements de base de classification audio

Amélioration

Si nous devions poursuivre ce projet, il y a un certain nombre de domaines supplémentaires qui pourraient être exploré :

• Comme mentionné précédemment, testez les performances des modèles avec l'audio en temps réel.

• Former le modèle pour les données du monde réel. Cela impliquerait probablement d'augmenter les données de formation de diverses manières telles que :

- Ajout d'une variété de sons de fond différents.

– Réglage des niveaux de volume du son cible ou ajout d'échos.

– Modification de la position de départ de l'échantillon d'enregistrement, par ex. la forme d'un aboiement de chien.

• Expérimentez pour voir si la précision par classe est affectée par l'utilisation de données d'entraînement de durées différentes.

• Expérimentez avec d'autres techniques d'extraction de caractéristiques telles que différentes formes de spectrogrammes.

BIBLIOGRAPHIE

[1] *Réseaux de neurones récurrents pour le traitement automatique de la parole ,thèse de doctorat de l'université paris-saclay préparée à l'université paris-sud*.

[2] *Le machine learning et deep learning*, (28 AOÛT 2020).

[3] *Logiciel très avancé de reconnaissance vocale*, (9 Février 2010).

[4] AUTHOT, *La reconnaissance vocale dans le monde mÉdical*, (04/09/2020).

[5] BASTIEN, *Comprendre le deeplearning – 2 / 3 : Fonctionnement*, (13 novembre 2019).

[6] ——, *Introduction aux réseaux de neurones – 3 / 3 : Apprentissage des réseaux de neurones*, (15 novembre 2019).

[7] L. BASTIEN, *Jupyter notebook : tout savoir sur le notebook préféré des data scientists*, (16 mars 2021).

[8] D. BELHAOUCI, *Doctorant en droit et responsable du développement de juri'predis*.

[9] S. BERTRAND-GASTALDY, *Université de montréal, dans le cadre du cours blt 6134 - analyse de textes et ordinateur -*, (2000).

[10] G. G. . F. BIMBOT, *la reconnaissance automatique du locuteur à la signature vocale*, (19/03/2007).

[11] G. G. F. BIMBOT, *la reconnaissance automatique du locuteur à la signature vocale*, (19/03/2007).

[12] M. BOUDRAA, *Thèse de doctorat sur : Reconnaissance automatique du locuteur*, (Septembre 2003).

[13] A. L. B. CHAIX, *ReprÉsentation du son*, (27/12/2016).

[14] E. CHAKER, *Systèmes et signaux*, (31/10/2014).

[15] S. CHRISTEL, *Le traitement du signal vocal*, (January 1995).

[16] J. K. DAS, *Urban Sound Classification Using Convolutional Neural Network and Long Short Term Memory Based on Multiple Features*, PhD thesis, November 2020.

[17] DELPHINE BOLUS, *Biométrie vocale et paiement via les assistants vocaux*, (29 mars, 2018).

[18] DELPHINE BOLUS, *Biométrie vocale et paiement via les assistants vocaux*, (29 mars, 2018).

[19] A. V. D'INFORMATION, *Origin-stt, l'application qui rend accessibles vos contenus audio*, (06/07/2020).

[20] G. P. (DIR), *Panorama de la physique*, (2007).

[21] Y. M. DJALOUL, *"deep learning pour la classification des images", mémoire de master, université de tlemcen, algérie*, (2017).

[22] DUTERTRE, *Conversions analogique - numérique et numérique - analogique.*, (10/08/2009).

[23] EDUSCOL, *Signal et information*, (Mars 2016).

[24] ——, *Signal et information*, (Mars 2016).

[25] J. P. . A. GIBSON, *Deep learning a practitioner's approach,1ère (ed),o'reilly media, inc., 1005 gravenstein highway north, sebastopol, ca 95472 , mike loukides timmcgovern, 532 p*, (2017).

[26] Y. GOLDBERG, *A primer on neural network models for natural language processing*, (2015).

[27] T. HERVÉ, *Les enjeux de la sécurité informatique*, PhD thesis, Haute école de gestion de Genève, 2011.

[28] Z. ISMAILI, *Machine learning ou apprentissage automatique*.

[29] JOEL DRAKES, *Responsable avant-vente - nuance communications dans les echos*.

[30] B. KARIMA, *Système sécurisé à base vocale -mémoire master en réseaux et système distribué -université abou bakr belkaid, tlemcen*, (23 Juin 2015).

[31] W. L. KOONTZ, *Introduction au traitement du signal audio*, (12/2016).

[32] D. M. LE DIPLOME, *Caractéristiques Biométrique pour l'identification*, PhD thesis, Université d'Oran, 2016.

[33] A. F. M. VACHER, F. PORTET AND N. NOURY., *Development of audio sensing technology for ambient assisted living : Applications and challenges. international journal of e-health and medical communications*, (march 2011).

[34] D. MALOWANY, *Classification audio avec les outils écosystémiques de pytorch*, (18 OCTOBRE 2020).

[35] D. C. MATHILDE GLÉNAT, *Principe du passage de l'analogique au numérique*, (29/06/2012).

[36] W. MCKINNEY, *Pandas*, (12 avril 2021).

[37] L. MEGHRAOUA, *Reconnaissance vocale : parle et je te dirai qui tu es*, (18 août 2019).

[38] E. MOUTOT, *La numÉrisation du signal audio*, (juin 2020).

[39] K. NADJAHI, *Découvrez les différentes couches d'un cnn*, (05/05/2021).

[40] A. NUTTINCK, *Les applications du deep learning*, (7/8/2018).

[41] G. PFOTZER, *Ingénieur informatique - c.n.a.m -reconnaissance vocale*.

[42] S. ROUQUETTE, *connaissance vocale, la cnil dit oui à michelin*.

[43] H. SAYOUD, *Thèse de doctorat sur reconnaissance automatique du locuteur approche connexionniste.*, (2003).

[44] SITE WEB, *https :/ /fr.mathworks.com /discovery /convolutional-neural-network-matlab.html*.

[45] (SITE WEB), *https :/ /fr.mathworks.com /discovery /deep-learning.html.*

[46] ——, *https :/ /fr.mathworks.com /discovery /deep-learning.html.*

[47] SITE WEB, *https :/ /link.springer.com /article /*.

[48] (SITE WEB), *https :/ /www.ionos.fr /digitalguide /web-marketing /search-engine-marketing /deep-learning /* .

[49] J.-P. STROMBONI., *Numériser le signal audio.*., (2/03/2005).

[50] VIVOKA, *La biométrie vocale est-elle un processus vraiment fiable ?*, (25 février 2021).

[51] W. G. W CHRISTOPHER ., S GERASIMOS, *., flexible deep neural network structure with application to natural language processing. department of knowledge engineering, maastricht university, the netherlands.*

[52] S. WEB, *https :/ /www.biometrie-online.net /technologies /voix*.

[53] ——, *https :/ /www.onelogin.com /fr /learn /biometric-authentication*.

[54] S. WEB), *https :/ /fr.quora.com /quelles-sont-les-diff %c3 %a9rences-entre-les-cas-dutilisation-des-ann-cnn-et-rnn*, (9 mars 2021).

[55] M. YASSINE, *These sur la reconnaisance de locuteurs par localisation dans un espace de locuteurs de référence - ecole nationale supérrieur des télécommunications,*, (octobre 2003).

[56] S. YEGULALP, *Comment jupyter notebook facilite l'analyse de données*, (11 Mars 2019).

[57] D. YOUNES, *Introduction au langage python*, (21 octobre 2018).

[58] ——, *Le module os en python*, (29 juillet 2019).