الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire

وزارة التعليم العالي والبحث العلمي
جامعة أكلي محند أولحاج
-البويرة-

Ministère de l'Enseignement Supérieur
et de la Recherche Scientifique
Université Akli Mohand Oulhadj - Bouira -
Tasdawit Akli Muḥend Ulḥağ - Tubirett -

جامعة البويرة

كلية العلوم والعلوم التطبيقية
Faculté des Sciences et des Sciences Appliquées

المرجع:...../م م/ 2023
Référence : ...../MM/ 2023

Mémoire de Master

Présenté au

Département : Génie Électrique
Domaine : Sciences et Technologies
Filière : Télécommunications
Spécialité : Systèmes des Télécommunications

Réalisé par :

AISSAOUI Amira **Et** **ZERROUKI Achouak**

Thème

Reconnaissance des émotions vocales basée sur l'apprentissage profond

Soutenu le : **02/07/2023**

Devant le Jury composé de :

Dr:	MEDJEDOUB ISMAIL	M.A.A	Univ. Bouira	Président
Dr :	ABDENNOUR ALIMOHAD	M.C.B	Univ. Bouira	Rapporteur
Dr :	NOURINE	M.C.A	Univ. Bouira	Examinateur

Année Universitaire : 2022-2023

الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire

وزارة التعليم العالي و البحث العلمي
جامعة أكلي محند أولحاج - البويرة

Ministère de l'Enseignement Supérieur
et de la Recherche Scientifique
Université Akli Mohand Oulhadj - Bouira
Tasdawit Akli Muḥend Ulḥağ - Tubirett

نموذج التصريح الشرفي الخاص بالالتزام بقواعد النزاهة العلمية لإنجاز بحث

انا الممضي اسفله،

السيد(ة) عيساوي أميرة الصفة: طالب، استاذ باحث Ali Mohad

الحامل(ة) لبطاقة التعريف الوطنية: 12027 1751 والصادرة بتاريخ 2021/4/7

المسجل(ة) بكلية: العلوم و العلوم التطبيقية قسم: الهندسة الكهربائية

والمكلف(ة) بإنجاز أعمال بحث(مذكرة، التخرج، مذكرة ماستر، مذكرة ماجستير، أطروحة دكتوراه).

عنوانها: Reconnaissance ~~Vocal~~ des émotions Vocal basée sur l'apprentissage Profond

تحت إشراف الأستاذ(ة): Ali Mohad Abd el nour

أصرح بشرفي اني ألتزم بمراعاة المعايير العلمية والمنهجية الاخلاقيات المهنية والنزاهة الاكاديمية المطلوبة في انجاز البحث المذكور أعلاه.

التاريخ: 2023/09/07

توقيع المعني(ة)

رأي هيئة مراقبة السرقة العلمية:

النسبة: itin % 11

الأمضاء:

الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire

وزارة التعليم العالي و البحث العلمي
جامعة أكلي محند أولحاج - البويرة

Ministère de l'Enseignement Supérieur
et de la Recherche Scientifique
Université Akli Mohand Oulhadj - Bouira
Tasdawit Akli Muḥend Ulḥağ - Tubirett

نموذج التصريح الشرفي الخاص بالالتزام بقواعد النزاهة العلمية لإنجاز بحث

انا الممضي اسفله،

السيد(ة) زروقي أسماء الصفة: طالب، استاذ باحث Ali Mohad

الحامل(ة) لبطاقة التعريف الوطنية: 400621 2F 1 والصادرة بتاريخ 2022,02.12

المسجل(ة) بكلية :العلوم و التكنولوجيا قسم: الهندسة الكهربائية

والمكلف(ة) بإنجاز اعمال بحث(مذكرة، التخرج، مذكرة ماستر، مذكرة ماجستير، اطروحة دكتوراه).

عنوانها: Reconnaissance des émotions locales basée sur l'apprentissage profond

تحت إشراف الأستاذ(ة): ABDENNOUR ALI MOHAD

أصرح بشرفي اني ألتزم بمراعاة المعايير العلمية والمنهجية الاخلاقيات المهنية والنزاهة الاكاديمية المطلوبة في انجاز البحث المذكور أعلاه.

التاريخ: 09/07/2023

توقيع المعني(ة)

رأي هيئة مراقبة السرقة العلمية:

النسبة: itin % 11

الامضاء:

Dédicaces 1

Je dédie

Ce modeste travail comme un témoignage d'affectation, de respect et d'admiration

A mon cher père ***Rabah***

L'homme qui m'a toujours encouragé à Poursuivre Mes études, source de tendresse. Tous les mots ne Pourront exprimer L'amour que je te porte. Tu es mon bonheur Et ma raison de vivre. Qu'dieu te garde pour moi pour toujours.

A *ma chère mère* ***Aicha***

En ce jour spécial, je voudrais dédier ces mots d'amour et de gratitude à ma mère. Tu es bien plus qu'une mère pour moi, tu es mon soutien indéfectible, ma meilleure amie. Depuis le premier jour où j'ai ouvert les yeux sur ce monde, tu as été là, guidant mes pas et illuminant mon chemin.

Que dieu garde mes parents et je les souhaite plus de bonheur et de joie dans leur vie.

A ma seule sœur ***Meriem***

Qui a partagé avec moi tous les moments d'émotion et d'amour lors de la réalisation de travail

A mes chers frères ***Alla*** *et* ***ABD EL Malek***

Pour leur aide tout en long de mon chemin, grâce à leur amour, leur compréhension et leur patience.

A mon fiancé ***OMAR***

Qui était présent à mes côtés durant tout au long de ce travail et ma guide et encourager à tout moment.

A ma famille et ma belle-famille ceux qui me donnent de l'amour et l'encouragement durant tout l'année, Que Dieu leur donne une longue et joyeuse vie.

A mes chéries amies qui je n'oublierais jamais. A toute personne de ma promotion que je leurs souhaite plein de succès dans le monde professionnel A tous ceux que j'aime.

A ma chère binômes ***Achouak***

Pour son amour, son compréhension et son aide pour terminer ce travail.

Amira

Dédicaces 2

Je dédie

Ce modeste travail comme un témoignage d'affectation, de respect et d'admiration.

A mon cher père ***Ahmed***

Qui n'a pas cessée de me conseiller, encourager et soutenir tout au long de mes études et qui a été derrière moi pour me guider dans la bonne direction pour que je puisse atteindre mes objectifs.

A ma chère mère ***Saliha***

La femme qui m'a toujours encouragé à Poursuivre Mes études, source de tendresse. Tous les mots ne Pourront exprimer L'amour que je te porte. Tu es mon bonheur Et ma raison de vivre.

Qu'dieu te garde pour moi pour toujours

A mes chers frères ***Aymen*** *et* ***Mohamed***

Pour leur aide tout en long de mon chemin, grâce à leur amour, leur compréhension et leur patience.

A mes amies et mes sœurs ***Ikram, Bouchera, Asma, Amira, Maissa, et Ikram*** *Je ne peux trouver les mots justes et sincères pour vous exprimer mon affection et mes pensées, vous êtes pour moi des sœurs et des amies sur qui je peux compter. Merci pour m'avoir toujours supporté dans mes décisions. Merci pour tout votre amour et votre confiance, pour m'avoir aidé à ranger mon éternel désordre et pour votre énorme support pendant la rédaction de mon projet.*

Je vous aime beaucoup.

A tous les membre de ma famille qui m'ont soutenu dans mon cheminement scolaire.

Je dédie mon diplôme et ma réussite à mes sœurs ***Donia*** *et* ***Hafsa*** *et à ma grand-mère, que Dieu ait pitié d'elles, qui ont toujours souhaité que leurs yeux soient heureux de me voir un jour comme celui-ci, à celle qui était couverte de terre avant que son souhait ne soit accomplie, dans le secret de ma lutte et de ma diligence.*

A ma chère binôme Amira.

Pour son amour, son compréhension et son aide pour terminer ce travail

Remerciements

Ce travail a été effectué au sein du Département de génie électrique de l'Université de Bouira.

Nous tenons tout d’abord à remercier le bon DIEU le tout puissant et miséricordieux qui nous a donné la force et la patience d’accomplir ce modeste travail.

Nos très vifs et sincères remerciements vont à nos chers parents qui ont toujours été là pour nous et qui nous ont donné un magnifique modèle de labeur et de persévérance. Nous espérons qu’ils trouveront dans ce travail toute notre gratitude, notre reconnaissance et notre amour.

Nos vifs remerciements à Mr ABDENNOUR ALIMOHAD, notre encadreur pour nous avoir fait l’honneur de nous encadrer, et nous guider par ses conseils avisés et son aide très précieuse, ses remarques, ses suggestions et sa disponibilité durant notre préparation de ce mémoire.

Nous remercions ensuite l'ensemble des membres de jury

qui nous ont fait l’honneur d'accepter d'évaluer et bien vouloir étudier avec attention notre travail.

Nous adressons nos profonds remerciements à toute notre famille qui a toujours été présente à notre côté au long de la thèse.

Un grand remerciement à nos enseignants et enseignantes qui ont contribué à notre formation, depuis le cycle primaire au cursus universitaire.

Notre remerciement s’adresse également à tous nos amis pour leur soutien moral et leurs encouragements.

Enfin, nous tenons à remercier toutes les personnes qui ont participé de près ou de loin à la réalisation de ce travail.

Résumé

La communication est l'un des moyens les plus répandus chez les êtres humains pour exprimer leurs états émotionnels internes. Par conséquent, il serait intéressant de développer un système capable de reconnaître automatiquement ces émotions. Dans notre projet, nous nous concentrons sur la création d'un système de reconnaissance des émotions vocales en utilisant des techniques d'apprentissage profond. Le système repose sur l'utilisation de plusieurs paramètres spectraux tels que les MFCC, ZCR, Chroma_stft, chroma_cqt, Mel Spectrogramme, ainsi que des paramètres prosodiques tels que RMS (énergie) et pitch. Chaque type d'émotion est représenté par un modèle CNN.

Afin d'optimiser les performances du système, nous avons réalisé de nombreux tests pour déterminer le taux de reconnaissance le plus élevé. En termes de paramètres, la combinaison des techniques MFCC et RMS a obtenu les meilleurs résultats avec un taux de reconnaissance de 84,44 %. Nous avons également constaté qu'en fusionnant les paramètres prosodiques et spectraux avec le modèle CNN, nous avons pu améliorer davantage les performances du système, atteignant ainsi un taux de reconnaissance de 85,88 %.

Mots clés : Reconnaissance, émotions, paramètres prosodiques, paramètres spectraux, CNN.

Table des Matières

Chapitre 2 : système de reconnaissance automatique des émotions basée sur l'apprentissage profond

Chapitre3 : Résultats et discussions

Liste des Figures

Liste des Tableaux

Listes des Acronymes

AND	Acide Désoxyribo Nucléique	
AI	Intellegence Arteficielle	
AMDF	Average Magnitude Difference Function	
API	Alphabets phonétiques international	
ATM	Asynchronous Transfer Mode	
CAH	Classification Ascendante Hiérarchique	
CNN	Convolutional Nueron Network	
CQT	Constant-q-transform	
Db	Décibel	
DCT	Discret Cosinus Transform	
DL	Deep Learning	
DNN	Dens neural network	
EEG	Électroencéphalographie	
FACS	Système de Codage de l'Action Facial	
FC	Fully connected	
FFT	Fast Fourier Transfer	
FN	False negative	
FP	False positive	
GMM	Gaussian Mixture Model	
HMM	Hidden Markov Models	
ISODATA	Iterative Self-Organizing Data Analysis Technique	
MFCC	Mel-Frequency Cepstral Coefficients	
RAE	Reconnaissance automatique emotionnelle	
RAVDESS	Rverson Audio-Visuel Data Base of Emotional Speech and Song	
RL	Reinforcement Learning	
RMS	Root Mean Square	
RNN	Recurrent neural network	
STFT	Short-time Fourier Transforme	
SVM	Support Vector Machine	
TP	True Positive	
TN	True Negative	
ZCR	Zero Crossing Rate	

Introduction Générale

La communication, est un outil indispensable pour l'être humain. Les émotions font partie intégrante de notre vie quotidienne. Elles influencent nos réactions, nos décisions et nos interactions sociales. Les émotions peuvent varier d'une personne à l'autre et d'une situation à l'autre, allant de la joie et l'excitation à la tristesse, la colère et la peur. Elles sont souvent accompagnées de manifestations physiologiques telles que des changements dans la voix, la respiration, les expressions faciales et le langage corporel [1].

La parole, est un moyen essentiel de communication humaine. Elle nous permet de transmettre des informations, de partager des idées et d'exprimer nos émotions. La production de la parole implique la coordination complexe de plusieurs systèmes, y compris les muscles du larynx, de la gorge, de la langue et des lèvres. Ces systèmes travaillent ensemble pour produire des sons vocaux qui sont ensuite façonnés en mots, en phrases et en discours [2].

La voix est un élément clé de la parole, et elle contient des informations riches sur l'état émotionnel d'une personne. Lorsque nous sommes émus, notre voix peut changer de plusieurs façons. Par exemple, lorsque nous sommes joyeux, notre voix peut devenir plus mélodieuse et rythmée, tandis que lorsque nous sommes en colère, notre voix peut devenir plus forte et plus tendue. Ces variations émotionnelles dans la voix sont perçues et interprétées par les auditeurs, ce qui leur permet de comprendre notre état émotionnel et de réagir en conséquence.

Dans cette étude, nous nous intéressons à la reconnaissance des émotions vocales, en mettant l'accent sur l'utilisation de techniques d'apprentissage automatique et d'intelligence artificielle. Nous examinerons les caractéristiques acoustiques et prosodiques de la voix qui sont associées aux émotions, ainsi que les modèles et les algorithmes utilisés pour la classification des émotions dans les signaux vocaux.

L'objectif de cette recherche est de contribuer à l'avancement de la reconnaissance des émotions vocales en exploitant le potentiel de l'apprentissage profond basé sur les CNN. Comme tout système de reconnaissance ; le nôtre comporte aussi l'étape d'extraction de paramètre dont lequel nous avons pris deux grandes catégories ; spectrale et prosodique. Nous espérons ainsi ouvrir de nouvelles perspectives pour une meilleure compréhension de l'expression émotionnelle dans la parole humaine et pour l'amélioration des interactions homme-machine, de la santé mentale et de l'analyse des sentiments dans les médias sociaux.

Notre projet est composé de trois chapitres, qui sont structurés de la manière suivante :

Le premier chapitre est un état de l'art qui résume le domaine traité dans notre travail : nous le consacrons à des généralités sur l'émotion, la production de la parole et la reconnaissance

émotionnelle, ainsi que ses applications. Par la suite, le deuxième chapitre se concentre sur l'exploration du système de reconnaissance automatique des émotions, où nous examinerons en détail les paramètres prosodiques tels que la fréquence fondamentale (pitch) et l'énergie, ainsi que les paramètres spectraux tels que les coefficients cepstraux de fréquence mel (MFCC, le Chroma_stft, le Mel spectrogramme de et le chroma_cqt). Ensuite nous allons introduire les méthodes de modélisation les plus utilisées comme CNN.

Dans le troisième chapitre, Nous implémenterons le système de reconnaissance émotionnelle automatique en utilisant PYTHON pour l'analyse des signaux vocaux, et nous présenterons les résultats des expériences réalisées pour évaluer son efficacité.

Enfin, nous terminerons ce mémoire par une conclusion générale et quelques Perspectives.

Chapitre 1 : Généralités sur la reconnaissance des émotions vocales.

1.1. Introduction

La communication est un élément fondamental de notre vie quotidienne, permettant aux individus de partager des informations, d'exprimer leurs sentiments et de se connecter les uns aux autres. Parmi les différentes formes de communication : la parole. La parole, c'est le langage articulé humain. Elle joue un rôle très important dans les systèmes de reconnaissances automatiques telle que la reconnaissance de locuteur.

Les émotions sont importantes dans nos vies. Elles permettent d'améliorer la communication entre les individus, de confirmer la compréhension du message véhiculé et de s'adapter à une situation particulière.

Dans ce premier chapitre, nous allons aborder la description de l'émotion et ses différentes bases puis la parole et la reconnaissance automatique des émotions.

1.2. Reconnaissance émotionnelle

1.2.1. Définition de l'émotion

L'émotion est une tâche difficile caractérisé par des effets à la fois physique et mentaux [1]. Elle a diverses ramifications physiques, par exemple la peur entraine des pleurs, la joie peut entrainer un sourire [2]. D'un point de vue morale, les émotions apparaissent comme des schémas de fonctionnement cérébral permettant de réagir rapidement à la décision à pendre [3].

Le point de départ de la recherche commence par la reconnaissance vocale et la manière dont les signaux vocaux sont perçus par les humains. [4]

A l'heure actuelle, l'émotion est un système à trois composants, la premier correspond à des réponses physiologiques qui se traduit par exemple, par la température corporelle élevée ou par une modification du rythme et l'intensité du système respiratoire ou tachycardie .Puis le deuxième, la composante examine les réponse comportementales et expressive, c'est-à-dire, les changements volontaires ou involontaires du visage, de la voix et la posture qui se produisent après une explosion émotionnelle. La troisième composante concerne les réponses cognitives et expériences dans les quelle le terme cognitif fait référence aux processus qu'un organisme reçoit des informations environnementales et interagit de manière appropriée. [5]

1.2.2. Classification des émotions

Les émotions sont des séquences courantes et vives qui interviennent tout au long de nos journées et en donnent le ton.

Ils sont classés en deux familles : les émotions primaires et les émotions secondaires.

1.2.2.1. Emotions primaires (émotions de base)

Les émotions primaires, aussi appelée « sentiments primaires », sont des sentiments généraux et innés mais aussi automatiques, ils sont inconscients et il a un démarrage rapide [1]. Il existe six émotions de base avec neutralité (pas d'émotion) dont chacune correspond à des expressions faciales qui sont les mêmes chez plusieurs personnes d'âges, de cultures ou des sexes différents ces émotions sont : la joie, le dégoût, la tristesse, la surprise, la colère, et la peur. [3]

La photo suivante montre les six émotions primaires :

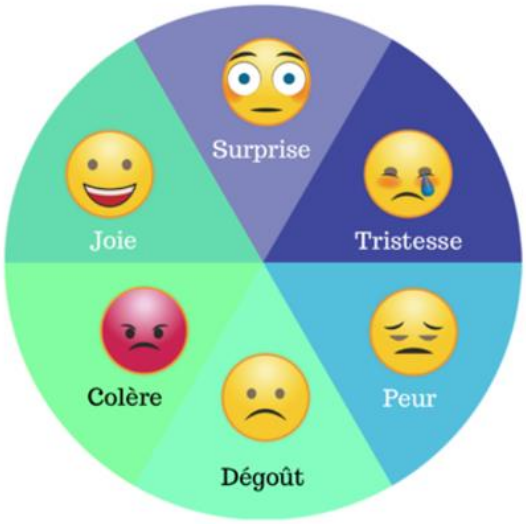

Figure 1.1 : les six émotions primaires [2].

1.2.2.2. Emotions secondaires (sociales et acquises)

Ces sentiments ne sont pas innés mais acquis tout au long de vie par l'influence de la famille, de la religion ou de la société elle-même. On peut citer : culpabilité, jalousie, honte, orgueil, vanité et gêne. On parle parfois d'émotions mixtes car ce sont des émotions complexes qui résultent d'un mélange d'émotions primaires. [2]

1.2.3. Catégories d'émotions

On va distinguer deux types de catégories d'émotions : les émotions positives et les émotions négatives.

1.2.3.1. Émotions positives

Dans le domaine de la psychologie positive, les sentiments positifs sont le résultat d'un bien-être individuel ou collectif. Les études scientifiques ont montré que ces sentiments sont toujours porteurs d'effets positifs.

Il s'agit de la joie, la gratitude, la sérénité, l'intérêt, l'espoir, la fierté, l'amusement, l'inspiration, l'admiration et l'amour. [6]

1.2.3.2. Émotions négatives

Dans le domaine de la psychologie négative, les sentiments négatifs sont le résultat d'un mal-être individuel ou collectif, des études scientifiques ont montré que ces sentiments sont toujours porteurs d'effets négatifs.

Il s'agit de la peur, la colère, le dégoût, la tristesse, la solitude, la dépression, la détresse, le désespoir, la frustration.

1.2.4. Représentation des émotions

1.2.4.1. Approche catégorielle (discrète)

Ces approches discrètes sont basées sur la présence d'un petit nombre d'émotions primaires discrètes, pour pouvoir distinguer ces sentiments. Par ces méthodes, d'autres émotions sont considérées comme un mélange de premières émotions, les approches discrètes de l'émotion est une tentative de visualiser les émotions majeures et de traiter chacune comme une émotion discrète. [7] un échantillon de cette approche est le modèle de Plutchilk illustré à la figure suivante : [1]

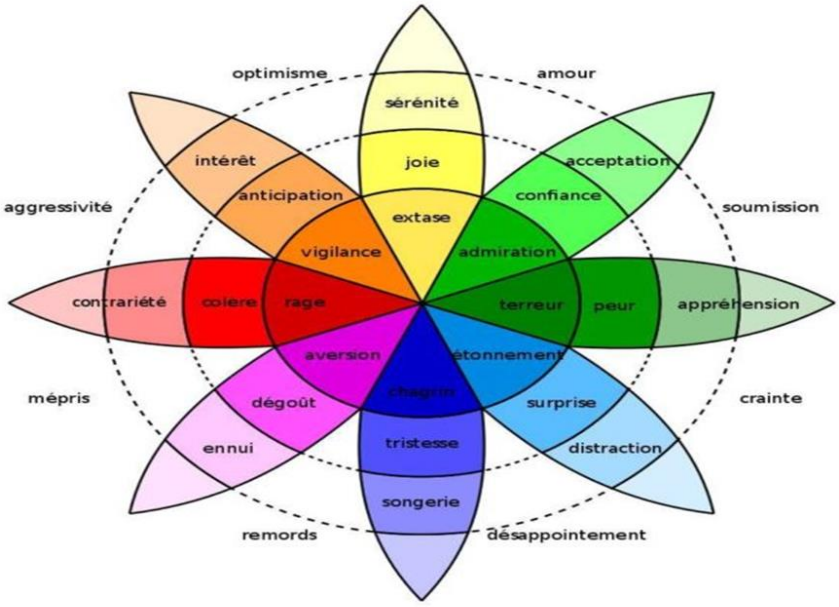

Figure 1.2 : la roue des émotions de Plutchilk. [2]

Cet échantillon se compose de 8 émotions de base, chacune est composées de 4 paires opposée en deux (heureux-triste, colère-peur, surprise-anticipation et dégoût-confiance) et de multiple variation. [2]

1.2.4.2. Approche dimensionnelle (continue)

Principalement, c'est la description verbale des sentiments subjectifs (subjective feeling). Dans cette imitation, différents états émotionnels sont cartographiés dans un espace à deux ou trois dimensions. Les deux dimensions principales sont la valence (agréable-désagréable) et l'activité (actif-passif). La troisième dimension, lorsqu'elle est utilisée, est souvent associée au contrôle ou au pouvoir intellectuel. [4]

On explique cette approche par une boussole de Russell suivante :

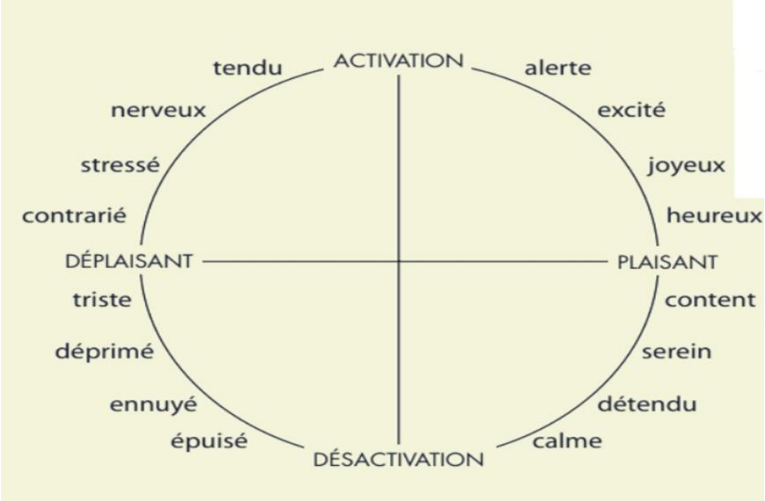

Figure 1.3 : La boussole de Russell. [2]

1.2.4.3. Approche hybride

Elle représente un compromis entre l'approche discrète et l'approche dimensionnelle. Ce modèle de perception des émotions est constitué de trois couches :

-La couche la plus abstraite a deux classes : « valence positive » et « valence négative ».

-La couche intermédiaire a des catégories d'émotions primaires : joie, colère, dégout, surprise, tristesse et peur.

-La couche inférieure est formée des émotions secondaires : adoration et tendresse pour l'amour ; l'enthousiasme et le zèle pour la joie ; agitation et le gêne pour la colère. [2]

1.2.5. Type de corpus des émotions

Il existe trois grandes catégories de corpus émotionnels utilisés dans le domaine de la reconnaissance automatique des émotions :

1.2.5.1. Corpus naturel (réaliste)

Ce sont des groupes d'enregistrements d'états émotionnels normaux et spontanés. Il a un très haut respect de l'environnement. Son inconvénient est que ces données sont très limitées en nombre de locuteurs, de courte durée, souvent de mauvaise qualité, et difficiles à classer en catégories d'émotions. [4]

1.2.5.2. Corpus induit

Il est appliqué dans le domaine de la psychologie et le domaine de la reconnaissance automatique des émotions. [4]. Les émotions évoquées sont généralement de faible intensité. Les émotions de cette catégorie font l'exposition du sujet à des taches difficiles pendant une courte période de temps. Pour induire le stress par exemple, visionnez des images, animées de films ou de jeux permettant d'induire cette émotion. [1]

1.2.5.3. Corpus acté (simulés)

Les émotions de cette collection sont créées par des acteurs professionnels ou semi-professionnels à partir de noms de catégories émotionnelles et /ou de scènes modelés [3]. Cependant, certaines critiques ont été faites contre ce type de corpus car les sentiments simulés sont plus forts que les sentiments naturels : [4]

1.2.6. Canaux de communication émotionnelle

Depuis les travaux de Picard en 1997, plusieurs systèmes ont été développés pour la reconnaissance des émotions. Sur la base de différents canaux de communication (Visage, voix, gestes, réactions physiologiques et neurologiques).

La photo suivante montre un exemple d'un Canaux de communication émotionnelle et capteurs associés. [8]

Figure 1.4 : Canaux de communication émotionnelle et capteurs associés. [8]

1.2.6.1. Expressions faciales

Le système de codage de l'action faciale (FACS) est une description des mouvements des muscles du visage et de la mâchoire/langue dérivée d'une analyse de l'anatomie faciale [9]. Le FACS se compose de quarante-quatre unités commerciales de base. Des groupes d'unités d'action indépendantes génèrent des expressions faciales. [10]

La reconnaissance automatique des expressions du visage constitue une autre source d'information importante dans le domaine de la reconnaissance automatiques des émotions. Soustraire système de mesure ou de codage de l'expression faciale (FACS) en tant que proposition pour contribuer au développement du domaine. Les profils de visage qui font référence à des expressions faciales émotionnelles peuvent être décryptés à l'aide de deux méthodes. La première approche se concentre sur les formes géométriques de certaines composantes du visage (par exemple, les traits du visage, la bouche et le nez. Dans la deuxième approche, il se concentre davantage sur l'apparence générale du visage comprennent également des rides et des rainures devant. [1]

Et voici des exemples sur l'expressions faciales :

Figure 1.5 : Des exemples sur l'expressions faciales. [11]

1.2.6.2. Signaux physiologiques

A la fin des années 1990, l'analyse des signaux physiologiques a été proposée comme approche alternative et complémentaire à l'identification des émotions chez les animaux. Rosalind Picard été l'un des premiers à mettre en évidence l'importance de ces indices dans la reconnaissance des émotions chez les humains.

De nos jours, de nombreux indicateurs physiologiques sont actuellement utilisés pour décrire et mesurer les émotions. Chaque signal n'est pas analysé seul mais en combinaison avec d'autres signaux du système nerveux central ou périphérique. Les exemples incluent la fréquence cardiaque,

la conductance cutanée, l'activité musculaire, les changements de température cutanée, les changements de pression artérielle, les signaux EEG et le volume respiratoire. La fréquence cardiaque et la conductivité électrique de la peau sont deux des signaux les plus courants. [8]

1.2.6.3. Voix

La voix est l'un des modèles et des sources les plus fiables pour la reconnaissance des émotions. Par exemple, un son faible et saccadé accompagné des mots " un peu excité " peut traduire des sentiments de mélancolie ou de tristesse, La voix peut être monotone, avec une modulation réduite, et des pauses prolongées peuvent être présentes. Ces caractéristiques peuvent indiquer une absence d'excitation émotionnelle ou un état de dépression. Ainsi qu'une voix forte et puissante Cela peut traduire des sentiments de colère. [12]

La voix est l'un des indicateurs les plus sensibles traduire et identifier l'état émotionnel de L'orateur. [13]

Le son peut être utilisé pour authentifier une personne comme en utilisant des empreintes digitales ou l'ADN. Ces derniers éléments font partie des composants du corps humain, ils évoluent peu ou pas dans le temps et ne peuvent être modifiés par l'intention de l'individu. En raison de ces caractéristiques, le terme biométrie est souvent utilisé. [14]

1.3. Parole et reconnaissance automatique

1.3.1. Parole

La parole est définie comme un changement de la pression de l'air excitante et libérée par le système articulatoire, Elle est considérée comme une tâche fondamentale dans le développement des techniques de télécommunication. Elle comporte deux formes d'informations permettant de distinguer les catégories d'émotions : les informations linguistiques et les informations paralinguistiques. [2]

1.3.1.1. Niveau acoustique

La phonologie acoustique travaille pour convertir ce signal en signal électrique avec le transducteur adéquat qui est le microphone. Actuellement, le signal conséquent est dans la plupart des cas discrétiser, puis soumis à un groupe de traitements statistiques pour montrer les taris acoustiques : sa fréquence fondamentale, son énergie, et son spectre. [15]

En termes de caractéristiques sonores, le son se compose de paramètres vocaux de base caractérisés par la fusion de l'intensité, la durée et le timbre. La hauteur est mesurée en hertz (vibrations par seconde) et est mesurée quantitativement par la longueur, la masse et la tension des cordes vocales d'un individu. L'intensité sonore se mesure en décibels (dB) et est définie par

l'amplitude des vibrations des cordes vocales d'un individu. La durée est mesurée en secondes (secs) et constitue la base de la synchronisation des messages. Le débit, c'est-à- dire le nombre de syllabes émises par seconde. Le timbre de la voix est précisé par les caractéristiques physiologiques au niveau de la longueur et de la masse des cordes vocales. [16]

1.3.1.2. Niveau phonétique

C'est est un niveau d'analyse de la langue qui étudie la production, la transmission et la perception des sons du langage parlé, à ce niveau, on s'intéresse à la manière dont les organes de la parole (lèvres, langue, les cordes vocales et le palais, sont utilisés pour produire les différents sons du langage.)

Il emploie le système des alphabets phonétiques international (API) qui relie des symboles phonétiques aux sons, de façon à permettre l'écriture compacte et universelle des prononciations. Aussi, la phonétique articulaire se soucie des classes phonétiques : les voyelles, les semi-voyelles, les liquides, et les consonnes. [15]

1.3.1.3. Niveau phonologie

Au niveau phonologique, nous étudions les sons et leurs organisations ; Ainsi, nous distinguons spécifiquement la phonologie segmentaire de la phonologie supra-segmentaire (présentation sonore).

La présentation générale correspond à tous les phénomènes supra-impératifs qui accompagnent et structurent la parole, et comprend des éléments tels que l'intonation, l'accent, le rythme, la mélodie, ou le timbre de la voix. La phonologie syllabique comprend les phonèmes et les monèmes. [16]

1.3.1.4. Niveau morphologie

La morphologie est le domaine de la linguistique, qui détaille comment les formes lexicales sont donnée à partir d'un ensemble réduit d'unités porteuses de sens, nommés morphèmes. [15].

La morphologie est l'étude de la formation des mots et de leurs variations. [17]

1.3.1.5. Niveau syntaxique

La structure syntaxique est la structure intermédiaire entre le sens, la structure hiérarchique, la perception phonémique des phrases et linéarité du message linguistique. [6]

Une phrase grammaticale est une unité de sens qui comprend au moins deux éléments obligatoires ; le sujet et le prédicat. Il peut aussi contenir un troisième composant le complément de phrase. [15]

1.3.1.6. Niveau sémantique

La différence entre sémantique et syntaxe reste relativement floue. Ainsi, une description est souvent porteuse de sens, tout analyse syntaxique basée sur un nombre important de classes d'éléments du discours possède inévitablement un caractère sémantique. [18]

La sémantique étudie le sens indépendamment du contexte d'utilisation des mots et des phrases. [15]

1.3.1.7. Niveau pragmatise

Pragmatique est défini comme selon du contexte, elle abordé des sujets : les présuppositions, les implications de dialogue, elles sont moins développées que la sémantique. [15]

La pragmatique étudie le sens qui est communique dans un contexte particulier. [19]

1.3.2. Production de la parole

La parole est traduite par une action volontaire de plusieurs muscles, L'air sert de source d'énergie appliquée aux poumons, ce qui produit une pression pulmonaire qui génère un son par phonation à travers la glotte qui est modulée par le conduit vocal en voyelles. [15].

La production de la parole est un processus fascinant qui implique la coordination de nombreux systèmes du corps humain.

Ce processus est essentiel à la communication humaine car il nous permet d'exprimer nos pensées, nos émotions et nos intentions à travers des sons et des mots.

Notamment le système respiratoire, le système phonatoire et le système articulatoire. [2]

Nous étayons cette explication par un description détaillée de l'appareil vocal.

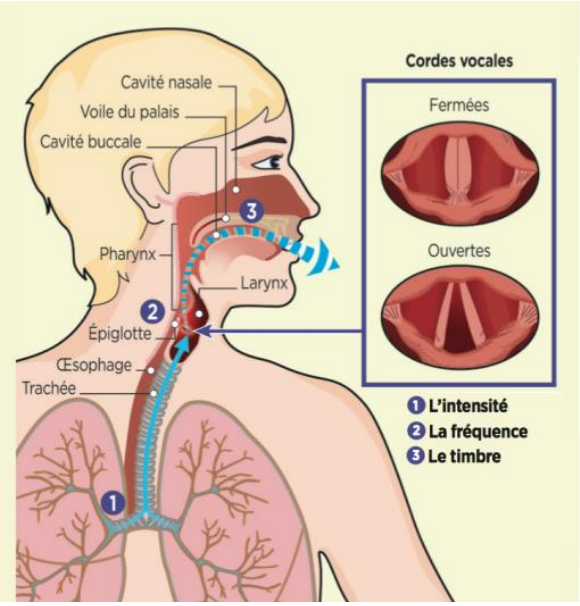

Figure 1.6 : Description détaillée de l'appareil vocal. [1]

1.3.2.1. Phase respiratoire (Les poumons et la trachée)

La phase de respiration est une étape essentielle du processus de respiration, inspirant et expirant de l'air. Pendant l'inspiration, le diaphragme se contracte et tombe, et les muscles intercostaux créent un espace dans les poumons qui est rempli d'air. Quant à l'expiration, le diaphragme se détend et permet à l'air de quitter les poumons, et ainsi il est dirigé vers les cordes vocales à travers la trachée. Les poumons sont le foyer du son. [20]

1.3.2.2. Phase phonatoire (cordes vocale et glotte)

La phase phonatoire se réfère à la production de sons vocaux pendant la parole ou le chant. C'est la partie du processus de production vocale où les cordes vocales se rapprochent, créant une obstruction partielle à l'air expiré par les poumons. L'air forcé à travers cette obstruction crée des vibrations des cordes vocales, produisant ainsi des sons phonétiques. Cette phase est essentielle pour la communication humaine et joue un rôle clé dans la production de la voix et des différents sons de la parole. [2, 21]

1.3.2.3. Phases d'articulatoire (les cavités supra glottiques)

Les résonateurs sont considérés comme des organes qui amplifient et modulent le son émis par le hautbois, car il plane et filtre à travers les cavités de la gorge, de la bouche et de la cavité nasale, et prend sa propre couleur et son propre timbre, ce qui lui permet de distinguer les sons.

Selon quatre cavités : pharynx, buccale, rénale et nasale. [2]

1.3.2.3.1. Cavité pharyngale

La cavité pharyngée mesure environ 8 cm de long. Il est vertical et se situe au-dessus du larynx. Le pharynx est le canal de la membrane musculaire qui relie le larynx à la cavité nasale et à la cavité buccale, et est le premier résonateur rencontré par les sons générés au niveau du larynx. Il peut être divisé en trois sous-parties et se situe de bas en haut : Larynx - pharynx qui se situe derrière l'épiglotte et descend jusqu'à l'œsophage ; le pharynx, qui s'étend de l'épiglotte dans la cavité buccale ; et le nasopharynx, qui s'étend du palais mou aux fosses nasales. [22]

1.3.2.3.2. Cavité orale

La cavité orale mesure environ 8 cm de long. Il représente la partie limitée du conduit vocal lèvres en avant. Les multiples articulations au-dessus du larynx, à la fois fixes et mobiles dans cette cavité buccale, modifient sa longueur et sa forme. Il aide à distinguer la cavité buccale et marque l'extrémité antérieure du conduit vocal. L'adduction forme la saillie des lèvres et l'adduction forme un anneau final le long du bord formé par les cavités buccale et pharyngée. [22]

1.3.2.3.3. Cavité labiale

La cavité labiale est située entre les incisives et les lèvres contractées sous une forme ou l'autre, elle est forme à l'extrémité antérieure du conduit vocal. Elle n'interfère avec la vocalisation que si les lèvres sont baissées vers l'avant. [2]

1.3.2.3.4. Cavité nasale (Les fosses nasales)

Elle rapproche les fosses nasales la cloison nasale sépare les cavités. Les cavités nasales sont incluses dans un groupement osseux. Il se compose principalement des os de la mâchoire supérieure, sphénoïde, palatine, sabot.

Les deux cavités tapissées de mucus sont séparées par la cloison nasale, Lors de l'émission de sons laryngés, le palais mou tombe de sorte que les cavités pharyngée, orale et nasale permettent à l'air de s'échapper par la cavité nasale, produisant un son nasal. [22]

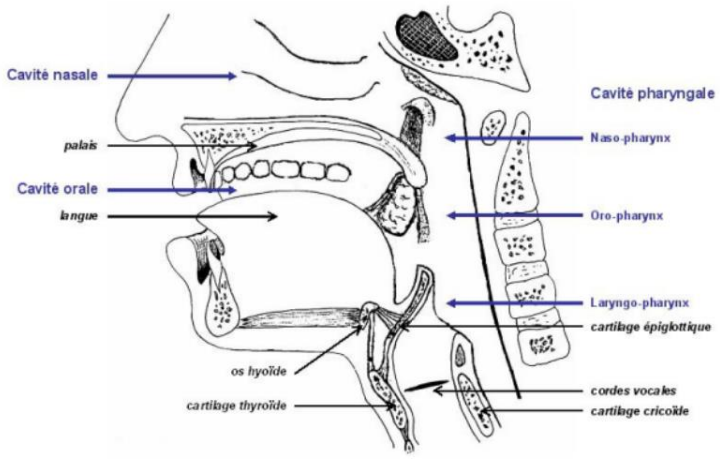

Figure 1.7 : les fosses nasales. [23]

1.3.3. Classifications des sons du langage

Le signal de parole dit parfois périodique et parfois aléatoire. Ce qui conduit à la classification des sons en deux types : Sons voisés et Sons non voisés.

1.3.3.1. Sons voisés

Les sons vocaux, comme les voyelles, par exemple, sont causés par l'air passant des poumons à travers la trachée, entrainer la vibration des cordes vocales. Ce sont généralement des sons semi-périodiques.

Ils se produisent principalement dans le temps d'expression et sont caractérisés par une énergie haute-basse fréquence d'environ kilohertz par kilohertz de bande passante, ce qui contribue de manière significative à l'information. [24]

1.3.3.2. Son non voisés :

Les sons non isolés sont acycliques. Son énergie est concentrée dans les hautes fréquences et est considérée comme du bruit. Les cordes vocales sont séparées et ne vibrent pas en leur sein ce pays. Ainsi, l'air circule librement vers le conduit vocal. Ensuite, l'air circule librement dans le conduit vocal. [3]

1.4. Reconnaissance automatique des émotions

La reconnaissance vocale est l'une des choses fondamentales de la vie humaine, en particulier dans le domaine de la communication avec la machine. Où il reconnaît ses sentiments à travers la légèreté de sa voix contenue dans le signal audio. Plusieurs applications de reconnaissance vocale existent dans notre quotidien et dans de nombreuses entreprises nous n'avons plus besoin d'un réceptionniste, mais nous parlons à des programmes. La reconnaissance automatique de l'émotion

s'est récemment améliorée dans de nombreux domaines, notamment : la médecine, l'enseignement à distance, le marketing et la sécurité.

1.4.1. Domaines d'application de la reconnaissance automatique des émotions

La reconnaissance automatique des émotions a de nombreux domaines d'application. Voici quelques-uns des principaux domaines où cette technologie est utilisée :

1.4.1.1. Reconnaissance des émotions pour l'enseignement à distance

L'enseignement à distance est devenu de plus en plus important et présente de nombreux avantages. La reconnaissance automatique des sentiments vocaux en temps réel est en cours de développement dans un système qui peut dire si un étudiant se sent frustré, ennuyé ou ennuyé par le matériel qu'il étudie. [2]

1.4.1.2. Reconnaissance des émotions pour le marketing

La reconnaissance émotionnelle vocal est la technique la plus appliqué pour le marketing, et de lui pour déterminer les sentiments des clients à partir du son. [2]

Il existe un système appelé 'le point de falaise' qui collecte des données sur les sentiments des clients entre la voix et les traits du visage afin de connaitre leur satisfaction et d'identifier s'ils aiment le produit. [25]

1.4.1.3. Reconnaissance des émotions pour la médecine

La reconnaissance vocale joue un rôle principal en psychopathologie car elle peut définir l'état d'une personne grâce à ses émotions par traiter ses paroles comme l'autisme. Cette recherche est effectuée par les scientifiques de l'institut de technologie et la Massachusetts [2].

Cette technologie la start-up beyond verbal appuyer les psychologues pour diagnostique l'état mental du patient. [25]

1.4.1.4. Reconnaissance des émotions pour la sécurité

C'est une structure automatique de reconnaissance des émotions, Il est exploité dans le contrôle des endroit public ou personnel pour détecter l'existence d'émotions extrême [2]. Il peut aussi utiliser dans les automobiles préciser que le conducteur en état et les différentes émotions des accidentel, le stress, abimé ou l'effet alcool. [3]

1.4.1.5. Reconnaissance des émotions dans les banques

Le développement de l'intelligence synthétique dans la reconnaissance vocale, de nos jours définir la contentement des clients par leur émotion dans leurs voix la joie ou la colère par exemple,

et simplifie a rassemblés les observation des clients et perfectionner les fonctions en analysant la voix de client quand il téléphone le centre de contact. [2]

Ils utilisent l'application de reconnaissance des émotions de la banque comme empreinte sonore émotionnelle pour les machines à mode de transfert Asynchrones (ATM).

1.5. Conclusion

Dans ce chapitre nous avons abordé présentation des concepts généraux sur notre domaine choisi.

Nous sommes d'abord partis d'une représentation générale de l'émotion et de sa classification et de ses différentes catégories, puis nous avons parlé de la voix et de la parole et défini leur relation avec la reconnaissance émotionnelle.

Ensuite nous avons parlé sur la reconnaissance automatique des émotions et leurs différons applications.

Au deuxième chapitre, nous allons parler plus détailler sur la reconnaissance automatique des émotions et ses étapes, puis nous allons parler sur l`apprentissage profond et l'architecteur de réseaux de neurones profond.

Chapitre 2 : Système de reconnaissance automatique des émotions basé sur l'apprentissage profond

2.1. Introduction

Le traitement automatisé des émotions est une partie importante de l'étude informatique des comportements de communication humaine. Des systèmes solides et fiables de reconnaissance des émotions sont nécessaires pour améliorer les capacités analytiques de la prise de décision humaine et des interfaces homme-machine pour une communication efficace [26]. Parmi ces systèmes, le système de reconnaissance automatique émotionnelle (RAE) qui compte sur le signal acoustique [4], il s'agit de rétablir l'information non linguistique qui représente l'état émotionnel de la lecture [2]. Ce système est basé sur des caractéristiques pertinentes en plus de leurs vitesses et accélérations. Le système RAE sera formé pour tester pour identifier les catégories d'ensemble des émotions des données utilisées [4]. Pour faire cette étape on va utiliser l'apprentissage profond (Deep Learning DL). Ce dernier est une branche de l'apprentissage automatique qui s'appuie sur les réseaux neuronaux [24].

2.2. Phase d'apprentissage

2.2.1. Signal acoustique

Le signal acoustique fonctionne en transmettant des informations sur l'état émotionnel du locuteur, pour détecter les incompréhensions entre le locuteur et le système de dialogue homme-ordinateur. La force d'un signal acoustique est perçue par son intensité, qui a un sens auditif. [4]

2.2.2. Prétraitement

Cette étape est appliquée avant l'extraction des paramètres du signal de parole. Il configure le signal acoustique (audio) et améliore la précision et l'efficacité du processus d'extraction des paramètres. Dans cette partie on peut faire appel aux étapes suivantes : filtrage, segmentation avec chevauchement et fenêtrage [2].

2.2.2.1. Filtrage

Le filtrage peut être appliqué pour supprimer les composants indésirables d'un signal audio, comme la réduction du bruit causé par les conditions environnementales ou d'autres perturbations [2]. Le signal audio est filtré puis échantillonné avec une fréquence donnée.[27]

2.2.2.2. Segmentation et chevauchement

Les méthodes de traitement du signal utilisées dans l'analyse des signaux audio sont basées sur des signaux statiques. Dans cette étape de segmentation, le signal précédemment accentué est divisé en N trames d'échantillons de parole. Généralement, N est défini de sorte que chaque trame corresponde à environ 20 à 30 millisecondes de parole.[27]

2.2.2.3. Fenêtrage

En segmentant le signal, on obtient des trames qui seront multipliées dans une fenêtre temporelle pour analyse, et cette fenêtre peut donner plusieurs formes par exemple la fenêtre de Hamming qu'est le plus utilisée en traitement de la parole [27]

2.2.3. Extraction de paramètre

Cela dépend des informations spectrales collectées dans les caractéristiques vectorielles. Ces caractéristiques peuvent affecter la précision du système RAE, il est donc nécessaire de sélectionner les meilleures caractéristiques pour la classification. On peut conclure de ces études que l'information paralinguistique est plus spécifique ou suffisamment pertinente pour reconnaître l'état émotionnel d'une personne à partir du son [2, 4].

2.2.3.1. Paramètres Prosodiques

C'est un canal parallèle du message parlé, Ils peuvent impliquer des segments de parole plus longs (syllabe, mot, phrase). Essentiellement caractérisé par : la fréquence fondamentale (pitch), l'énergie, les formants, la durée, et le débit de la parole [2].

2.2.3.1.1. Fréquence fondamentale (pitch)

Cette fréquence est le phénomène prosodique le plus représentatif en termes d'état émotionnel. Il exprime le rythme du cycle d'ouverture et de fermeture des cordes vocales du larynx en parlant. Le domaine de reconnaissance affective utilise la fréquence fondamentale pour effectuer une classification parmi les voix des locuteurs et est appelé F0 [3].

Les valeurs de F0 peuvent être estimées avec la forme de la fonction d'auto corrélation. [4]

2.2.3.1.1.1. Détection de pitch par autocorrélation

L'autocorrélation est défini comme une convolution d'un signal utilisée pour distinguer un son voisé et d'un son non voisé, elle détermine la fréquence fondamentale d'un signal parole [2].

$$\mathbf{C}\ (\boldsymbol{\tau}) = \lim_{T->\infty} \frac{1}{T} \int_{-T\backslash 2}^{T\backslash 2} x(t)x(t-\tau)dt \qquad (2.1)$$

2.2.3.1.1.2. Détection de pitch par AMDF (Average Magnitude Différence Fonction) :

Il s'agit d'un algorithme temporel utilisé pour la détection de fréquence, l'un des algorithmes les plus avancés pour les utilisateurs de traitement du signal afin qu'un signal de différence soit généré entre l'original et l'audio retardé appelé signal de déférence. [2]

La Figure (2.1) représente les blocs d'algorithme de AMDF

$$\mathbf{D(k)} = \sum_{n=1}^{N-1-k} |x(n) - x(n+k)| \ ; \quad \mathbf{k=0, 1, \ldots\ldots, K} \qquad (2.2)$$

X(n) : signal original

X (n+k) : le signal décalé par la valeur de k dans le temps

K : c'est le nombre des points de la fonction AMDF

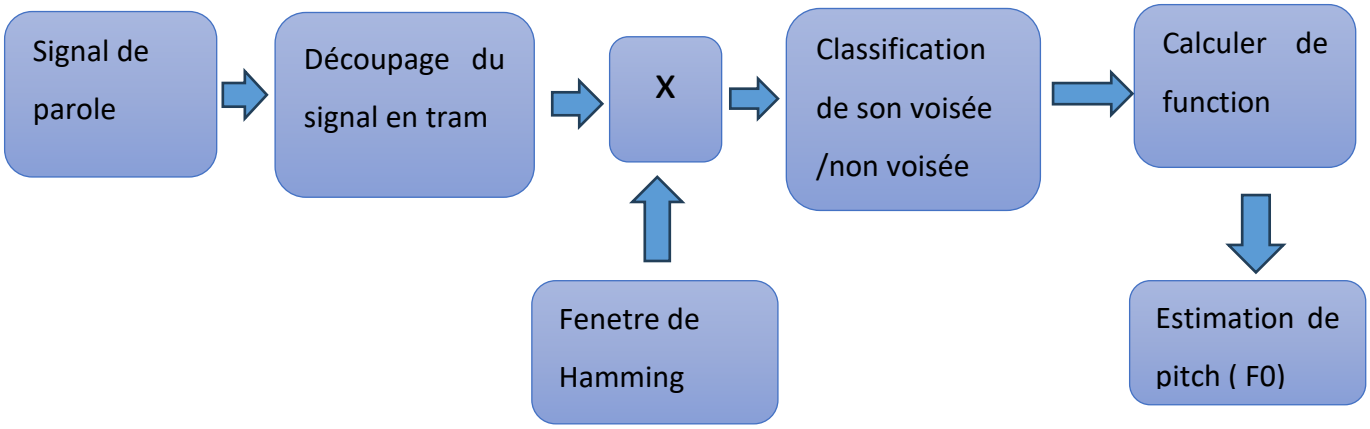

Figure 2.1: schéma Bloc d'algorithme AMDF.

2.2.3.1.2. Intensité (Energie)

L'intensité est la variation de l'amplitude du signal de parole qui résulte de plus ou moins d'énergie, Ainsi, il est possible de distinguer un son fort d'un son faible, de sorte qu'il représente l'amplitude de l'onde sonore résultant d'une énergie forte, inférieure ou supérieure.

L'intensité est calculée en décibels (dB) comme suit : [2]

$$\mathbf{I = 10 \log \sum_{n=1}^{N} S_p^2 (n)} \qquad \mathbf{(2.3)}$$

2.2.3.1.3. Formant

L'énergie vocale produite par les cordes vocales passe par les cavités du conduit vocal (cavités nasale, buccale, buccale et pharyngée). Cette énergie sonore est résonnée et filtrée**.**

Les structures spectraux des sons voisés qui nous permettent d'identifier le type de son prononcé (voyelles, consonnes ou autres sons vocaliques), Il est généralement utilisé en phonétique ou en acoustique pour décrire les vibrations des tractus vocaux ou des instruments de musique [2].

Voir la Figure 2.2

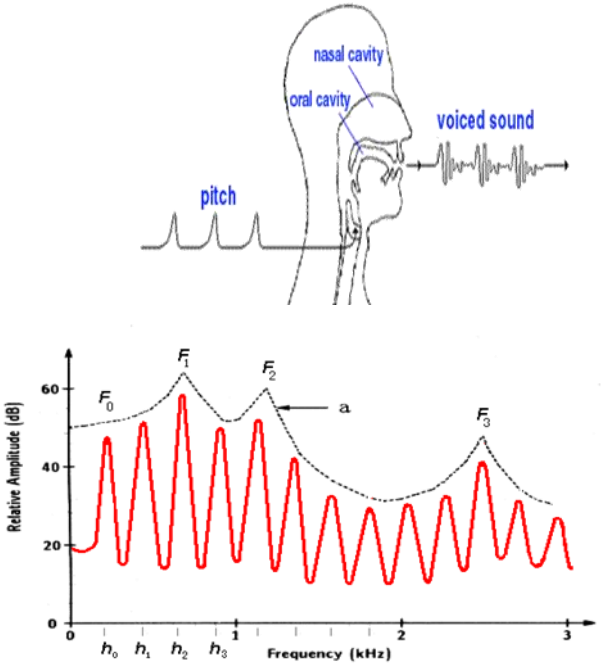

Figure 2.2 : Structure spectrale des sons voisés et son spectre. [2]

2.2.3.2. Paramètres spectraux

Les paramètres spectraux jouent un rôle primordial dans la distinction des catégories d'émotions. Il est absolument essentiel qu'une grande variété de paramètres puisse être mesurée dans les recherches futures. Ces paramètres viennent de technologies de compression de la parole comme ZCR, MFCC, STFT, CQT et Spectrogramme. [4]

2.2.3.2.1. Coefficients cepstraux sur l'échelle Mel (MFCC)

Les paramètres MFCC (Mel Frequency Cepstral Coefficients) sont extraits par une technique couramment utilisée en traitement du signal audio et en reconnaissance automatique de la parole pour représenter les caractéristiques spectrales d'un signal audio. Il s'agit d'une transformation qui permet de capturer les informations les plus pertinentes du spectre de fréquence d'un signal. [28, 29]

Les coefficients MFCC sont définis comme la transformée en cosinus logarithmique inverse du spectre de puissance du segment de parole. La puissance spectrale est calculée en appliquant un ensemble de filtres uniformément espacés à une échelle de fréquence modifiée, appelée échelle de pente. Elle se compose à des étapes. [2]

La Figure (2.3) résume les différentes étapes de MFCC

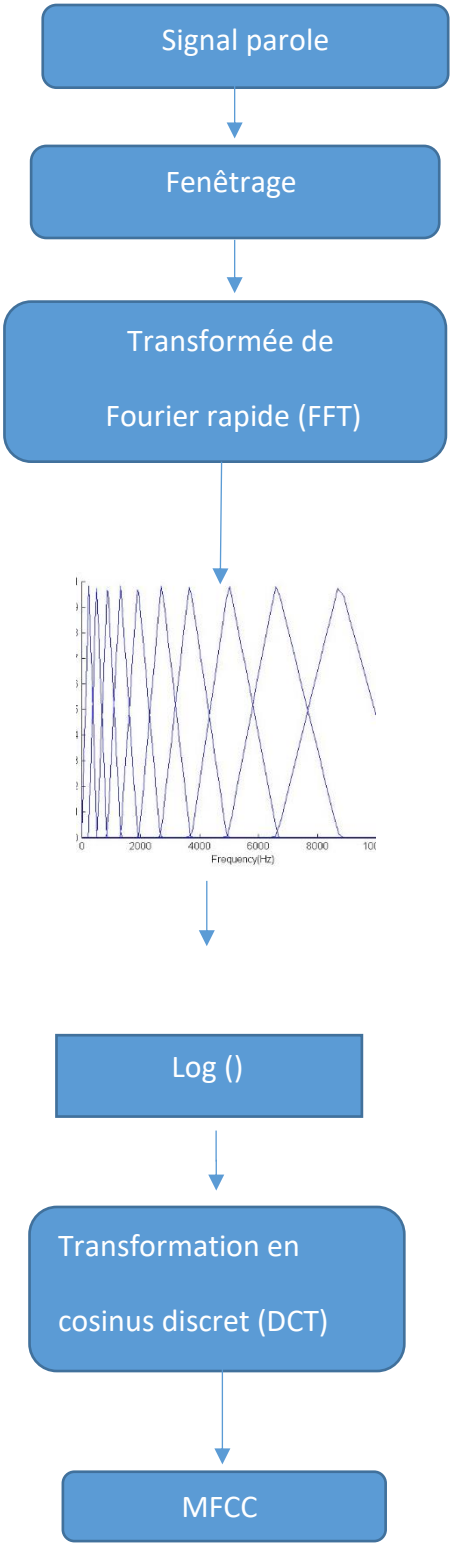

Figure 2.3 : Schéma bloc de MFCC.

Fenêtrage

D'abord, le signal est découpé en trames chevauchées de faible durée où il est considéré comme quasi stationnaire. Après, chaque trame est multipliée par une fenêtre temporelle d'analyse qui prendre des formes, uniformes, triangulaires, gaussiennes, etc. la fenêtre de Hamming est la plus utilisée dans le domaine du traitement de la parole. [30]

Transformée de Fourier rapide (FFT)

Est un algorithme pour calculer rapidement la transformée de Fourier discrète. La FFT sera exercée pour chaque fenêtre d'analyse, afin d'effectuer le passage du signal du domaine temporel au domaine fréquentiel. [2]

Transformation en cosinus discret

Cette transformée consiste à multiplier les logarithmes des réponses en énergie des filtres de Mel par des fonctions sinusoïdes de fréquences variées. Le but est de liée ces valeurs d'énergie pour constituer par la suite les coefficients de notre vecteur MFCC final.[30]

La DCT est écrire par La forme suivante : [3]

$$\mathbf{Cn} = \sum_{K=1}^{N} \log S_k \cos\left(n * \left(k - \frac{1}{2}\right) \frac{\pi}{N}\right) \quad \textbf{(2.4)}$$

2.2.3.2.2. Zero Crosing Rate (ZCR)

Le passage par zéro se produit dans un signal échantillonné lorsque deux échantillons successifs ont des polarités opposées. On peut estimer le taux de passage par zéro à court terme en utilisant la formule suivante [31]

$$\mathbf{ZCR\ (X)} = \frac{1}{N} \sum_{N-1}^{N=1} an \quad avec \begin{matrix} an=1\ Si\ Xn+1 \ldots\ldots < 0 \\ an=0\ Sinon \end{matrix} \qquad \textbf{(2.5)}$$

n : Nombre d'échantillon du signal de parole.

N : Nombre d'échantillon par fenêtre.

X_n : Échantillon du signal.

X_{n+1} : Échantillon suivant du signal.

Le taux de passage par zéro (ZCR) est un indicateur facile à calculer qui représente le nombre de fois où le signal traverse la valeur centrale (zéro) dans sa représentation amplitude/temps. Les sons non voisés, en raison de leur caractère aléatoire, ont généralement un taux de passage par zéro plus élevé que les sons voisés (comme nous expliquons dans le chapitre 1). Cette caractéristique est utilisée pour classifier les signaux en sons voisés et non voisés. [32]

2.2.3.2.3. Transformée de Fourier à court terme (STFT Short-time Fourier transform)

STFT est une variante de la transformée de Fourier qui est appliquée à des portions limitées du signal en utilisant une fenêtre qui se déplace le long de l'axe temporel. En ce qui concerne les signaux discrets, la STFT est utilisée pour analyser les caractéristiques fréquentielles du signal sur des intervalles de temps spécifiques

Ainsi, on obtient la formule de la transformée de Fourier discrète dans laquelle une fenêtre est introduite et positionnée en un certain point. La STFT est donc une représentation bidimensionnelle du signal qui dépend du temps et de la fréquence, d'où le nom de "représentation temps-fréquence". L'objectif de cette transformée est d'analyser les caractéristiques du signal dans le domaine temps-fréquence, et non de reconstruire le signal temporel, donc la transformée inverse n'est pas utile.

Lorsque l'on représente une STFT, on utilise uniquement son module, plus précisément le carré du module, appelé "spectrogramme" de la STFT. Le spectrogramme fournit des informations sur la quantité d'énergie présente dans le signal autour de la fréquence et de l'instant donné. [33]

2.2.3.2.4. Mel spectrogramme

Un spectrogramme est une représentation visuelle du spectre de fréquence d'un signal qui varie dans le temps. Il est généré en utilisant une transformée de Fourier courte (STFT) du signal, qui convertit le signal du domaine temporel en domaine fréquentiel. Contrairement à la transformée de Fourier (TF) d'un signal complet, la STFT calcule le spectre pour des segments courts et les organise en une séquence. Par exemple, pour un son d'une seconde, on peut diviser en 10 intervalles de 100 ms, puis calculer la TF pour chaque intervalle de temps et concaténer les séquences spectrales agrégées. Il y a un chevauchement possible entre les intervalles, par exemple avec le premier intervalle.

Le spectrogramme Mel diffère du spectrogramme classique en ce sens que les fréquences sont converties à l'échelle de Mel. Cette échelle est obtenue par une transformation logarithmique appliquée aux composantes du domaine fréquentiel. L'objectif principal de cette transformation est de représenter les composantes sonores en accordant une plus grande importance à celles qui sont plus sensibles à l'oreille humaine. L'idée sous-jacente est de réduire la quantité d'informations corrélées tout en préservant les informations pertinentes. La transformation utilisée pour passer du domaine fréquentiel à l'échelle de Mel est donnée par : [34]

$$\text{Mel}(f) = 2595 \times \log 10\,(1 + \frac{f}{700}) \qquad 2.6$$

2.2.3.2.5. Transformée en Q constant (CQT constant-q transform)

La transformée en Q constant, telle qu'introduite dans [Brown, 1991], est étroitement liée à la transformée de Fourier. Tout comme la transformée de Fourier, une transformée en Q constant est une banque de filtres, mais contrairement à celle-ci, elle a des fréquences centrales espacées de manière géométrique, fk = f0 · 2^(k/b) (k = 0, ...), où b détermine le nombre de filtres par octave.

La transformée en Q constant présente une autre caractéristique intéressante : sa résolution temporelle augmente à mesure que les fréquences deviennent plus élevées. Cela est similaire à notre système auditif, où il faut plus de temps non seulement à un ordinateur numérique, mais aussi à notre perception auditive pour reconnaître les fréquences basses. Cela est dû au fait que la musique est généralement moins complexe dans les registres inférieurs. [35]

2.2.4. Modélisation

Dans un système REA, différentes stratégies sont utilisées pour concevoir le classificateur de type statistique afin de reconnaître des modèles. Les classificateurs font partie de l'approche paramétrique ou non paramétrique [2]

L'approche probabiliste exprime la catégorie dans le cas de l'émotion comme source et la conçoit avec une densité de probabilité connue, dont les plus importantes sont le modèle de mélange Gaussien (GMM) et le modèle caché Markov (HMM) [2, 3].

Le classifieur est considéré comme non paramétrique lorsqu'il n'exploite aucune distribution statistique paramétrique et se base uniquement sur la distance spectrale. Cette catégorie englobe notamment des méthodes telles que la minimisation de distance, les k plus proches voisins, les k-means, ISODATA, ainsi que des méthodes plus récentes telles que les réseaux neuronaux et les machines à vecteurs de support (SVM).[36]

2.2.4.1. Apprentissage profond

2.2.4.1.1. Définition de l'apprentissage profond (Deep Learning)

Le terme "Deep Learning" ou Apprentissage profond, a été introduit pour la première fois au Machine Learning (ML) par Dechter en 1986, et aux réseaux neuronaux artificiels par Aizenberg et AL en 2000.

L'apprentissage profond (en anglais Deep Learning, Deep structured Learning, hierarchical Learning) est un ensemble de stratégies d'apprentissage automatique affriolant de modéliser avec une haute classe d'abstraction des données à l'aide des structures discrètes de différentes transformations non linéaires. [37]

Le Deep Learning ou bien apprentissage profond est une partie de l'intelligence artificielle provenue du machine Learning ou le matériel est capable d'apprendre tout seul, contrairement à la

programmation ou le contenu consiste à mettre en œuvre des règles prédéfinies par caractère. [39] voir la Figure (2.4).

L'apprentissage est fondé sur la technologie analogique des neurones artificiels qui se compose de milliers d'unités qui exécutent chacune de petites opérations simples. Les effets d'une première couche de « neurones » servent d'entrée aux calculs d'une deuxième couche et ainsi de suite. [37]

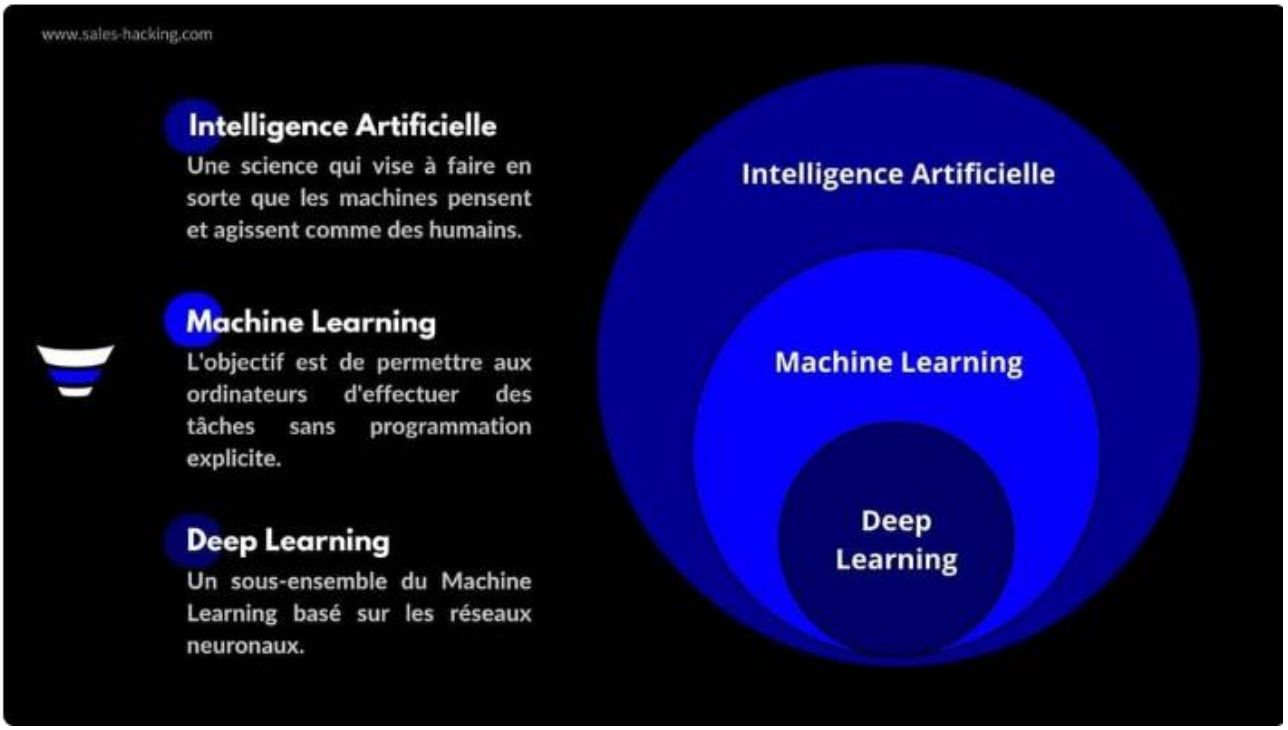

Figure 2.4 : La relation entre l'intelligence artificielle, le Machine Learning et le Deep Learning [38]

2.2.4.1.2. Fonctionnement de l'apprentissage profond

Le Deep Learning est alimenté par des couches de réseaux neuronaux, L'apprentissage profond repose sur réseau du neurone artificiels inspires du cerveau humain, il est constitué de dizaines de couches de neurones, chacune récit et explique les notions de la couche précédent. [41]

Les réseaux d'apprentissage en profondeur sont formés sur la base des structures de données complexe qu'ils rencontrent. Ils construisent des modèles informatiques composés de plusieurs couches de traitement pour crées plusieurs niveaux d'abstraction pour représente les données. [37]

2.2.4.1.3 Pourquoi l'Apprentissage profond ?

L'IA émotionnelle rassemble un ensemble de cas d'utilisation de machine Learning particulièrement bien adaptés aux techniques de Deep Learning. [42]

Différents algorithmes d'apprentissage profond n'ont émergé que lorsque l'apprentissage automatique n'a pas réussi à résoudre une variété de problèmes d'IA afin de :

- Perfectionner le développement d'algorithmes traditionnels dans les tâches d'intelligence artificielle.
- Évoluer une grande quantité de données comme le big data.
- Pour s'adapter à tout type de problème. Extraire les fonctionnalités automatiquement. [43]

2.2.4.1.4. Types d'apprentissage profond

Il existe trois types d'apprentissage :

2.2.4.1.4.1. Apprentissage supervisé

L'algorithme est guidé par une connaissance préalable de ce que devraient être les résultats de sortie de modèle. Ainsi, le modèle ajuste ses paramètres pour réduire la déférence entres les résultats attendus obtenus et les résultats attendus. Pour diminuait la marge d'erreur au fur et à mesure que le modèle est en trainé qu'il peut être appliqué à de nouveaux cas.[44]

Enseignement supervisé bien que les deux types d'apprentissage soient soumis à l'intelligence artificielle, dans le premier cas il y a un chercheur pour "guider" l'algorithme dans le chemin de l'apprentissage en lui fournissant des exemples qu'il juge probants après qu'il ait été pré-classé avec les résultats attendus. L'IA apprend alors de chaque exemple en ajustant ses paramètres (poids des neurones) afin de réduire l'écart entre les résultats obtenus et les résultats attendus. Ainsi, la marge d'erreur est réduite lors des séances d'entraînement, dans le but de pouvoir généraliser son apprentissage à des situations nouvelles. [45]

2.2.4.1.4.2. Apprentissage non supervisé

L'apprentissage non supervisé n'utilise pas de donnés étiquètes. Ils devenaient alors impossibles pour l'algorithme de calcules avec certitude la note de passage. Son objectif est donc de réduit quels groupe son présent dans nos données, il existe deux principaux domaines de paradigmes dans l'apprentissage non supervisé pour récupérer des ensembles :

- Les méthodes par partitionnement : les algorithmes des k-means.
- Les méthodes de regroupement hiérarchique classification ascendante hiérarchique (CAH). [44]

Dans le cas de l'apprentissage non supervisé, l'apprentissage automatique se déroule de manière totalement indépendante. Les données sont alors envoyées à l'appareil sans lui fournir d'exemples de résultats attendus en sortie. [45]

2.2.4.1.4.3. Apprentissage par renforcement

L'apprentissage par renforcement est type d'apprentissages statistique dérivé à la fois de l'apprentissage humain et animal. [46]

L'apprentissage par renforcement (RL) est la science de la prise de décision. Il s'agit d'apprendre un comportement optimal dans un environnement pour maximiser la récompense. Ce comportement

optimal s'apprend par le biais d'interactions avec l'environnement et d'observations de la façon dont il interagit, de la même manière que les enfants explorent le monde qui les entoure et apprennent des actions qui les aident à atteindre un objectif. [47]

2.2.4.1.5. Architecteur de réseaux de neurones profond

Il existe un grand nombre de variation d'architecteur profond. Les algorithmes d'apprentissage en profondeur et les réseaux de neurones profonds (DNN), les réseaux de neurones convolutifs en réseau (CNN) et les réseaux de neurones récurrents (RNN) ...etc., qui sont expliqué dans la figure (2.16).

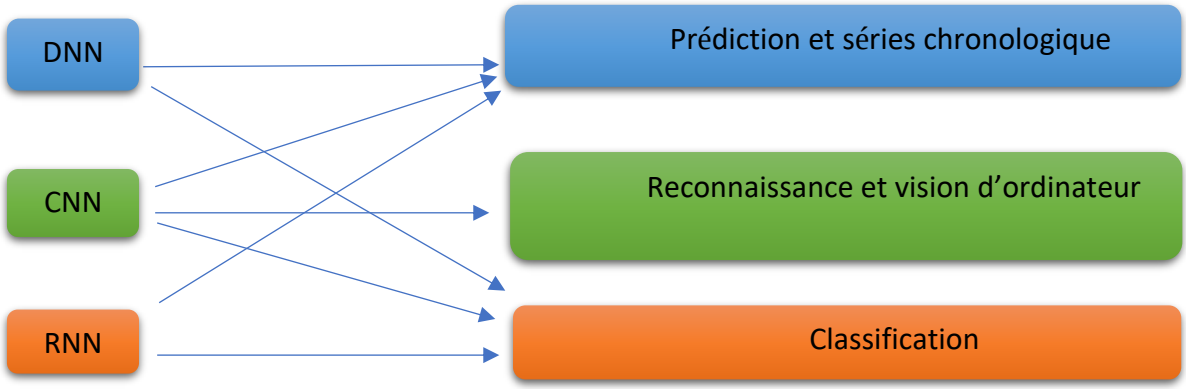

Figure 2.5: Types de modèles utilisant des architectures d'apprentissage en profond.

2.2.4.1.5.1. Perceptions multi couche

Cette couche est la plus vieux et la plus simple permis les réseaux de neurones [48]

Structure composée de différentes couches cachées de neurones dont la sortie sert d'entrée aux neurones de la couche suivante. Elle est le premier réseau de neurones qu'a trouvé de nombreuses application pratique telles que la reconnaissance de fleurs.

Peut l'utiliser pour toutes les taches de classifications supervisées. A l'heure actuelle c'est le modèle le plus populaire et il est implémenté par des nombreuses bibliothèques comme tensor flow et wek.

La particularité topologique de ce réseau est que tous les neurones d'une couche sont connectés à tous les neurones de la couche suivants. [49]

2.2.4.1.5.2. Réseaux neurones profond (DNN)

Un réseau de neurones profond est un réseau d'un neurone artificiel multicouche a plus d'une couche cachée entre son entrée et sa sortie et des millions de paramètre

Une caractéristique très prometteuse des DNN est qu'ils peuvent apprendre des caractéristiques invariantes de haut niveau à partir de données brutes et les classer efficacement, des données qui peuvent être utiles pour la reconnaissance des émotions. [50]

2.2.4.1.5.3. Réseaux de neurones convolutifs (CNN) :

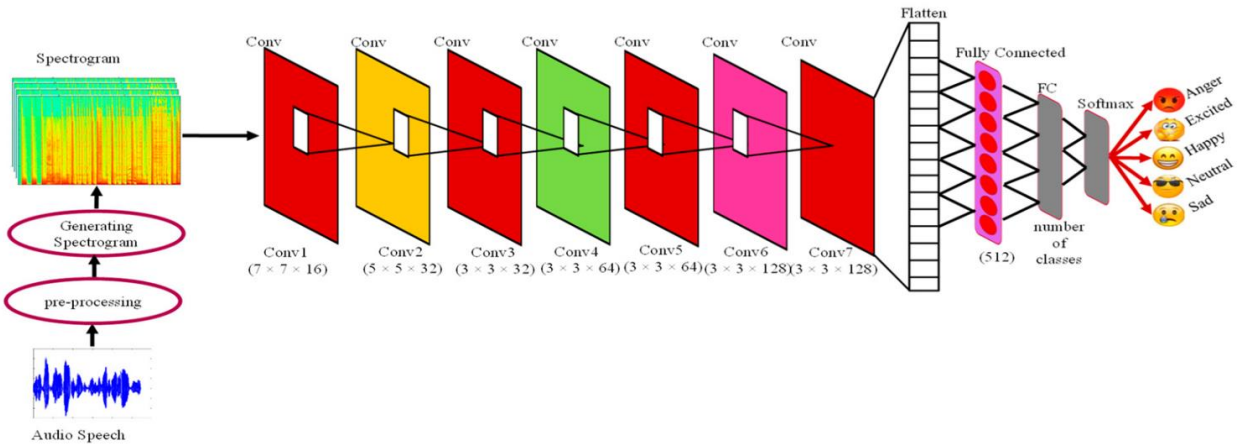

Figure 2.6 : schéma de modèle CNN pour la reconnaissance des émotions. [40]

Les réseaux de neurones convolutifs sont utilisés dans grand nombre d'apprentissage avec un grand succès. L'un des premières applications est la reconnaissance de l'écriture manuscrit et il donnée des bons résultats dans les taches de détection et de classification d'objet, ils appliqués aussi sur la reconnaissance facile et la connaissance de texte, le CNN est inspiré de cortes visuel des vertébrés, il est développé par le Cun en 1990. Les CONV Nets sa destinée a traité les données sous forme de tableaux de valeurs en N dimension pour N^{+*}.

Le Conv Nets est basé sur quatre idées principaux qui exploitent les propriétés signes normaux.

-les connexions locales.

-les poids partagé et la couche de regroupement (pooling). [26]

2.2.4.1.5.3.1. Architecture de réseaux de neurone convolutif

Les CNN se composent de trois types de couches : les couches convolutives, les couches de regroupement et les couches entièrement connectées.

- La couche de convolution (CONV), analyser les données d'un champ récepteur.
- La couche de pooling (POOL), compresser l'information en réduisant la taille de l'image intermédiaire.
- La couche de correction (Relu).
- La couche "entièrement connectée" (FC), C'est une couche similaire au sens perceptron.
- La couche de perte LOSS.

Couche de convolution (CONV)

La couche convolution sont consistée d'un ensemble de noyaux, les poids peuvent être appris processus de convolution c'est un produit scalaire en vecteur [50]. Le but de cette couche est d'analyser les images fournies en entrée et de détecter la présence d'un ensemble de factures. [44]

Couche pooling

La couche de pooling est une opération généralement appliquée entre deux couches de convolution, son but est de réduire la taille des images, tout en préservant leur caractéristique les plus essentielles. Parmi les plus utilisée on retrouver le max pooling [44]. L'un des avantages est que les entrés ne sont pas spatialement une meilleure généralisation de caractéristique et diminué l'échantillonnage et le rend des entrées les calculs suivants sont plus efficaces. [50]

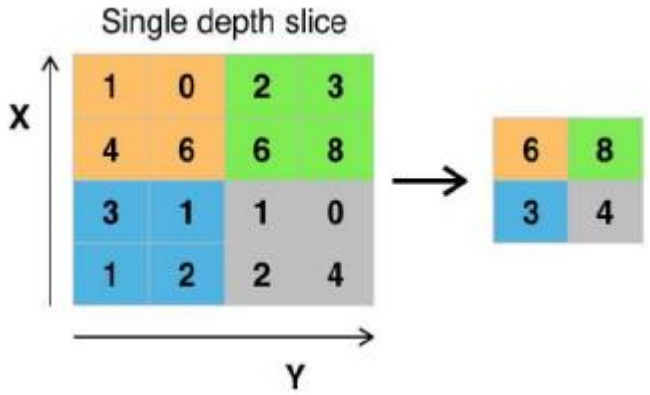

Figure 2.7 : Pooling avec un filtre 2x2 et un pas de 2. [36]

Couche d'activation

Le but principal de la couche d'activation est d'ajouter du non linéaire au réseau. Ce dernier est nécessaire pour approximer les problèmes non linéaires .il y a plusieurs fonctions d'activation non linéaire comme sigmoïde, tanch et ReLU. Les ReLU sont la plus populaire utiliser dans le CNN.

$$\sigma(x) = \max(0,x) \qquad (2.7)$$

Avec cette méthode les valeurs négative sont supprimées et les positive a une correspondance linéaire, ce qui rend ReLU non linéaire avec décalage. [50]

Couche de fully connected (FC)

Sont places en fin d'architecteur de CNN et sont entièrement connectées à tous les neurones de sortie. Le FC applique successivement une combinaison dans le but final de classifier l'input image. [44]

Ces couches sont placées à la fin de la structure CNN et sont entièrement connectées à tous les neurones de sortie (d'où le terme entièrement connecté). Après réception du vecteur d'entrée, la

couche FC applique successivement une combinaison linéaire puis une fonction d'activation dans le but ultime de classifier l'image d'entrée. Enfin, il renvoie en sortie un vecteur de taille d correspondant au nombre de classes dans lesquelles chaque composante représente la probabilité que l'image d'entrée appartienne à une classe. [44]

Couche de perte (LOSS)

La couche de perte est responsable de définir comment le réseau pénalise l'écart entre la prédiction du signal et la vérité terrain lors de l'entraînement. Généralement, elle se trouve en dernière position dans le réseau. Différentes fonctions de perte adaptées à différentes tâches peuvent y être utilisées. Par exemple, la fonction de perte "Softmax" est utilisée pour prédire une seule classe parmi un ensemble de K classes mutuellement exclusives. La fonction de perte de l'entropie croisée sigmoïde est utilisée lorsque l'on souhaite prédire K valeurs de probabilité indépendantes comprises dans l'intervalle [0,1]. Enfin, la fonction de perte euclidienne est utilisée pour réaliser une régression vers des valeurs réelles. [36]

2.2.4.1.5.4. Réseaux de neurones récurrent (RNN)

Dans le traitement des données séquentielle comme les flux vidéo, il existe une technique d'apprentissage qui traite la séquence entière comme un seul grand entrée, cette méthode ne fonctionne pas pour des périodes variables séquences.

Alternativement, on peut former un modèle sur clips vidéo et le compiler plus tard sur la sortie de clips pour chaque vidéo. Cette approche ignore les dépendances entre les segments de chaque vidéo. Les réseau RNN sont un ensemble de réseaux de neurones qui traiter ces problèmes en lorsque en ajoutant des connexions toroïdales dans le graphe de réseau. [51]

Le réseau récurrent en complément permet u traitement fréquentiel des données, les séquence d'entrées X0 …… X_m. En fait, au temps t, ils calculent leur sortie en fonction de l'entrée Xt, mais aussi pour l'état de la couche cachée la dernière fois. [52]

2.2.5. Base de données

Il s'agit d'une basse qui conserve de manière structurée et avec une minimisation de la redondance les données des catégories émotionnelles obtenues lors de l'étape précédente (modélisation). Elle stocke les modèles traités afin d'améliorer la détection et la reconnaissance du modèle déclaré lors de la phase de test.

2.3. Phase de test

Pendant la phase de test, il y a deux parties qui visent à évaluer l'état émotionnel de l'orateur et à analyser les résultats pour une comparaison et une utilisation fiable. Les deux premières parties, à

savoir le prétraitement et l'extraction des paramètres, revêtent une importance cruciale lors de cette phase d'évaluation.

2.3.1. Comparaison

Le modèle de reconnaissance des émotions compare les caractéristiques extraites du signal vocal d'entrée avec les modèles de référence des émotions appris lors de la phase de formation. Cette comparaison peut être basée sur des métriques de similarité ou de distance pour déterminer quelle émotion correspond le mieux au signal vocal donné.

2.3.2. Décision

Dans cette étape, en fonction de la comparaison des émotions, le système de RAE prend une décision finale sur l'émotion exprimée dans le signal vocal d'entrée. Cela peut être sous la forme d'une classe émotionnelle spécifique (joie, tristesse, colère, peur, etc.) ou d'une évaluation continue de l'intensité de l'émotion. La prise de décision est basée sur les scores de probabilité de chaque classe émotionnelle et peut être utilisée pour prendre des décisions en temps réel.

Dans ce schéma en résumé les étapes président :

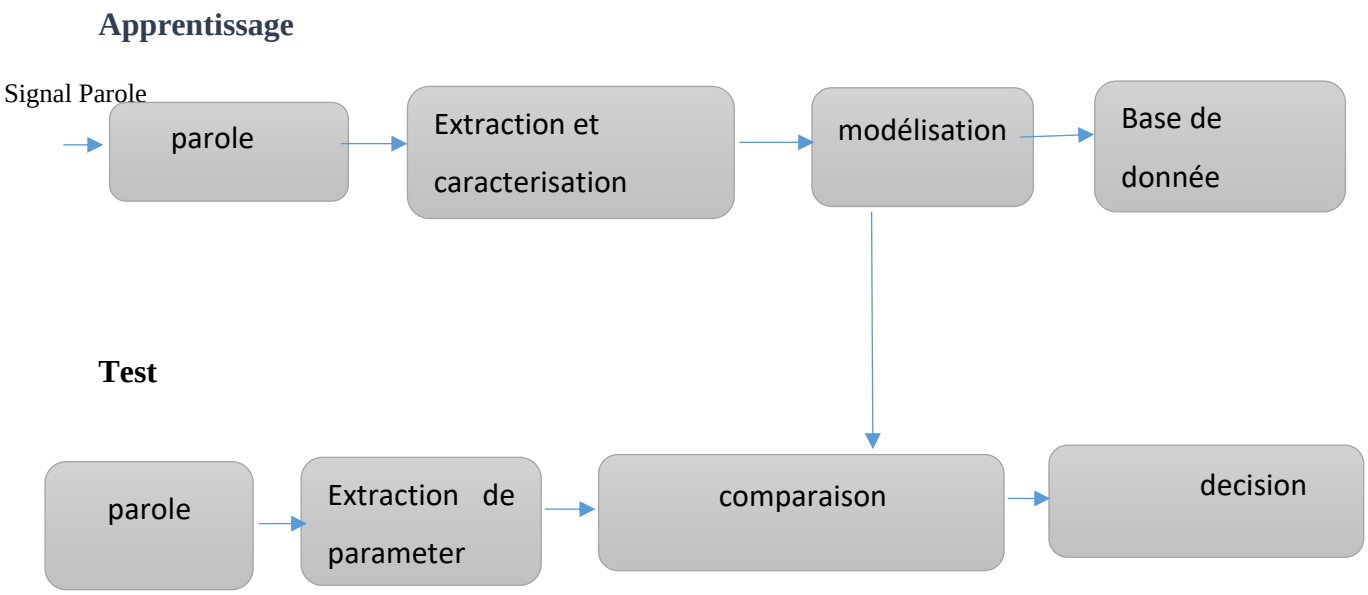

Figure 2.8 : la structure d'un système de RAE.

2.4. Conclusion

Dans ce chapitre, nous avons présenté les deux phases du système de reconnaissance émotionnels. La phase d'apprentissage et la phase de test, et aussi l'origine de l'apprentissage en profondeur, qui basés sur les réseaux de neurones. On commence par donnée les notions importantes qui sont en relation avec l'apprentissage profond (définition, architectures…etc.). Nous avons ensuite concentré notre attention sur les réseaux de neurones convolutifs CNN qui sont exploités pour atteindre notre objectif. Dans ce contexte, nous commencerons par montrer la structure générale de CNN. Après, nous détaillerons les différentes architectures communes de type CNN.

Chapitre 3 :

Résultats et discussions

3.1. Introduction

Dans le cadre de ce projet, nous avons utilisées toutes les notions théoriques essentiel à notre travail avec le but de : développement d'un système de reconnaissance automatique de l'émotion à partir de l'apprentissage profond. Pour implémenter se système en pratique, nous avons utilisé une base de données contenant des échantillons audios enregistré. L'objectif était d'analyser ces échantillons afin de traitées des caractéristiques acoustiques pertinentes pour la reconnaissance automatique des émotions. Pour ce faire, nous avons applique différente méthodes spectrales et prosodiques et nous avons utilisé le modèle CNN pour classifier les répétitions des audios avec leur émotion, avec un critère de performance qui est taux correct de reconnaissance, à la fin nous exécutons le code sur le logiciel python, pour réaliser notre projet, afin d'obtenir le meilleur taux correct de reconnaissance possible.

3.2. Base des données

Dans notre projet, nous utilisons un ensemble de donnée en accès libre appelé RAVDESS (Ryerson Audio-Visual Data base of Emotional Speech and Song). C'est une base de données audio-visuelle développée par le département de psychologie de l'Université Ryerson à Toronto, Canada. Utilisée pour la recherche sur la reconnaissance et la compréhension des émotions exprimées dans la parole et le chant. La base de données RAVDESS contient des enregistrements audio et vidéo d'acteurs et d'actrices professionnels qui ont été invités à exprimer différentes émotions primaires telles que la colère, la peur, le dégoût, la joie, la tristesse et la surprise. Les émotions sont exprimées à travers la parole et le chant, permettant ainsi aux chercheurs d'étudier la façon dont les émotions sont communiquées à travers ces modes d'expression, qui contient 7356 fichiers de taille totale égale à 24,8 Go. Elle est disponible en trois formats de modalité : audio-uniquement, vidéo uniquement et audio-vidéo. RAVDESS a été créé par 24 acteurs professionnels (12 femmes, 12 hommes), qui prononcent deux déclarations lexicalement assorties avec un accent nord-américain. La parole comprend des expressions calmes, heureuses, tristes, en colère, peur, surprises et dégoûtées, et chaque expression est produite à deux niveaux d'intensité émotionnelle (normal, fort), avec une expression neutre supplémentaire.

Nous utilisant les fichiers audio-uniquement pour notre simulation. Cette partie de RAVDESS contient 1440 dossiers dont 60 essais par acteur. Chacun de ces fichiers à un nom de fichier unique qui est composé d'un identificateur numérique en 7 parties

Par exemple : 03-01-06-01-02-01-12.wav.

Ces identificateurs définissent respectivement les caractéristiques de l'acteur comme suit :

-Modalité : 01 = audio-vidéo, 02 = vidéo uniquement, 03 = audio uniquement.

-Canal vocal : 01 = parole, 02 = chanson.

-Émotion : 01 = neutre, 02 = calme, 03 = heureux, 04 = triste, 05 =colère, 06 = effrayé 07=dégoûté, 08 = surpris.

-Intensité émotionnelle : 01 = normal, 02 = fort. « Pour émotion neutre y a pas une intensité forte ».

-Énoncé : 01 = « Les enfants parlent près de la porte », 02 = « Les chiens sont assis près de la porte ».

-Répétition : 01 = 1ère répétition, 02 = 2ème répétition.

-Les acteurs (1 à 24) : les actrices paires sont des femmes et les impaires sont des hommes.

3.3. Protocole

Notre travail consiste à mettre en place un système de reconnaissance des émotions vocal basé sur l'apprentissage profond. Pour cela, on utilise une base de données de 24 acteurs, Chacun de ces derniers a enregistré 8 émotions en plusieurs fois.

Le système est composé des parties principales ; une plateforme de développement pour la création d'un modèle du convolution CNN utilisant douze premiers acteurs (de 1 à 12).

Le système effectue une classification d'émotions à partir de fichiers audios. Des caractéristiques sont extraites à partir des données audios en utilisant les techniques de taux de passage par zéro, le Chroma_stft, Chroma_cqt, le MFCC, la valeur RMS, le pitch et le Mel Spectrogramme. Les caractéristiques extraites sont stockées dans une DataFrame appelée Features.

En phase d'apprentissage, nous avons pris les répétitions d'enregistrements pour créer les 96 modèles représentant toutes les émotions avec tous les acteurs, ensuite, ont utilisées des techniques d'extraction de caractéristiques pour extraire des informations utiles du signal audio vocal. Cela inclut les propriétés temporelles (durée, débit de parole), fréquentielles (spectre de puissance, fréquence fondamentale). Après, le modèle d'apprentissage en profondeur est en cours de construction, basé sur des réseaux de neurones convolutifs CNN. Le modèle est formé sur un ensemble de données annotées, où chaque échantillon vocal est associé à une émotion spécifique. Les résultats de comparaison sont donnés par des scores représentant la probabilité conditionnelle que l'émotion testée provienne du modèle généré la technique réseaux de neurones convolutifs CNN avec les techniques le taux de passage par zéro, le Chroma_stft, le MFCC, la valeur RMS, le Mel Spectrogramme, chroma_cqt et le

pitch. L'évaluation du système se calcule sous forme d'un taux correct de reconnaissance « Accuracy ».

Avec : $\text{Accuracy} = \frac{T_P+T_N}{T_P+T_N+F_P+F_N}$

T_P : True positive.

T_N : True negative.

F_P : False positive.

F_N : False negative.

3.4. Python

Python est un langage de programmation qui peut s'utiliser dans de nombreux contextes et s'adapter à tout type d'utilisation grâce à des bibliothèques spécialisées. Il est cependant particulièrement utilisé comme langage de script pour automatiser des tâches simples mais fastidieuses, Python est souvent utilisé pour le développement de l'apprentissage profond et bien d'autres domaines Il offre une large gamme de bibliothèques qui facilite le développement de divers types de projets. On l'utilise également comme langage de développement de prototype lorsqu'on a besoin d'une application fonctionnelle avant de l'optimiser avec un langage de plus bas niveau. Il est particulièrement répandu dans le monde scientifique, et possède de nombreuses bibliothèques optimisées destinées au calcul numérique.

Python est un langage de programmation polyvalent, facile à apprendre et à lire, qui offre de nombreuses fonctionnalités et une grande flexibilité pour le développement d'applications dans divers domaines.

3.5. Résultats et discussions

3.5.1. Utilisation de CNN sur le système de reconnaissance

La convolution est l'acte consistant à prendre les données d'origine et à en créer des cartes. Le Polissage est un sous-échantillonnage, le plus souvent sous la forme de "pooling maximal", où nous sélectionnons une région, puis prenons la valeur maximale dans cette région, et cela devient la nouvelle valeur pour toute la région. Les couches entièrement connectées sont des réseaux de neurones typiques, dans lesquels tous les nœuds sont "entièrement connectés".

Les couches convolutives ne sont pas entièrement connectées comme un réseau de neurones traditionnel. La structure CNN de base est la suivante : La couche de convolution (CONV) > La

couche de pooling (POOL) > La couche de correction (Relu) > La couche ‘‘entièrement connectée’’ (FC).

3.5.1.1. Paramètres spectraux

Dans cette partie nous avons testé individuellement les différents types de paramètres spectraux. En utilisant les paramètres suivants : MFCC, ZCR, Chroma_stft, Mel Spectrogramme, chroma_cqt, le nombre de coefficients varie entre 12 à 24 dans le paramètre MFCC.

Dans le tableau 3.1, on montre le taux correct de reconnaissance de MFCC dans chaque coefficients (12 à 24).

MFCC	12	14	16	18	20	22	24
Taux globale (%)	69,07	69,26	69,99	71,11	**75,93**	71,48	73,14

Tableau 3.1 : taux correct de reconnaissance de reconnaissance en fonction de paramètre MFCC.

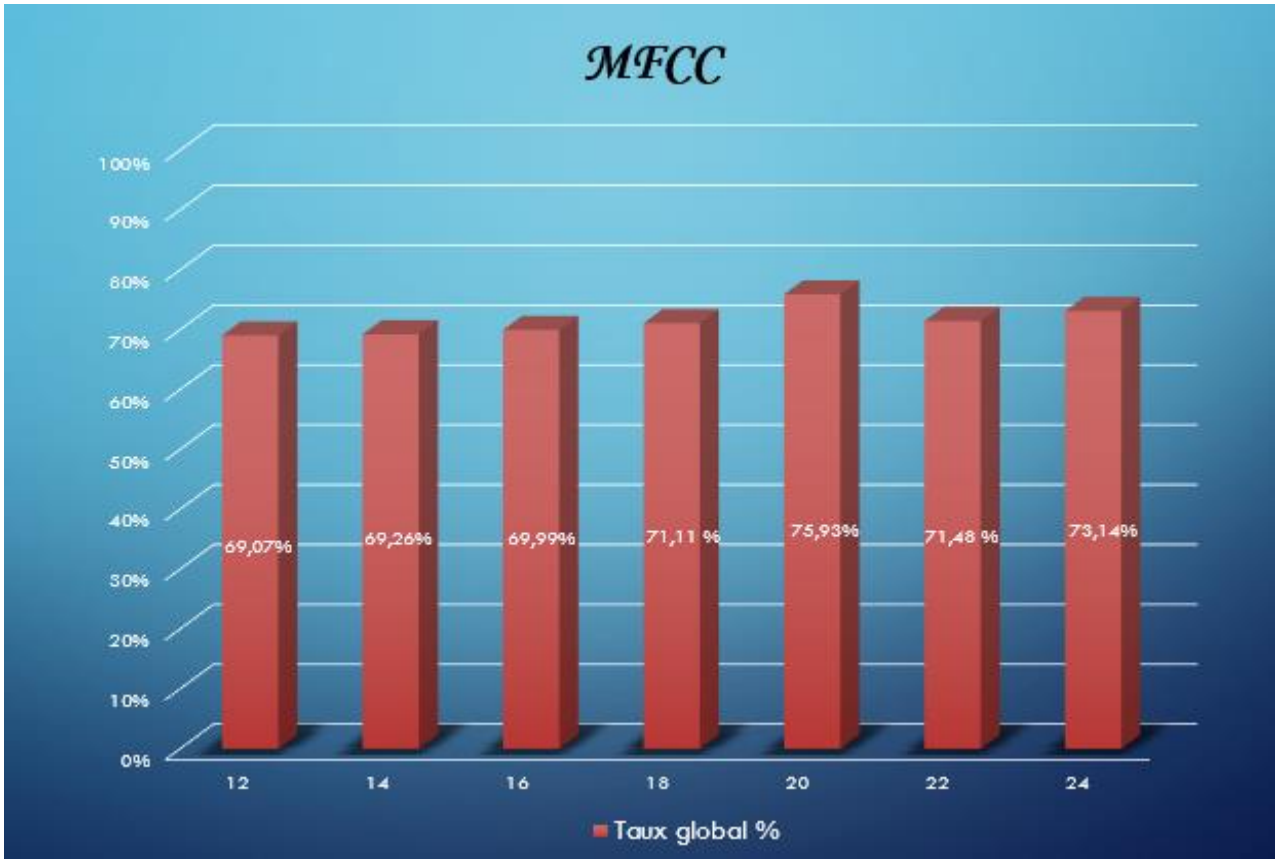

Figure 3.1 : taux correct de reconnaissance de reconnaissance en fonction de paramètre MFCC

On remarque que le meilleur coefficient de reconnaissance est obtenu pour le paramètre MFCC est 20 avec un taux correct de reconnaissance de **75,93%.**

Le tableau 3.2 montre le taux correct de reconnaissance des autres paramètres spectraux (ZCR, Chroma_stft, Mel Spectrogramme, chroma_cqt).

Paramètre	ZCR	Chroma_stft	Mel Spectrogramme	Chroma_cqt
Taux global (%)	16,48	30,93	**51,30**	34,81

Tableau 3.2 : taux correct de reconnaissance de reconnaissance en fonction des paramètres.

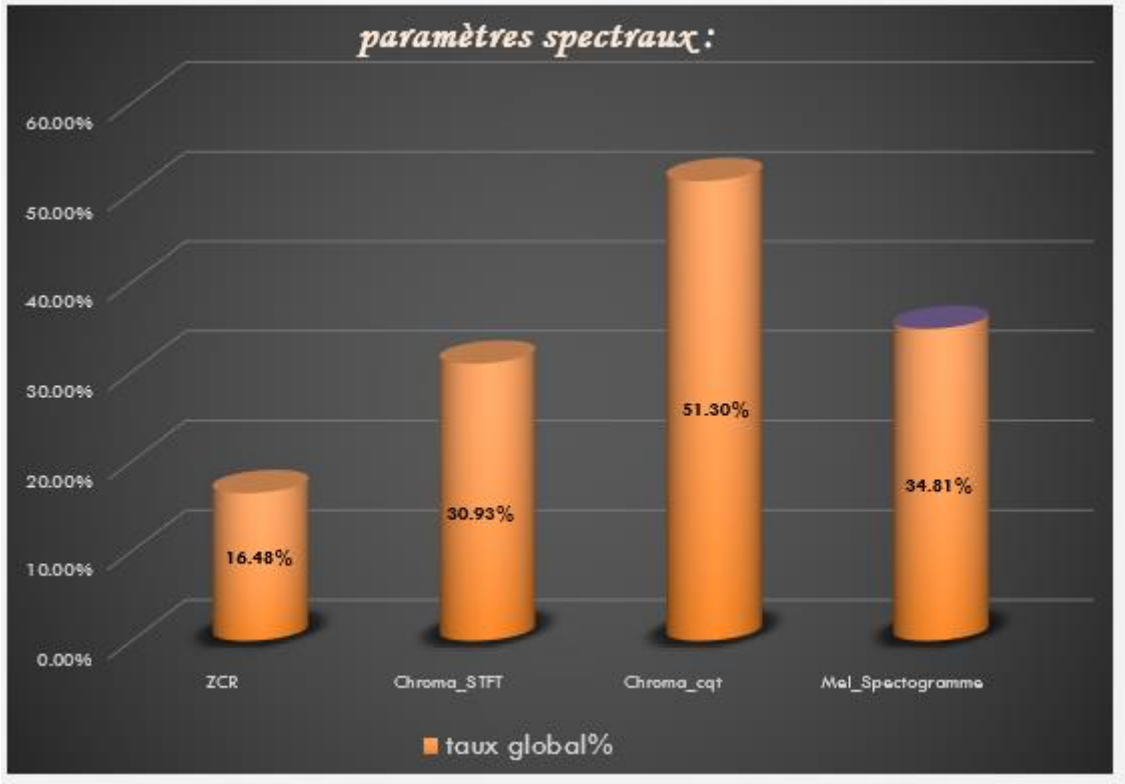

Figure 3.2 : taux correct de reconnaissance de reconnaissance en fonction des paramètres spectraux.

D'après les résultats des paramètres on remarque que Mel Spectrogramme donne un taux correct de reconnaissance le plus élevé.

3.5.1.2. Paramètre prosodique

Dans cette expérience nous avons teste les fréquences fondamentale F-min et F-max de pitch avec plusieurs valeurs pour trouver un meilleur résultat. A la fin en conclu que la valeurs F-min=50Hz et F-max=500Hz donne le meilleur taux correct de reconnaissance.

Le tableau suivant montre l'utilisation individuelle des paramètres prosodiques.

Paramètre	Pitch	RMS
Taux global (%)	22.59	**29,63**

Tableau 3.3 : taux correct de reconnaissance en fonction des paramètres prosodiques.

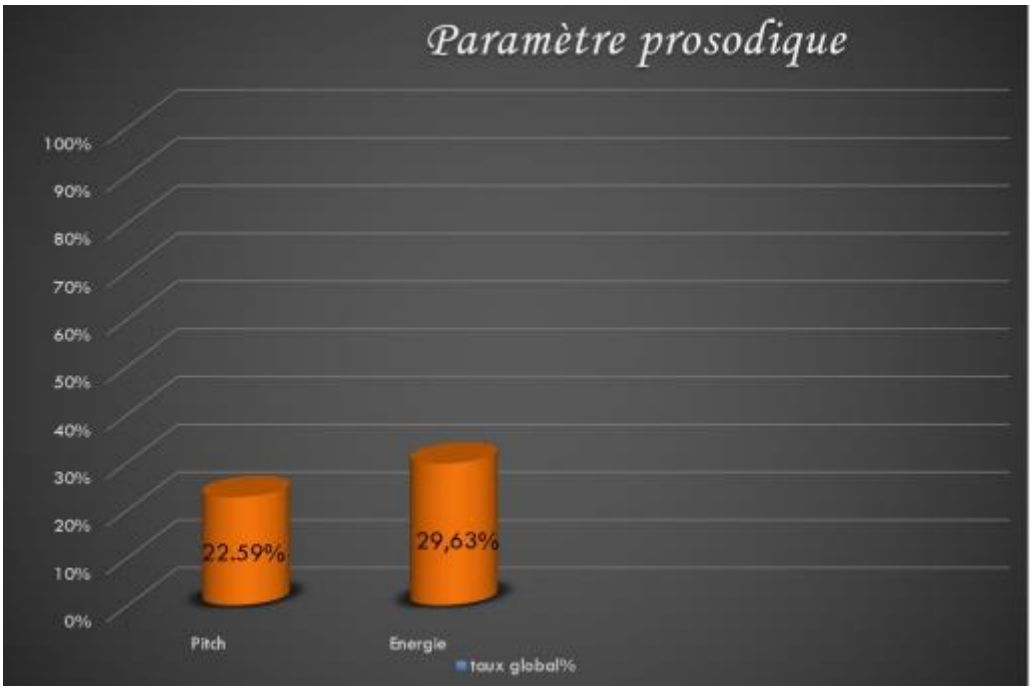

Figure 3.3 : taux correct de reconnaissance en fonction des paramètres prosodiques.

Le résultat montre que les paramètres prosodiques pitch et RMS donne un taux correct de reconnaissance très faible.

3.5.1.3. Fusion des paramètres prosodique

Dans cette expérience, nous exposons les divers tests réalisés en utilisant une combinaison de paramètres prosodiques pitch et RMS, en fixons les fréquences de pitch : F-min=50 et F-max=500.

Le tableau suivant montre résultat de la fusion entre pitch et RMS :

Paramètres	Pitch et RMS
Taux global (%)	**31.11**

Tableau 3.4 : Effet de la diffusion des paramètres prosodiques sur le système RAE.

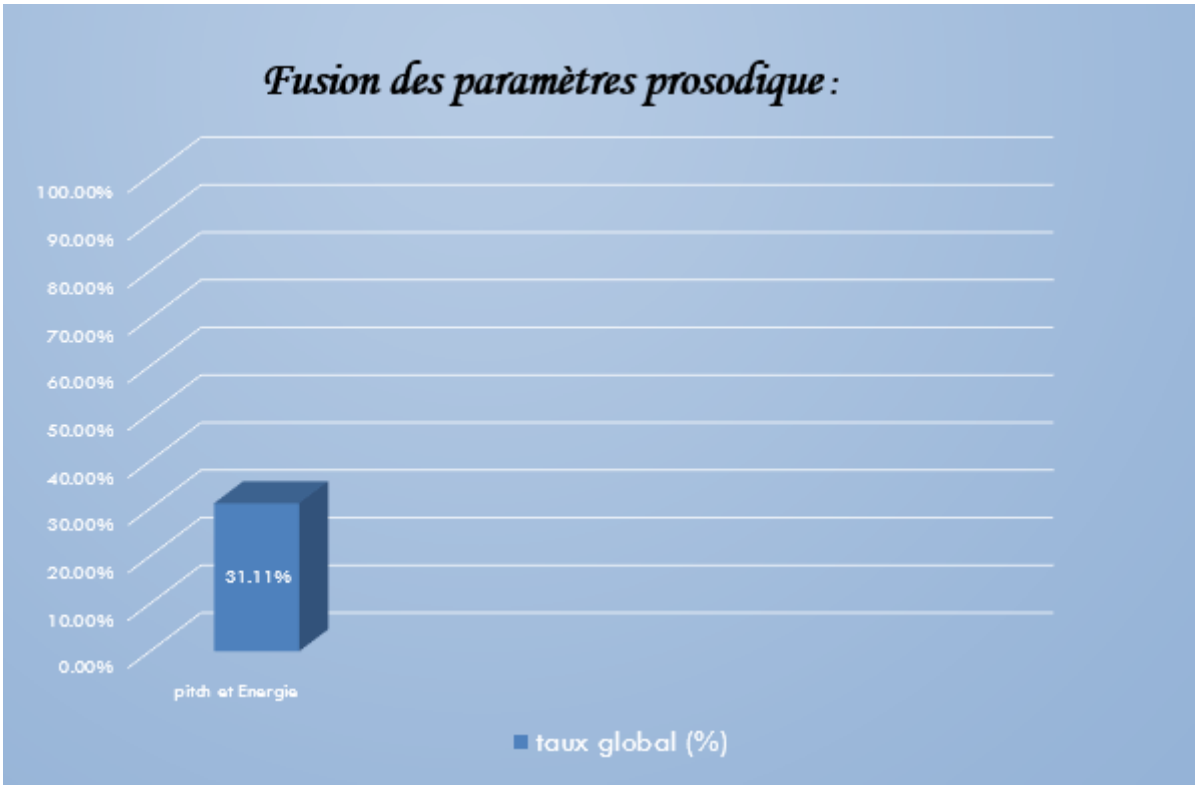

Figure 3.4 : Effet de la diffusion des paramètres prosodiques sur le système RAE.

D'après cette expérience en remarque que, la fusion des paramètres prosodique Pitch-RMS n'améliore pas le taux correct de reconnaissance.

3.5.1.4. Fusion des paramètres spectraux et prosodiques

Dans la section précédente, nous avons teste individuellement les différents types des paramètres. Toutefois, afin d'améliorer notre taux de reconnaissance, il est primordial d'étudier la combinaison des caractéristiques prosodiques et spectrales. À cet effet, nous avons effectué plusieurs combinaisons du paramètre MFCC avec un des paramètres restant puis deux ainsi de suite jusqu'à ce que toutes les combinaisons soient testes. On a constaté que la combinaison deux à deux donnait le meilleur résultat. Dans le paragraphe suivant on présentera ces résultats.

3.5.1.4.1. Fusion de paramètre MFCC avec les paramètres spectraux

Dans cette étape nous avons fixé le coefficient de paramètres MFCC à 20, puis en va tester le MFCC avec chaque paramètre spectral.

Le tableau suivant montre les résultats de la fusion entre les paramètres spectraux :

Fusion des Paramètres	MFCC avec ZCR	MFCC avec Chroma_stft	MFCC avec Chroma_cqt	MFCC avec Mel spectrogramme
Taux global (%)	72.04	73.33	71.82	**74.07**

Tableau 3.5 : taux correct de reconnaissance en fonction de paramètre spectraux.

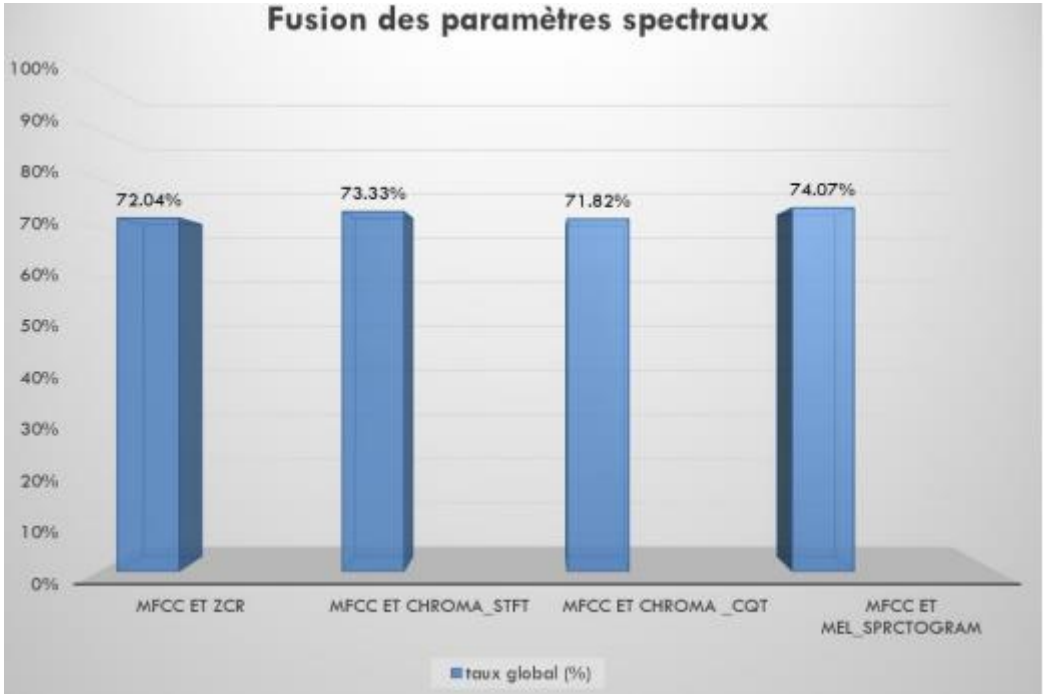

Figure 3.5 : taux correct de reconnaissance en fonction de paramètre spectraux.

On remarque que la fusion entre les paramètres spectraux n'améliore pas le taux correct de reconnaissance.

3.5.1.4.2. Fusion de paramètre MFCC avec les paramètres prosodiques

Dans cette étape nous avons fixé le coefficient de paramètres MFCC à 20 et les fréquences fondamentales de pitch à F-min=50Hz et F-max=500Hz, puis en va tester MFCC avec chaque paramètre prosodique.

Le tableau suivant montre les résultats de la fusion entre les paramètres prosodiques :

Paramètre	MFCC et RMS	MFCC et pitch
Taux global (%)	**76.48**	73.70

Tableau 3.6 : taux correct de reconnaissance en fonction de paramètre prosodique.

On remarque que le meilleur résultat obtenu dans ce cas correspond à la fusion de RMS avec les coefficients MFCC. Cependant, cette fusion améliore le taux correct de reconnaissance, donc en peut dire que l'énergie donne un effet d'amélioration.

3.5.2. Effet des modèles CNN

3.5.2.1. Effet de (karnel_size)

Dans cette expérience nous avons utilisées les deux paramètres MFCC et RMS, puis nous avons modifié la taille du noyau (kernel_size) dans chaque couche de convolution en utilisant les variables suivantes : [1, 3, 5, 7, 9], cela signifie que nous avons testé différentes tailles de noyau pour chaque couche de convolution.

CNN (Karnel_size)	1	3	5	7	9
Taux global (%)	31.11	68.52	**76.48**	75.93	75.55

Tableau 3.7 : taux correct de reconnaissance en fonction de CNN.

Figure 3.7 : taux correct de reconnaissance en fonction de CNN.

Il est intéressant de constater que vous avez remarqué que la taille de noyau (kernel_size) pour chaque couche a un impact sur les résultats de reconnaissance. En fixant la taille de noyau à 5 (kernel_size=5), nous avons obtenu un taux correct de reconnaissance de 76,48%.

3.5.2.2. Effet de pool_size

Dans cette expérience nous avons utilisées les deux paramètres MFCC et RMS, puis nous avons modifié la taille de la fenêtre de pooling dans chaque couche de pooling en utilisant les variables suivantes : [1, 2, 3, 4, 5], cela signifie que nous avons testé différentes tailles de fenêtre pour chaque couche de pooling.

CNN (Pool_size)	1	2	3	4	5
Taux global (%)	84.26	**84.44**	82.41	78.89	76.48

Tableau 3.8 : Effet de CNN (pool_size) sur le taux correct de reconnaissance

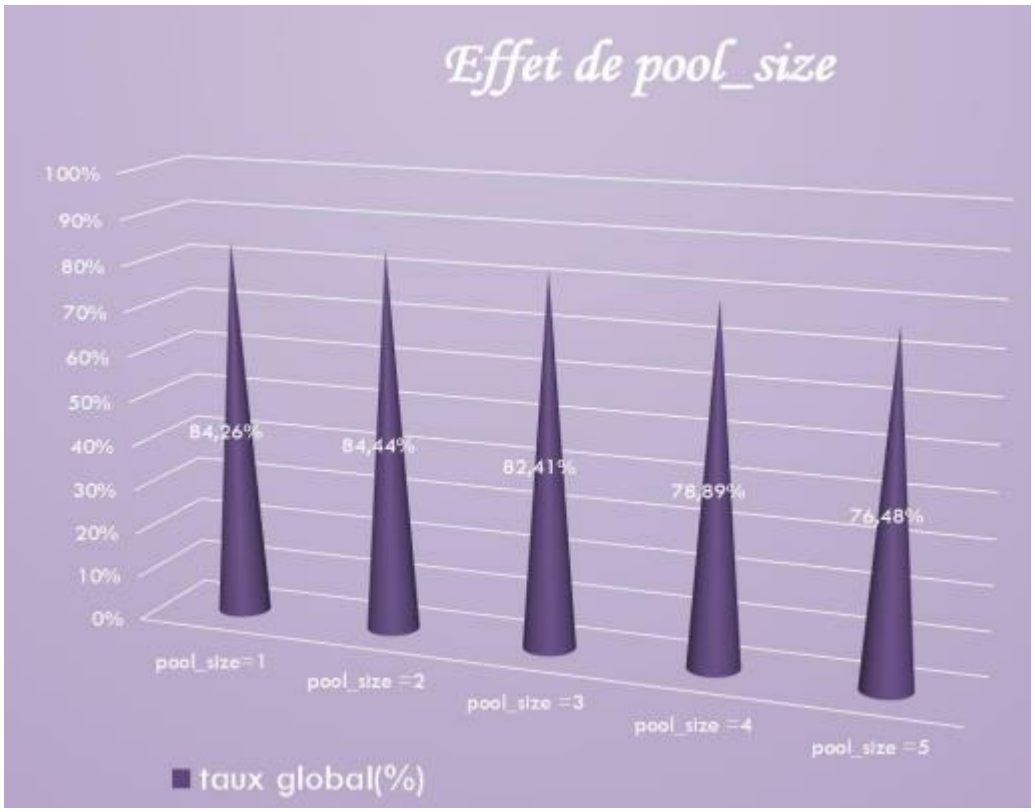

Figure 3.8 : Effet de CNN (pool_size) sur le taux correct de reconnaissance

Il est intéressant de constater que nous avons remarqué que la taille de la fenêtre de pooling a un impact sur les résultats de reconnaissance. En fixant la taille de la fenêtre de pooling à 2 (pool_size=2), nous avons obtenu un taux correct de reconnaissance de 84,44%.

3.5.2.3. Effet de Dropout

Dans cette expérience nous avons utilisées les deux paramètres MFCC et RMS, puis nous avons modifié le dropout (la régularisation et la prévention du surapprentissage) en utilisant les variables suivantes : [0.2, 0.3, 0.4, 0.5].

CNN (Dropout)	0.2	0.3	0.4	0.5
Taux global (%)	74.26	**76.48**	74.26	74.07

Tableau 3.9 : Effet de CNN (Dropout) sur le taux correct de reconnaissance.

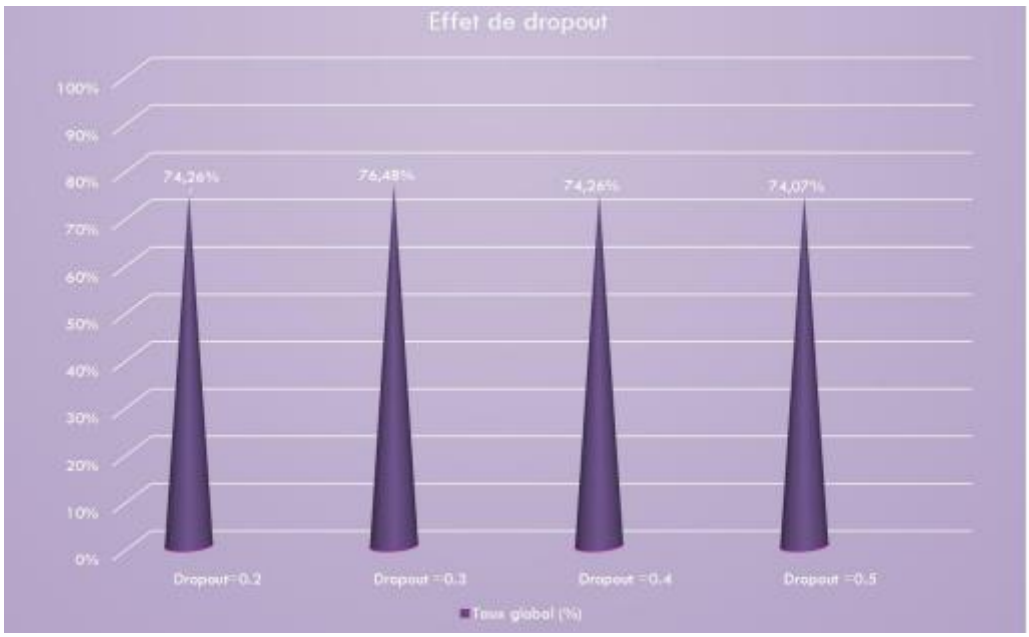

Figure 3.9 : Effet de CNN (Dropout) sur le taux correct de reconnaissance.

Il est intéressant de constater que nous avons remarqué que Dropout a un impact sur les résultats de reconnaissance. En fixant Dropout a 0.3 (Dropout=0.3), nous avons obtenu un taux correct de reconnaissance de 76.48%.

3.5.2.4. Effet de Epochs

Dans cette expérience nous avons utilisées les deux paramètres MFCC et RMS, puis nous avons modifié le nombre d'itération (Epochs), en utilisant les variables suivantes : [30, 40, 50, 60, 70, 80, 90, 100], cela signifie que nous avons testé différents nombres d'itérations pour entraîner votre modèle.

CNN (Epochs)	30	40	50	60	70	80	90	100
Taux global (%)	69.81	74.81	76.48	74.81	77.96	76.48	**78.15**	75.93

Tableau 3.10 : Effet de CNN (Epochs) sur le taux correct de reconnaissance.

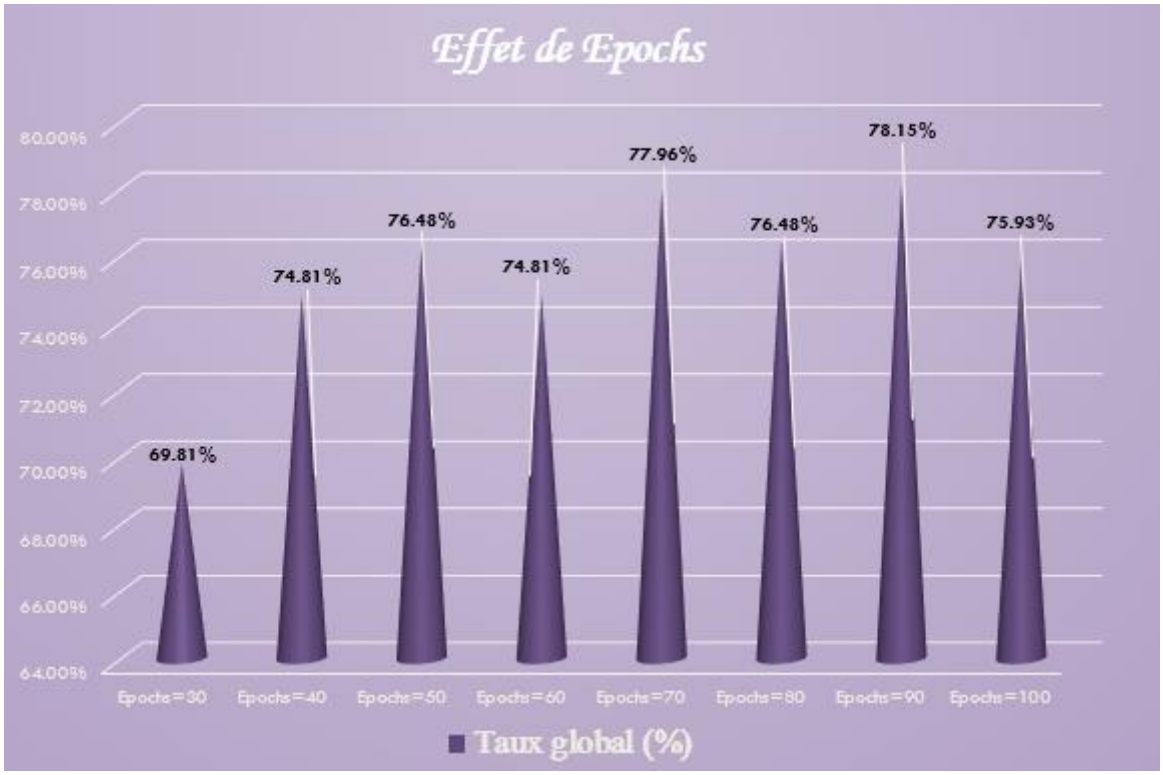

Figure 3.10 : Effet de CNN (Epochs) sur le taux correct de reconnaissance.

Il est intéressant de constater que nous avons remarqué que Epochs a un impact sur les résultats de reconnaissance. En fixant Epochs a 90 (Epochs=90), nous avons obtenu un taux de reconnaissance correct de 78.15%.

3.5.2.5. Effet de la fusion entre les couches CNN

Pour améliorer les résultats de notre expérience précédente, nous avons fait plusieurs testes sur les couches de CNN. Ces tests incluent un meilleur résultat de 85.88% et un résultat plus élevé dans l'epoch 52/80 de 87.04%, avec les coefficients suivants : la taille de noyau 5, la taille de fenêtre de regroupement 1, le taux de dropout 0,3 et le nombre d'époques d'entraînement 80. (Karnel_size=5, pool_size=1, Dropout=0.3, Epochs=80.

Paramètres	Karnel_size=5	Pool_size=1	Dropout=0.3	Epochs=80
Taux global (%)	**85.88%**			

Tableau 3.11 : Effet des paramètres CNN sur le taux correct de reconnaissance.

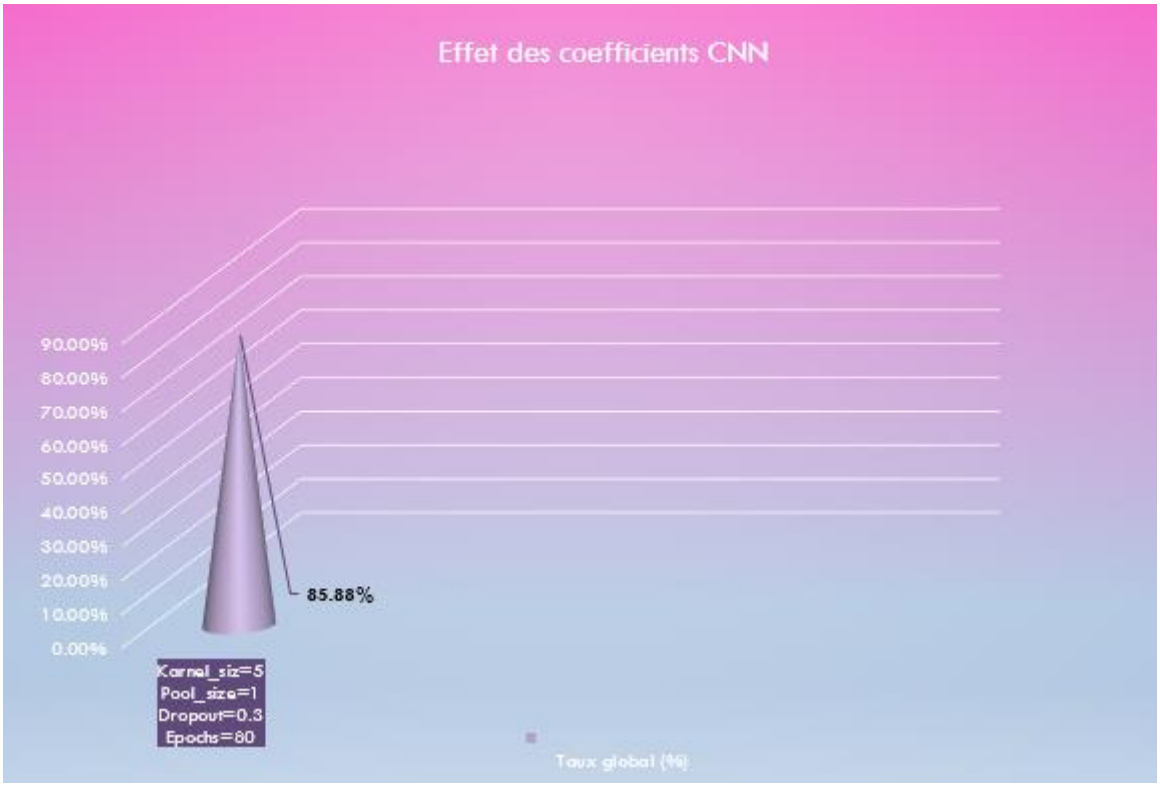

Figure 3.11 : Effet des paramètres CNN sur le taux correct de reconnaissance.

3.5.3. Taux de reconnaissance par émotions

Afin de présenter les résultats de détection individuelle des émotions obtenus grâce à notre système, nous avons mené plusieurs expériences en variant les configurations des paramètres. Nous avons identifié la meilleure configuration de détection en utilisant MFCC=20 et RMS. De plus, nous avons utilisé des paramètres spécifiques pour un réseau de neurones convolutif (CNN), ce qui a abouti à un taux correct de reconnaissance satisfaisant de 85,88% et dans l'epoch 52/80 nous avons trouvé un taux de reconnaissance plus élevé 87.04%.

Les résultats de ces expériences sont présentés dans la matrice qui offre une visualisation des performances du système en termes de détection individuelle des émotions.

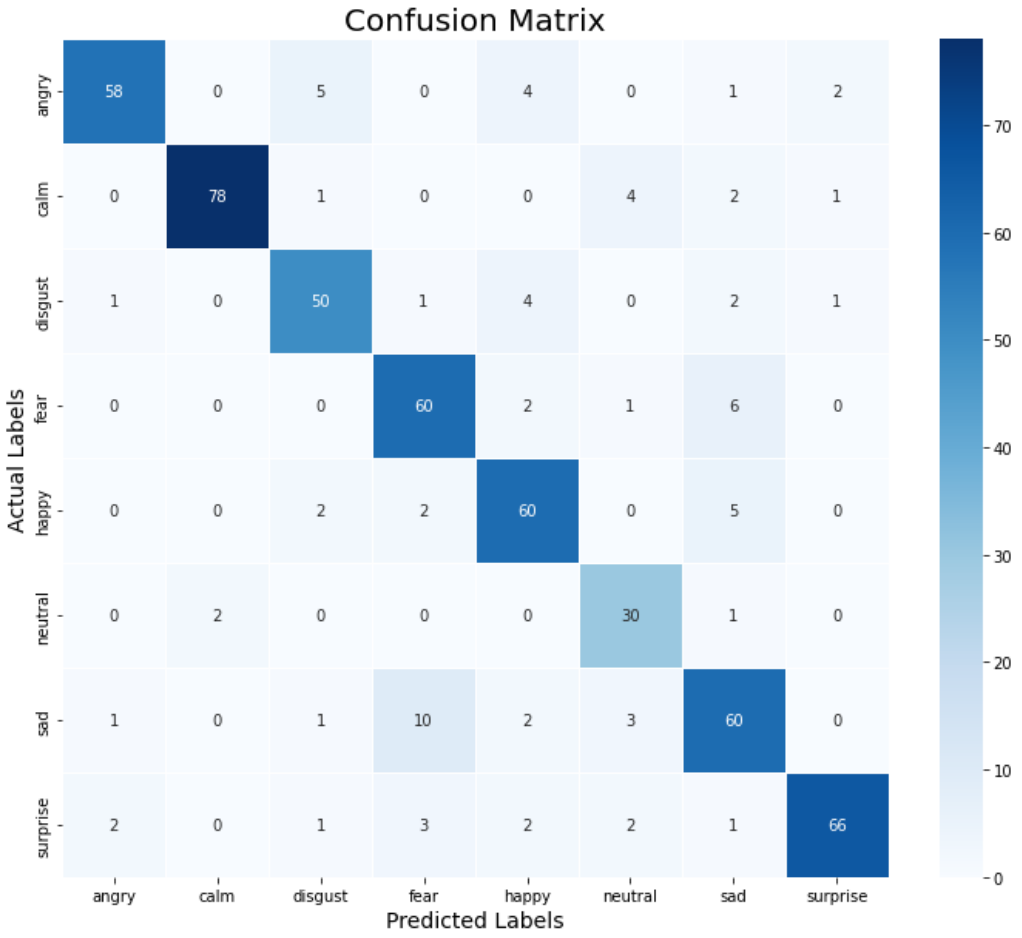

Figure 3.12 : matrice de taux correct de reconnaissance par émotion.

A partir de cette matrice, nous pouvons calculer le taux correct de reconnaissance pour chaque émotion par la méthode suivante :

$$\text{Taux global} = \frac{nombre\ de\ teste\ pour\ émotion}{nombre\ total\ de\ teste\ des\ émotions}\ (\%)$$

Example : Taux global de heureux :

$$\text{Taux global} = \frac{60}{0+0+2+2+60+0+5+0} = \frac{60}{69} = 86.96\%$$

Les résultats pour chaque émotion sont résumés dans le tableau ci-dessous :

Emotions	Colère	Calme	Dégoût	Effrayé	Heureux	Neutre	Triste	Surprise
Taux global (%)	82.86	90.97	84.74	86.96	86.96	90.90	77.92	85.71

Tableau 3.12: Taux correct de reconnaissance par émotion.

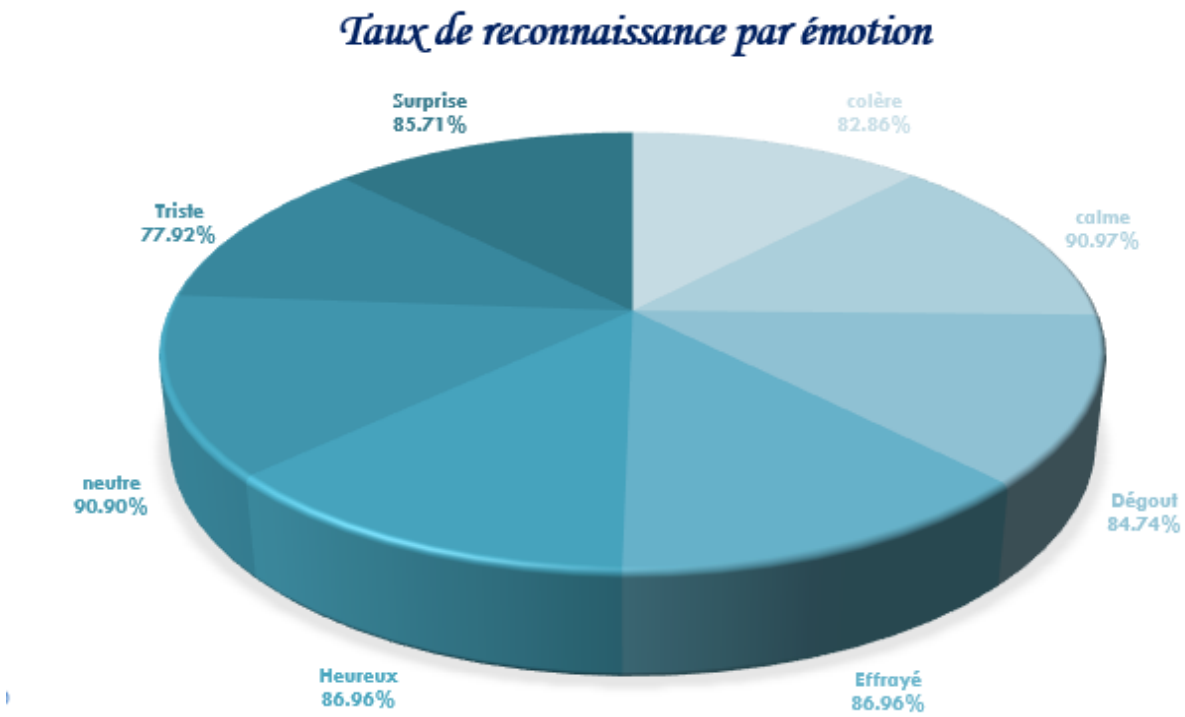

Figure 3.12: Taux correct de reconnaissance par émotion.

D'après les résultats fournis dans la matrice précédente, il est évident que le modèle CNN améliore la capacité à détecter les émotions individuelles. De plus, il est intéressant de noter que le système RAE se distingue par une excellente détection des émotions neutres et calme, avec des taux de reconnaissance de 90,90 % et 90,97 % respectivement. En revanche, il présente une moins bonne performance en ce qui concerne la détection du la tristesse, avec des taux de reconnaissance de seulement 77,92 %.

Pour trouver le taux global de cette expérience à partir de ces résultats de matrice, en fait le calcul suivant :

$$\text{Taux global} = \frac{82.86+90.97+84.74+86.96+86.96+90.90+77.92+85.71}{8} = 85.8775\%$$

3.6. Conclusion

Dans ce chapitre, nous avons présenté notre système de reconnaissance des émotions basé sur l'apprentissage profond à partir de la voix. Ce système a été validé par des phases d'apprentissage et de test en utilisant des techniques basées sur les réseaux CNN. Une étape d'extraction des caractéristiques est effectuée en utilisant des paramètres spectraux tels que MFCC, Chroma_stft, mel spectrogramme et chroma_cqt, ainsi que des paramètres prosodiques tels que l'énergie (RMS) et le pitch. Dans notre système de RAE, nous avons testé la détection des émotions en utilisant ces

paramètres individuellement, ainsi qu'en combinant différentes combinaisons entre eux. Nous avons également exploré différentes configurations du nombre de coefficients et de MFCC avec le modèle CNN dans le but d'améliorer ce système. Les résultats les plus prometteurs ont été obtenus en utilisant la combinaison de MFCC et l'énergie (RMS) avec le modèle CNN qui donne un taux correct de reconnaissance 85,88%.

Conclusion Générale

Dans notre travail, nous avons initialement développé un système complet de reconnaissance automatique de l'état émotionnel d'un locuteur. Dans cette optique, nous avons entrepris une étude théorique qui a débuté par la présentation des concepts fondamentaux liés aux émotions et à la parole dans le premier chapitre. Par la suite, dans le deuxième chapitre, nous avons examiné en détail les différentes étapes nécessaires à la mise en place du système de reconnaissance automatique d'émotions, telles que l'extraction des caractéristiques acoustiques à l'aide des paramètres spectraux (MFCC, ZCR, STFT, CQT et le spectrogramme), ainsi que des paramètres prosodiques (énergie, et pitch). De plus, nous avons étude l'apprentissage profonde en se basant sur les réseaux de neurone convolutifs (CNN).

Dans le dernier chapitre, nous nous sommes concentrés sur la mise en pratique du système RAE en utilisant le langage de programmation Python. Comme pour tout système de reconnaissance, une étape préliminaire consistait à organiser les données en trois parties distinctes : une partie de développement, une partie d'apprentissage et une partie de test.

Ensuite, nous avons procédé à l'extraction des caractéristiques acoustiques en utilisant à la fois des paramètres prosodiques et spectraux. Cela nous a permis de capturer des informations pertinentes pour décrire les aspects temporels et fréquentiels des signaux audios liés aux émotions.

Par la suite, nous avons créé des modèles d'état émotionnel pour chaque apprenant en utilisant la technique du réseau de neurones convolutifs (CNN). Cette approche basée sur les CNN nous a permis de bénéficier des avantages de l'apprentissage profond et de la capacité des CNN à extraire des caractéristiques significatives des données audios.

Enfin, nous avons testé les émotions en utilisant le système RAE et évalué ses performances en mesurant le taux de reconnaissance correct. Cette évaluation nous a permis de quantifier l'efficacité et la précision du système dans la reconnaissance des états émotionnels.

Différentes expérimentations réalisées sur le système nous ont permis de constater que le système MFCC est le plus performant. Avec un nombre de coefficients de 20, le taux de reconnaissance correct atteint 75,93%. Lors de l'utilisation du paramètre spectral obtenu à l'aide du paramètre Mel Spectrogramme, le taux de reconnaissance correct le plus élevé est de 51,30 % et lors de l'utilisation des paramètres généraux, il donne le taux de reconnaissance correct de 29,63 % avec RMS. De plus, la fusion du MFCC avec les paramètres spectraux donne le meilleur résultat de 76,48% lorsqu'on le combine au paramètre RMS.

L'effet des modèles CNN sur les performances du système de reconnaissance a été aussi considéré. Nous avons modifié la taille du noyau (kernel_size) à chaque fois, et nous avons obtenu le

meilleur taux de 76,48 %. Le changement du Dropout nous a conduit à 76,48% de taux de reconnaissance correct. L'effet Epoch donne un taux de reconnaissance correct de 78,15%. Lors de la modification dans la fenêtre de pooling, nous obtenons de taux de reconnaissance correct avec 84,44%. A la fin nous avons fixe karnel_size a 5, pool_size a 1, dropout a 0.3 et epoch à 80 le meilleur résultat que nous avons obtenu dans notre expérience est 85,88%.

La reconnaissance automatique des émotions basée sur l'apprentissage profond offre de larges perspectives pour améliorer notre compréhension des émotions humaines et pour développer des applications scientifiques dans de nombreux domaines.

Les développements futurs dans ce domaine sont susceptibles de permettre des avancées majeures dans la façon dont nous interagissons les uns avec les autres et avec la technologie.

Références bibliographiques

[1] **:** Kerkeni.L, " Analyse acoustique de la voix pour la détection des émotions du locuteur. Vision par ordinateur et reconnaissance de formes", thèse de doctorat, Université du Maine, Université du Centre, Tunisie, 2020.

[2] : Silem, A., & Daoui, H. (2022). " Reconnaissance automatique des émotions par la voix " mémoire de master, université Akli Mohand oulhadj-bouira.

[3] : Allache, O., & Habache, S. (2021). " Détection et Analyse de l'Etat Emotionnel du Locuteur" mémoire de master, université Akli Mohand oulhadj-bouira

[4] : Yazid, A. (2008). Reconnaissance automatique des émotions à partir du signal acoustique (Doctoral dissertation, PhD thésis, Ecole de technologie supérieure, Université du Québec).

[5] : J. SASSERATH, "Impact de l'intensité émotionnelle sur la hauteur tonale en fonction de l'étendue fréquentielle des participants", thèse de Master, Faculté des Sciences Psychologiques et de l'Éducation Unité de logopédie de la voix, Université de Liège, 2017 – 2018.

[6] : KOTSOU, I., & ALTENLOH, E. (2013). Les émotions positives dans le monde des organisations. Psychologie positive en environnement professionnel, 177.

[7] : X. Hung Le, "Indexation des émotions dans les documents audiovisuels à partir de la modalité auditive", thèse de doctorat, Institut National Polytechnique de Grenoble, Institut Polytechnique de Hanoi, 2009.

[8] : Hamza Hamdi Mémoire Présenté en vue de l'obtention du grade de Docteur de l'Université d'Angers sous le label de L'Université de Nantes Angers Le Mans Discipline : Informatique Spécialité : Laboratoire : Laboratoire d'Ingénierie des Systèmes Automatisés (LISA) Soutenue le 03 décembre 2012.

[9] : Mauri.A (2022). Prédiction et évitement d'obstacle bases Deep Learning :application à la mobilité ferroviaire et routière (Doctoral dissertation, Normandie université).

[10]: F. Parke. A parametric model for human faces. page 109. The University of Utah, 1974. 21.

[11] : https ;/www.cairn.info. Consulte le 13/04/2023 à 23 :13h.

[12] : Trabelsi, A. (2010). Configuration et exploitation d'une machine émotionnelle.

[13] : Menahem, R. (1983). La voix et la communication des affecta. L'année psychologique, 83(2), 537-560.

[14]: ZERDALI, A. O., HAMMA, S. E., & KROBBA, A. (2021). Étude de la performance du système de reconnaissance du locuteur basé sur des paramètres acoustiques multi-résolution en communication mobile.

[15] : R. Boite, H.bourlard, T.Dutoit, J.Hancq et H.Leich, "Traitement de parole", Presses Polytechnique universitaires Romandes, Lausanne,2000.

[16] : Intérêt de la musicothérapie pour les troubles de l'expression et de la reconnaissance des émotions vocales et faciales dans la maladie de Parkinson afin d'améliorer les capacités communicationnelles. 2 Mémoire présenté et soutenu par Lara Vieira Universités de Paris Descartes - Septembre 2021-Juin 2022.

[17] : https//www.universalis.fr/ consulte le 13/04/2023 à 23 :20h.

[18] : https://deci.ai/deep-learning-glossary/deep-neural-network-dnn/ consulte le 15/03/2023 à 14 :36h.

[19] : https ;/www.cairn.info. Consulte le 13/04/2023 à 23 :13h.

[20] : http://www.ac-grenoble.fr/ecoles/vienne2/IMG/pdf/la_respiration-2.pdf. Consulte le 29/04/2023 à 20 :49h.

[21] : https://www.em-consulte.com/article/19333/physiologie-de-la-phonation. Consulte le 29/04/2023 à 21 :50h.

[22] : Résonance sonore et cavités supralaryngées Alain Ghio, Serge Pinto 2007 - hal. Science.

[23] : Ikram, N. (2022). implémentation des Détecteurs de contours à Base de réseau Neurones Convolutifs CNN (Doctoral dissertation, faculté des sciences et de la technologie univ bba).

[24] : Jemâa, I. (2013). Suivi de Formants par analyse en Multirésolution (Doctoral dissertation, Université de Lorraine; Faculté des Sciences de Tunis).

[25] : https://mbamci.com/la-reconnaissance-des-emotions-par-lia-bonne-ou-mauvaise-idee Consulter le 17/02/2023.

[26] : Caroline ETIENNE APPRENTISSAGE PROFOND appliqué à la reconnaissance des émotions dans la voix. Paris _sud 18 décembre 2019.

[27] : A.Hacine Gharbi,"Sélection de paramètres acoustiques pertinents pour la reconnaissance de la parole", thèse de doctorat, à la Faculté de Technologie Département d'Électronique, Université Ferhat Abbas-Sétif Algérie, Soutenue le 09 Décembre 2012.

[28]: Logan, B. (2000, October). Mel frequency cepstral coefficients for music modeling. In Ismir (Vol. 270, No. 1, p. 11).

[29]: Huang, X., Acero, A., Hon, H. W., & Reddy, R. (2001). Spoken language processing: A guide to theory, algorithm, and system development. Prentice hall PTR.

[30] : Mohammed Senoussaoui Amelioration De La Robustesse Des Systemes De Reconnaissance Automatique Du Locuteur Dans L'espace Des I-Vecteurs Montreal, Le 10 Juin 2014.

[31] : TIGZIRI née IBRAHIM Noura « Reconnaissance automatique de la parole dans un contexte multilocuteur ».

[32] : Madame AMRANE – STIET Malika. « Paramètrisation et segmentation du signal de parole » UMMTO-2002. (Bibliothèque centrale Hasnaoua).

[33] : https://vincmazet.github.io/signal1/fourier/temps-frequence.html CONSULTER le 6 juin 2023 à 13 :06

[34]: Deep Learning and Mel-spectrograms for Physica Violence Detection in Audio Tiago B. Lacerda1, Pericles Miranda ´ 2, Andre C´ amara ˆ 2, Ana Paula C. Furtado1,2. Consulter le 18/06/2023 à 13:39h

[35] : https://doc.ml.tu-berlin.de/bbci/material/publications/Bla_constQ.pdf. Consulte le 13/04/2023 à 15 :38h

[36] : Boughba Mohammed et Boukhris Brahim « l'apprentissage profond (Deep Learning) pour la classification et la recherche d'image par le contenu »universite KASDI MERBAH OUARGLA 2016/2017

[37] : Abdelaziz, H. A. B. B. A., Omar, I. S. H. A. K., & OUAHAB, A. (2019). La classification des images satellitaires par l'apprentissage profonde (deep learning) (Doctoral dissertation, Université Ahmed Draïa-Adrar.

[38]:https://www.sales-hacking.com/post/intelligence-artificielle-vs-machine-learning-vs-deep-learning

[39] : https://www.netapp.com/fr/artificial-intelligence/what-is-deep-learning/ Consulte le 11/03/2023 à 15 :33h.

[40]: Mustaqeem, and Soonil Kwon. "A CNN-assisted enhanced audio signal processing for speech emotion recognition." Sensors 20.1 (2019) : 183. Consulte le 06/06/2023 à 10 :34h.

[41]: https://www.futura-sciences.com/tech/definitions/intelligence-artificielle-deep-learning-17262/. Consulte le 11/03/2023 à 15 :43h.

[42] : Rebahi Ghediri Imane Semri KhawLa « La Reconnaissance Des émotions de base Par Les réseaux de neurones : Application de Deep Learning » Université L'arbi Ben M'Hidi Oum El Bouaghi.

[43] : ZOUBIRI Sefouane HAFSI Mohamed Abderrahmane. Reconnaissance des émotions par l'analyse visuelle du visage. UNIVERSITE YAHIA FARES DE MEDEA.

[44]: https :// datascientest .com /deep learning. Consulte le 13/03/2023.

[45] : https ://www.actuia.com. Consulte le 20/03/2023 à 21 :30h

[46] : http://larevueia.fr/apprentissage par renforcement Consulte le 14/03/2023 à 11 :13h

[47] : https ://www.synopsys.com/ai/what-is-reinforcement-learning. Consulte le 21/03/2023 à 7 :39h

[48] : Learning pour et par les nuls C. Ambroise. Rennes. Juillet 2018.

[49] : HTTPS:// kongakur.fr/article/le perception-multi couche- Deep Learning Consulte le 14/03/2023.

[50] : Youssef BARKAOUI CLASSIFICATIO PROFOND ET RESEAUX DE NEURONS, APPLICTION EN SCIENCE DES DONN2ES. Mars 2022.université du Qubec.

[51] : Samira EBRAHIMI KAHOU. « Émotion recognition white Deep neural network » université de Montréal, juillet 2016.

[52] : Tristan sérine. Réseau de neurones récurrents et mémoire application a la musique septembre 2016.

ملخص

التواصل هو أحد أكثر الوسائل انتشارًا للبشر للتعبير عن حالاتهم العاطفية الداخلية. لذلك، سيكون من المثير للاهتمام تطوير نظام قادر على التعرف تلقائيًا على هذه المشاعر. في مشروعنا، نركز على إنشاء نظام التعرف على المشاعر الصوتية باستخدام تقنيات

التعلم العميق. يعتمد النظام على استخدام العديد من المعلمات الطيفية مثل MFCC، وZCR، وChroma_stft، وchroma_cqt، وMel Spectrogramme، بالإضافة إلى معلمات عرضية مثل RMS (الطاقة) والخطوة(pitch). يتم تمثيل كل نوع من أنواع المشاعر بواسطة نموذج CNN.

من أجل تحسين أداء النظام، أجرينا العديد من الاختبارات لتحديد أعلى معدل التعرف. من حيث المعلمات، حصل الجمع بين تقنيات MFCC وRMS على أفضل النتائج بمعدل التعرف 84.44٪. وجدنا أيضًا أنه من خلال دمج المعلمات العروضية والطيفية مع نموذجCNN، تمكنا من زيادة تحسين أداء النظام، وتحقيق معدل التعرف بنسبة 85.88%.

الكلمات المفتاحية: التعرف، العواطف، المعلمات العروضية، المعلمات الطيفية، CNN

Résumé

La communication est l'un des moyens les plus répandus chez les êtres humains pour exprimer leurs états émotionnels internes. Par conséquent, il serait intéressant de développer un système capable de reconnaître automatiquement ces émotions. Dans notre projet, nous nous concentrons sur la création d'un système de reconnaissance des émotions vocales en utilisant des techniques d'apprentissage profond. Le système repose sur l'utilisation de plusieurs paramètres spectraux tels que les MFCC, ZCR, Chroma_stft, chroma_cqt, Mel Spectrogramme, ainsi que des paramètres prosodiques tels que RMS (énergie) et pitch. Chaque type d'émotion est représenté par un modèle CNN.

Afin d'optimiser les performances du système, nous avons réalisé de nombreux tests pour déterminer le taux de reconnaissance le plus élevé. En termes de paramètres, la combinaison des techniques MFCC et RMS a obtenu les meilleurs résultats avec un taux de reconnaissance de 84,44 %. Nous avons également constaté qu'en fusionnant les paramètres prosodiques et spectraux avec le modèle CNN, nous avons pu améliorer davantage les performances du système, atteignant ainsi un taux de reconnaissance de 85,88 %.

Mots clés : Reconnaissance, émotions, paramètres prosodiques, paramètres spectraux, CNN.

Abstract

Communication is one of the most widespread means for human beings to express their internal emotional states. Therefore, it would be interesting to develop a system capable of automatically recognizing these emotions. In our project, we focus on creating a voice emotion recognition system using deep learning techniques. The system is based on the use of several spectral parameters such as MFCC, ZCR, Chroma_stft, chroma_cqt, Mel Spectrogramme, as well as prosodic parameters such as RMS (energy) and pitch. Each type of emotion is represented by a CNN model.

In order to optimize the performance of the system, we have carried out numerous tests to determine the highest recognition rate. In terms of parameters, the combination of MFCC and RMS techniques obtained the best results with a recognition rate of 84.44%. We also found that by merging the prosodic and spectral parameters with the CNN model, we were able to further improve system performance, achieving a recognition rate of 85.88%.

Keywords: Recognition, emotions, prosodic parameters, spectral parameters, CNN.