

Filière : Master en informatique et télécommunications
Projet de fin d'étude

Analyse acoustique de la voix pour la détection automatique des émotions

Réalisé par :

Mr. Abdellah Agnaou

Soutenu le 25 Septembre 2024 devant le jury :

Pr. HOUARI Soumaya	Professeur à la faculté des sciences Rabat	Encadrante
Pr. BELMAJDOUB Hanae	Professeur à la faculté des sciences Rabat	Encadrante
Pr. MINAOUI Khalid	Professeur à la faculté des sciences Rabat	Examinateur
Mme. Hmaida Sanaa	Doctorante à la faculté des sciences Rabat	Examinateur

Année universitaire : 2023-2024

REMERCIEMENTS

C'est grâce à Dieu que tout a commencé, et c'est à lui que nous rendons grâce.

Ce n'est pas la réalisation d'un tel travail qui exige un remerciement, si nous remercions quelqu'un c'est par ce qu'il mérite.

Mes remerciements vont en premier lieu à mon encadrant, Madame **Soumaya Houari** professeur à la FSR Rabat, pour la qualité de son encadrement, ses conseils, son enthousiasme, sa disponibilité, ses encouragements et son attention du début à la fin de mon projet de fin d'études.

Mes vifs remerciements vont à mon encadrante Madame **Hanae Belmajdoub** pour son aide précieuse, sa disponibilité, sa gentillesse et ses précieuses directives qui ont permis une bonne orientation dans la réalisation de ce modeste travail.

Je remercie également les honorables membres du jury qui ont accepté de juger ce travail.

Mes remerciements s'adressent également à tous les professeurs de la faculté des sciences de Rabat généralement et à tous les professeurs du Master Informatique et Télécommunications spécialement.

Finalement, je remercie toutes les personnes qui m'ont apporté leurs encouragements et qui ont contribué de près ou de loin à la réalisation de ce projet.

DÉDICACE

*Je dédie ce travail à mes chers parents.
À ma soeur et mon frère.
À toute la famille.
À tous mes amis.
À toute la promotion MIT 2018.
À toute la promotion MIT 2020.
À toute l'aimable promotion MIT 2021.
À toute l'aimable promotion MIT 2023.
À toute les personnes de la faculté.*

Je dédie ce travail

RÉSUMÉ

Ce rapport présente un système d'analyse acoustique de la voix pour la détection automatique des émotions, structuré autour de plusieurs sections principales. La première partie introduit les caractéristiques fondamentales de la parole, en abordant les différences entre les sons voisés et non voisés, et en décrivant l'appareil phonatoire humain. Elle explore également les caractéristiques acoustiques essentielles telles que l'énergie, l'amplitude, l'intensité, et le pitch (tonalité), qui sont cruciales pour l'analyse de la voix.

La deuxième partie détaille les étapes nécessaires à la reconnaissance automatique des émotions à partir des signaux vocaux. Elle couvre les techniques de prétraitement du signal, l'extraction de caractéristiques pertinentes telles que les MFCC et les caractéristiques prosodiques, ainsi que l'utilisation d'algorithmes d'apprentissage automatique et profond pour la classification des émotions.

Enfin, la troisième partie se concentre sur la méthodologie utilisée pour construire et entraîner les modèles de détection des émotions. Elle décrit les critères d'évaluation des performances, analyse les résultats obtenus à partir des jeux de données d'entraînement et de test, et discute de la robustesse des modèles face aux variations inter-orateurs et intra-orateurs, ainsi qu'aux conditions environnementales.

Mots-clés

Détection automatique des émotions, analyse acoustique, signal vocal, traitement de la parole, MFCC, apprentissage automatique, réseaux de neurones, classification, reconnaissance des émotions.

ABSTRACT

This report presents an innovative acoustic voice analysis system designed for automatic emotion detection. Focusing on the fundamental characteristics of speech, the report examines the differences between voiced and unvoiced sounds and provides a detailed description of the human phonatory apparatus. It also explores essential acoustic features such as energy, amplitude, intensity, and pitch, which are crucial for in-depth voice analysis.

The report details the steps required for emotion recognition from vocal signals, including signal preprocessing techniques, extraction of relevant features such as MFCCs and prosodic features, and the application of machine learning and deep learning algorithms for emotion classification.

Finally, the methodology used to build and train emotion detection models is thoroughly examined. Performance evaluation criteria are described, and the results are analyzed to assess the robustness of the models against inter-speaker and intra-speaker variations as well as changing environmental conditions. This report provides a comprehensive overview of the challenges and approaches associated with automatic emotion detection through vocal analysis.

Keywords

Automatic emotion detection, acoustic analysis, vocal signal, speech processing, MFCC, machine learning, neural networks, classification, emotion recognition.

NOMENCLATURE

ANN : Artificial Neural Networks

CNN : Convolutional Neural Networks

CREMA-D : Crowd-sourced Emotional Multimodal Actors Dataset

CSS : Cascading Style Sheets

DCT : Discrete Cosine Transform

EMO-DB : Berlin Database of Emotional Speech

FFT : Fast Fourier Transform

IDE : Integrated Development Environment

IEMOCAP : Interactive Emotional Dyadic Motion Capture Database

KNN : k-Nearest Neighbors

MFCC : Mel-Frequency Cepstral Coefficients

MSP-IMPROV : Multimodal Spontaneous Emotion Corpus

PC : Personal Computer

RAVDESS : Ryerson Audio-Visual Database of Emotional Speech and Song

ReLU : Rectified Linear Unit

RMSE : Root Mean Square Energy

ROC : Receiver Operating Characteristic

Tanh : Hyperbolic Tangent

UML : Unified Modeling Language

ZCR : Zero Crossing Rate

TABLE DES MATIÈRES

1	Introduction au signal de parole	17
1.1	Appareil phonatoire humain	18
1.1.1	Organes de la phonation	19
1.1.2	Cordes vocales	20
1.1.3	Cavités de l'appareil phonatoire	21
1.1.4	Points d'articulation	22
1.2	Caractéristiques de la voix	23
1.2.1	Fréquence fondamentale	24
1.2.2	Énergie	24
1.2.3	Spectre	25
1.3	Troubles de la voix	26
1.3.1	Troubles organiques	26
1.3.2	Troubles fonctionnels	26
1.3.3	Troubles psychogènes	26
1.4	Évaluation des voix acoustiques	27
1.4.1	Méthodes objectives	27
1.4.2	Méthodes subjectives	28

2 Détection automatique des émotions avec la voix	29
2.1 Base de données des voix émotionnelles	30
2.2 Pré-traitement de données	30
2.3 Techniques d'analyse des caractéristiques des signaux audios	33
2.3.1 Signal d'onde	33
2.3.2 Spectrogrammes	34
2.3.3 Chroma	35
2.4 Extraction des caractéristiques	37
2.4.1 MFCC (Mel-Frequency Cepstral Coefficients)	37
2.4.2 Root Mean Square Energy (RMSE)	39
2.4.3 Zero Crossing Rate (ZCR)	39
2.5 Classification	40
2.5.1 Apprentissage automatique	40
2.5.2 Forêt aléatoire	43
2.5.3 k-Nearest Neighbors (k-NN)	44
2.5.4 Apprentissage profond	46
2.6 Méthodes d'évaluation	50
2.6.1 Matrice de confusion	50
2.6.2 Courbe ROC	51
2.6.3 Validation croisée	53
2.6.4 Calcul de confiance	54
2.7 Exemples des travaux en détection automatique des émotions avec la voix	54
3 Contexte expérimental et résultat	57
3.1 Environnement de travail	58
3.1.1 Technologies utilisées pour le développement du site web de détection des émotions	58
3.1.2 Outils et bibliothèques utilisés pour la création du modèle de détection des émotions	59
3.2 Conception de l'application pour la détection des émotions	61

3.2.1	Diagramme de classe	61
3.2.2	Diagramme de cas d'utilisation	62
3.3	Bases de données utilisées	64
3.4	Représentation des caractéristiques pour chaque émotion	66
3.4.1	Représentation des caractéristiques de l'émotion " Neutral "	66
3.4.2	Représentation des caractéristiques de l'émotion " Happy "	67
3.4.3	Représentation des caractéristiques de l'émotion " Sad "	68
3.4.4	Représentation des caractéristiques de l'émotion " Angry "	69
3.4.5	Représentation des caractéristiques de l'émotion " Fear "	70
3.4.6	Représentation des caractéristiques de l'émotion " Disgust "	71
3.4.7	Représentation des caractéristiques de l'émotion " Surprise "	72
3.5	Augmentation de données	73
3.6	Expérimentations et résultats	74
3.6.1	Techniques d'apprentissage automatique utilisées	74
3.6.2	Techniques d'apprentissage profond utilisées	78
3.7	Interfaces utilisateurs pour une application de détection des émotions par la voix	86
3.7.1	Page d'accueil avant authentification	86
3.7.2	Page d'accueil après authentification	88
3.7.3	Page pour enregistrer un audio	89
3.7.4	Page pour importer un audio	89
3.7.5	Affichage de résultat	90
Annexe		98

TABLE DES FIGURES

1.1	Appareil phonatoire humain [1]	18
1.2	Organes de la phonation [2]	19
1.3	Cordes vocales [2]	20
1.4	Cavités de l'appareil phonatoire [3]	21
1.5	Points d'articulation [4]	22
1.6	Fréquence fondamentale [5]	24
1.7	Spectres [5]	25
1.8	Troubles de la voix [7]	27
2.1	Normalisation et échantillonnage [11]	31
2.2	Signal avec le bruit [12]	31
2.3	Segmentation d'un audio [13]	32
2.4	Augmentation des données audio [14]	32
2.5	Normalisation des caractéristiques audios [15]	33
2.6	Signal d'onde [16]	34
2.7	Spectrogramme [16]	35
2.8	Architecture de chroma [17]	36
2.9	Architecture de MFCC [18]	38
2.10	Illustration de RMSE [20]	39
2.11	Illustration de ZCR [22]	40

2.12	Architecture de forêt aléatoire [31]	44
2.13	Architecture de KNN [32]	45
2.14	Architecture de ANN [33]	49
2.15	Architecture de CNN [33]	50
2.16	Matrice de confusion [35]	51
2.17	Courbe ROC [36]	52
2.18	Validation croisée [37]	53
3.1	Bootstrap [44]	58
3.2	Flask [45]	58
3.3	Kaggle [46]	59
3.4	Python [47]	60
3.5	Tensorflow [48]	60
3.6	Librosa [49]	61
3.7	Diagramme de classe	62
3.8	Diagramme de cas d'utilisation	62
3.9	Architecture générale de l'application	63
3.10	Distribution des enregistrements d'émotions après regroupement . .	65
3.11	Caractéristiques d'émotion " Neutral "	66
3.12	Caractéristiques d'émotion " Happy "	67
3.13	Caractéristiques d'émotion " Sad "	68
3.14	Caractéristiques d'émotion " Angry "	69
3.15	Caractéristiques d'émotion " Fear "	70
3.16	Caractéristiques d'émotion " Disgust "	71
3.17	Caractéristiques d'émotion " Surprise "	72
3.18	Augmentation de donnée audio	74
3.19	Matrice de confusion résultant de classification par KNN ($k=1$) entre les classes des émotions	75
3.20	Courbe ROC résultant de classification par KNN ($k=1$) entre les classes des émotions	76

3.21	Matrice de confusion résultant de classification par forêt aléatoire entre les classes des émotions	77
3.22	Courbe ROC résultant de classification par forêt aléatoire entre les classes des émotions	77
3.23	Matrice de confusion résultant de classification par ANN entre les classes des émotions	79
3.24	Courbe ROC résultant de classification par ANN entre les classes des émotions	80
3.25	Courbe de précision & de perte	83
3.26	Matrice de confusion résultant de classification par CNN entre les classes des émotions	83
3.27	Courbe ROC résultant de classification par CNN entre les classes des émotions	84
3.28	Comparaison des étiquettes prédites et réelles pour un modèle de classification des émotions avec CNN	85
3.29	Page d'accueil avant authentification	86
3.30	Page pour inscription	87
3.31	Page pour connexion	88
3.32	Page d'accueil après authentification	88
3.33	Page d'enregistrement d'un audio	89
3.34	Page d'importation d'un audio	90
3.35	Affichage de résultat	91

LISTE DES TABLEAUX

3.1	Résultats de classification des émotions avec KNN	75
3.2	Performance de classification des émotions avec forêt aléatoire	76
3.3	Architecture du modèle ANN avec Batch Normalization	78
3.4	Performance du modèle ANN sur l'ensemble de test	79
3.5	Architecture du modèle CNN	81
3.6	Performance du modèle CNN sur la classification des émotions	82
3.7	Comparaison des performances des algorithmes de classification des émotions	85
3.8	Comparaison des valeurs ROC des algorithmes de classification des émotions	86

*The future belongs to those who believe
in the beauty of their dreams.*

— Eleanor Roosevelt

INTRODUCTION GÉNÉRALE

La reconnaissance des émotions par des systèmes automatiques rencontre plusieurs défis majeurs. Les caractéristiques émotionnelles varient largement non seulement entre individus, mais aussi pour un même individu selon les contextes et états internes. Cette variabilité, à la fois inter-orateurs et intra-orateurs, complique la tâche de classification des émotions. Par ailleurs, les signaux émotionnels obtenus à partir de modalités comme la voix et les expressions faciales peuvent être altérés par le bruit de fond et les conditions environnementales changeantes, ce qui impacte la précision et la robustesse des systèmes de reconnaissance émotionnelle. Enfin, pour une reconnaissance plus précise, il est nécessaire d'intégrer plusieurs modalités, ce qui exige des techniques sophistiquées pour le traitement et la fusion des données.

En améliorant la reconnaissance des émotions, ces systèmes ont le potentiel de transformer divers domaines tels que la santé mentale, l'éducation, et les interactions homme-machine. Une meilleure compréhension des émotions par les machines peut enrichir l'expérience utilisateur, rendre les interactions plus intuitives et personnalisées, et, en fin de compte, améliorer la qualité de vie.

Les objectifs de ce projet visent principalement à détecter les émotions à partir de la voix pour diverses catégories émotionnelles telles que la joie, la tristesse, la colère, la surprise et la neutralité. Pour atteindre cet objectif, le projet se concentre sur trois aspects clés : l'extraction de caractéristiques, la modélisation et l'apprentissage, ainsi que l'optimisation et le déploiement des modèles.

Tout d'abord, l'extraction de caractéristiques consiste à identifier et extraire des caractéristiques pertinentes à partir des signaux audio, telles que les coefficients cepstraux en fréquences mélodiques (MFCC), les caractéristiques prosodiques, et les spectrogrammes. Ensuite, la modélisation et l'apprentissage impliquent la construction et l'entraînement de modèles d'apprentissage automatique et de modèles de deep learning capables de classifier efficacement les émotions. Enfin, l'optimisation et le déploiement des modèles sont réalisés pour garantir leur utilisation en temps réel dans des applications pratiques, rendant la détection d'émotions à partir de la voix plus rapide et plus précise.

Ce rapport est structuré en trois chapitres principaux, chacun abordant des aspects essentiels du traitement et de la reconnaissance des émotions à partir de la parole. Le **chapitre 1** présente une introduction générale au traitement de la parole, en détaillant ses caractéristiques et types, tels que les sons voisés et non voisés, et en décrivant l'appareil phonatoire ainsi que les caractéristiques sonores comme l'énergie, l'amplitude, l'intensité et le pitch. Le **chapitre 2** se concentre sur le processus de reconnaissance automatique des émotions, en détaillant les étapes clés telles que le prétraitement des données, l'extraction des caractéristiques pertinentes et la classification des émotions à l'aide d'algorithmes d'apprentissage automatique et profond. Le **chapitre 3** explique la méthodologie de travail adoptée et propose une analyse approfondie des performances des modèles, discutant des résultats obtenus pour chaque algorithme utilisé.

Enfin, le rapport se conclut par une conclusion générale qui résume les résultats finaux, offre une perspective sur l'avenir de la reconnaissance des émotions multimodale et propose des pistes pour de futures recherches dans ce domaine.

CHAPITRE 1

INTRODUCTION AU SIGNAL DE PAROLE

Introduction

Dans ce chapitre, nous explorerons le cadre général de l'étude de la voix humaine, un domaine fascinant qui englobe l'anatomie et la physiologie de l'appareil phonatoire, les caractéristiques acoustiques de la voix, ainsi que les troubles qui peuvent affecter la production vocale.

Nous commencerons par examiner en détail l'appareil phonatoire humain, en nous intéressant aux organes de la phonation, aux cordes vocales, aux cavités de l'appareil phonatoire et aux points d'articulation. Ensuite, nous aborderons les caractéristiques fondamentales de la voix, telles que la fréquence fondamentale, l'énergie et le spectre.

Nous discuterons également des différents types de troubles de la voix, notamment les troubles organiques, fonctionnels et psychogènes, ainsi que des méthodes d'évaluation des voix acoustiques, qu'elles soient objectives ou subjectives. Ce chapitre jettera les bases nécessaires pour une compréhension approfondie de la voix humaine et de ses mécanismes.

1.1 Appareil phonatoire humain

L'appareil phonatoire, aussi appelé appareil vocalique, est l'ensemble des organes de la parole et des muscles qui les actionnent. Il permet la production des phones, ou sons propres à la langue parlée.

Voici les principaux composants de l'appareil phonatoire :

- **Organes de la phonation** : Ils comprennent les poumons, qui agissent comme une soufflerie pour fournir l'air nécessaire à la production de la plupart des sons. Les sons sont produits pendant l'expiration.
- **Cordes vocales** : Situées dans le larynx, elles vibrent pour produire des sons. Le larynx contient également le pharynx et la glotte.
- **Cavités de l'appareil phonatoire** : L'appareil phonatoire comprend plusieurs cavités que l'air pulmonaire traverse, notamment la glotte, les cavités supraglottiques, la cavité pharyngale, la cavité buccale, la cavité labiale et la cavité nasale.
- **Points d'articulation** : Pour permettre la phonation, les différents organes mobiles entrent en contact ou se resserrent contre d'autres organes, pour former un point d'articulation [1].

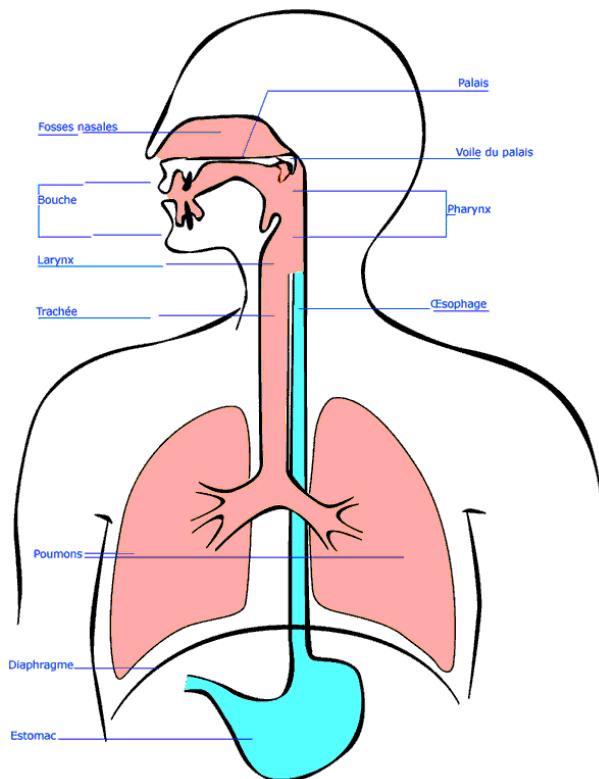


FIGURE 1.1 – Appareil phonatoire humain [1]

1.1.1 Organes de la phonation

Les organes de la phonation sont essentiels pour la production de la parole. Ils comprennent notamment :

- **Les poumons** : Situés dans le thorax, au-dessus du diaphragme et derrière les muscles costaux, ils agissent comme une soufflerie pour fournir l'air nécessaire à la production de la plupart des sons. Les sons sont produits pendant l'expiration.
- **La trachée** : Elle descend dans le thorax et se divise en deux bronches prolongées par de nombreuses bronchioles qui pénètrent dans les deux poumons jusqu'aux alvéoles.
- **Le larynx** : Situé dans le cou, à hauteur de la pomme d'Adam, il abrite les cordes vocales.
- **Le pharynx et la glotte** : Ils se trouvent dans la gorge, entre les deux plis vocaux.
- **La cavité orale et la cavité nasale** : Elles se trouvent dans la tête, tout comme le nez (et ses fosses nasales), la bouche (qui contient la langue, et sa pointe, et son dos), le palais (et sa voûte, et son voile), la luette, les dents (enracinées dans les alvéoles dentaires), les lèvres (inférieure et supérieure), et les joues [2].

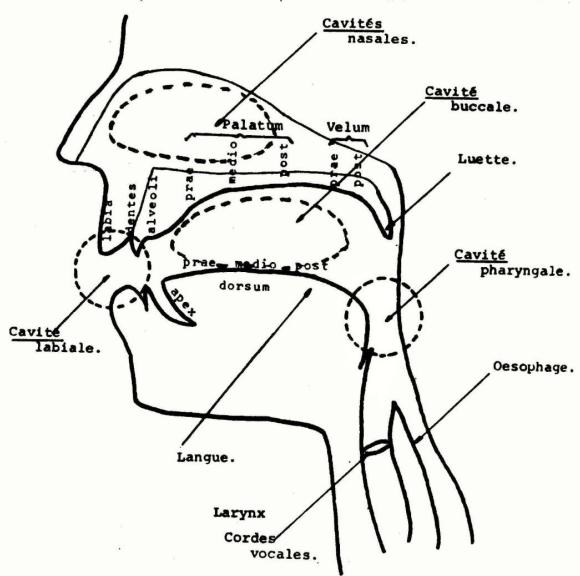


FIGURE 1.2 – Organes de la phonation [2]

1.1.2 Cordes vocales

Les cordes vocales, plus justement appelées les plis vocaux, sont un des organes musculaires de la phonation constitué de replis fermes et souples à la fois des membranes muqueuses du larynx dont la vibration produit les sons vocaux — cri, langage, chant.

Elles sont composées de plusieurs couches de structures différentes : le muscle vocal puis la lamina propria qui regroupe trois couches : profonde, moyenne et superficielle. Les couches profondes et moyennes sont formées par le ligament vocal. La couche superficielle est un espace de glissement appelé espace de Reinke. Ces différentes épaisseurs sont recouvertes de muqueuse plus ou moins visqueuse.

Les cordes vocales sont tendues, dans le larynx, de l'angle rentrant du cartilage thyroïde (celui qui, ayant un angle de 90° chez l'homme donne le relief de la pomme d'Adam) à l'apophyse vocale des cartilages arytenoïdes.

Elles sont écartées lorsque le sujet respire, rapprochées quand il déglutit. Les cordes vocales sont innervées par les branches du nerf vague que sont le nerf laryngé supérieur (se divisant lui-même en nerf laryngé interne et nerf laryngé externe) et le nerf laryngé récurrent (ou nerf laryngé inférieur). Le rameau interne du nerf laryngé supérieur est essentiellement sensitif tandis que le rameau externe est moteur et innervé le muscle cricothyroïdien [2].

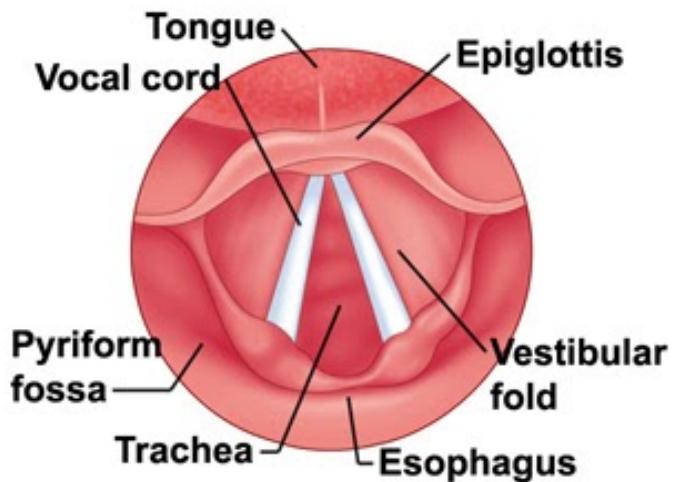


FIGURE 1.3 – Cordes vocales [2]

1.1.3 Cavités de l'appareil phonatoire

L'appareil phonatoire comporte plusieurs cavités que l'air pulmonaire traverse. Voici les principales cavités de l'appareil phonatoire :

- **La glotte** : C'est l'espace de forme triangulaire circonscrit par les plis vocaux, par lequel l'air pulmonaire peut s'échapper. Pour la phonation, la glotte doit se fermer le long de sa ligne médiane.
- **Les cavités supraglottiques** : Aussi appelées cavités suprapharyngales, ou résonateurs supraglottiques, elles jouent un rôle principal dans l'articulation de la parole. Elles servent en effet de résonateurs au ton laryngien.
- **La cavité pharyngale** : Elle est limitée vers le bas par le pharynx, et vers le haut par la racine de la langue et le voile du palais.
- **La cavité buccale** : Aussi appelée cavité antérieure, elle est limitée à l'avant par des incisives et à l'arrière par le point d'articulation, c'est-à-dire le lieu de resserrement le plus étroit du chenal buccal pendant l'articulation. Elle peut changer de forme grâce aux mouvements de la langue et à la position des lèvres.
- **La cavité labiale** : Elle est comprise entre les incisives et les lèvres plus ou moins contractées. Elle n'intervient dans la phonation que si les lèvres sont projetées vers l'avant, qui s'accompagne généralement d'un arrondissement de ces dernières. Les sons produits par ce processus seront caractérisés comme labialisés (ou labials).
- **La cavité nasale** : Elle intervient dans la phonation quand l'abaissement de l'extrémité du voile du palais, ou la luette, permet l'écoulement libre d'une partie de l'air issu du larynx par les fosses nasales.

Ces cavités sont essentielles pour la production de la parole car elles modifient le son produit par les cordes vocales, ce qui permet de créer les différents sons de la langue parlée [3].

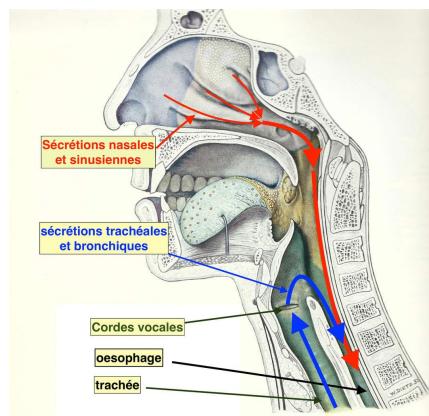


FIGURE 1.4 – Cavités de l'appareil phonatoire [3]

1.1.4 Points d'articulation

En phonétique articulatoire, un point d'articulation d'une consonne désigne l'endroit où s'effectue l'obstruction (soit partielle, soit totale puis relâchée) au passage de l'air injecté ou éjecté par la voie buccale¹. Voici les principaux points d'articulation :

- **Labial** : L'obstruction se fait avec les lèvres (exo-labial et endo-labial).
- **Dental** : L'obstruction se fait avec les dents.
- **Alvéolaire** : L'obstruction se fait avec les alvéoles.
- **Post-alvéolaire** : L'obstruction se fait juste derrière les alvéoles.
- **Pré-palatal** : L'obstruction se fait à l'avant du palais.
- **Palatal** : L'obstruction se fait avec le palais.
- **Vélaire** : L'obstruction se fait avec le voile du palais.
- **Uvulaire** : L'obstruction se fait avec l'uvule.
- **Pharyngal** : L'obstruction se fait avec le pharynx.
- **Glottal** : L'obstruction se fait avec la glotte.

On distingue deux points distincts pour cette obstruction :

- Au niveau supérieur, le lieu d'articulation contre lequel glisse l'air injecté ou éjecté, pendant (obstruction partielle) ou après (obstruction totale) l'articulation.
- Au niveau inférieur, l'organe articulatoire. Un seul point peut jouer les deux rôles obstructeurs simultanément : la glotte qui ne peut que se fermer ou s'ouvrir elle-même, sans intervention d'un second organe dans l'articulation [4].

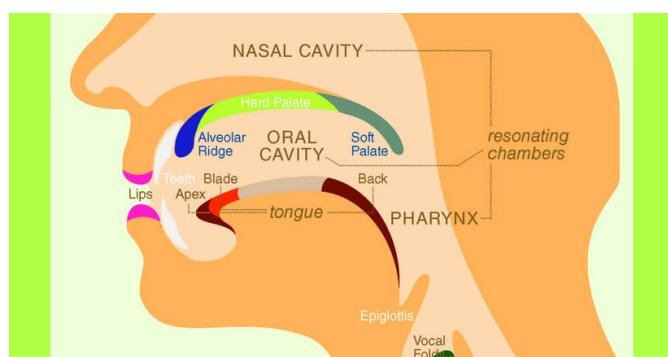


FIGURE 1.5 – Points d'articulation [4]

1.2 Caractéristiques de la voix

La voix humaine est une manifestation complexe de communication, résultant de la vibration des cordes vocales, modulée par les mouvements de divers organes tels que la langue, les lèvres, et le palais. Ses caractéristiques comprennent plusieurs aspects distincts qui contribuent à sa richesse et sa diversité.

- **La tonalité de la voix**, souvent associée à la hauteur, est déterminée par la fréquence des vibrations des cordes vocales. Une voix peut être grave ou aiguë en fonction de cette fréquence, et elle peut varier considérablement d'une personne à l'autre.
- **L'intensité de la voix**, ou son volume, est dictée par l'amplitude des vibrations des cordes vocales et la force de l'air expulsé des poumons. Cette caractéristique permet à une voix d'être perçue comme forte ou douce, et elle est essentielle pour transmettre des émotions et attirer l'attention.
- **Le timbre de la voix** est la qualité distinctive qui permet de différencier les voix individuelles, même si elles produisent la même note à la même intensité. Il est influencé par divers facteurs, notamment la forme et la taille des cavités buccales et nasales, ainsi que par les caractéristiques physiques uniques de chaque personne.
- **Le débit vocal**, ou la vitesse à laquelle les mots sont prononcés, est une caractéristique importante de la voix qui peut varier selon le contexte et l'émotion. Un débit rapide peut indiquer de l'enthousiasme ou du stress, tandis qu'un débit lent peut suggérer du calme ou de la réflexion.
- **L'expressivité de la voix** englobe une gamme d'éléments tels que l'intonation, le rythme, et l'articulation, qui permettent de transmettre des nuances de sens et d'émotion. Une voix expressive est capable de captiver l'auditeur et de rendre le discours plus vivant et engageant.

Dans l'ensemble, la voix humaine est un instrument remarquablement adaptable et expressif, capable de communiquer une vaste gamme d'émotions, d'idées et de significations. Ses caractéristiques variées offrent un potentiel infini pour la communication et l'expression créative [5].

1.2.1 Fréquence fondamentale

La fréquence fondamentale est une caractéristique essentielle de la voix humaine. Elle représente la fréquence de base des vibrations des cordes vocales lors de la production de sons vocaux. Cette fréquence est déterminée par la tension et la longueur des cordes vocales, ainsi que par la quantité d'air expulsée des poumons. La fréquence fondamentale est étroitement liée à la perception de la hauteur tonale de la voix. Les sons avec une fréquence fondamentale plus élevée sont perçus comme des sons aigus, tandis que ceux avec une fréquence fondamentale plus basse sont perçus comme des sons graves. Chez les individus adultes, la fréquence fondamentale moyenne se situe généralement entre 85 et 180 Hertz chez les hommes et entre 165 et 255 Hertz chez les femmes, bien que ces chiffres puissent varier considérablement d'une personne à l'autre [5].

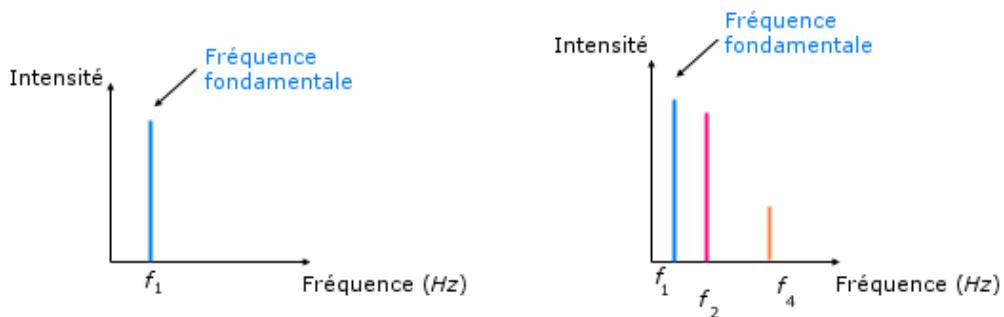


FIGURE 1.6 – Fréquence fondamentale [5]

1.2.2 Énergie

Dans le contexte de la voix humaine, l'énergie se réfère à l'intensité acoustique des sons produits lors de la parole. Elle représente la quantité de puissance sonore émise par l'émission vocale et est étroitement liée au volume perçu de la voix.

L'énergie vocale est produite par la force de l'air expulsée des poumons à travers les voies respiratoires et les cordes vocales. Plus l'air est expulsé avec force, plus l'énergie acoustique générée est grande, ce qui se traduit par une voix plus forte ou plus intense.

L'énergie vocale est mesurée en décibels (dB) et peut varier considérablement d'une personne à l'autre en fonction de facteurs tels que la force des muscles respiratoires, la taille des cavités buccales et nasales, et le degré de tension des cordes vocales.

La modulation de l'énergie vocale est essentielle pour la communication expressive, car elle permet à l'orateur de transmettre des nuances d'émotion, d'accentuer des mots ou des phrases importants, et d'attirer l'attention de l'auditeur. Un contrôle efficace de l'énergie vocale est donc crucial pour une communication vocale claire, expressive et engageante [5].

1.2.3 Spectre

Le spectre vocal fait référence à la distribution des composantes fréquentielles dans le son produit par la voix humaine. Il est représenté graphiquement sous forme de spectre, où l'axe horizontal représente les différentes fréquences et l'axe vertical représente l'intensité ou l'énergie des différentes fréquences.

Le spectre vocal d'une personne est unique et peut varier en fonction de facteurs tels que la physiologie individuelle, le sexe, l'âge et même l'émotion exprimée. Il est généralement composé de multiples harmoniques, qui sont des multiples entiers de la fréquence fondamentale. Ces harmoniques contribuent à donner à chaque voix sa qualité distinctive ou son timbre.

L'analyse du spectre vocal permet de caractériser différentes propriétés de la voix, telles que la hauteur tonale, l'énergie vocale à différentes fréquences et les caractéristiques de la résonance. Par exemple, un spectre vocal avec une forte énergie dans les basses fréquences peut indiquer une voix grave, tandis qu'un spectre avec une énergie plus élevée dans les hautes fréquences peut indiquer une voix plus aiguë [5].

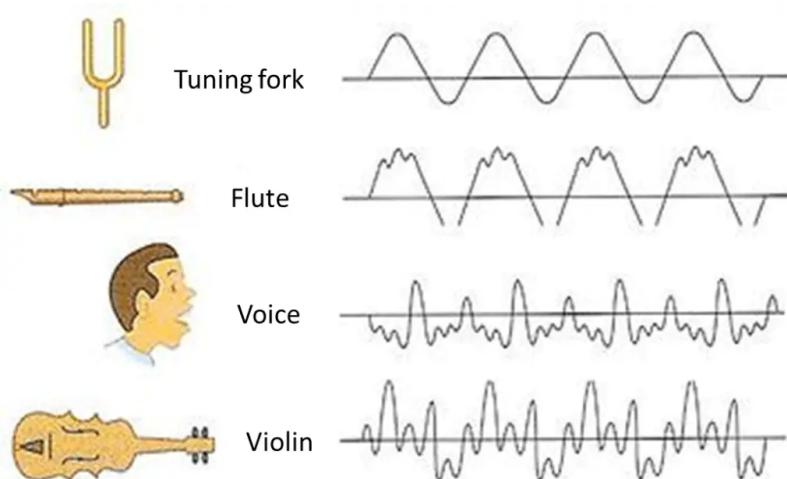


FIGURE 1.7 – Spectres [5]

1.3 Troubles de la voix

Les troubles de la voix, également connus sous le nom de troubles vocaux ou dysphonies, sont des conditions médicales qui affectent la production normale de la voix. Ces troubles peuvent être causés par divers facteurs, notamment des problèmes physiologiques, des lésions des cordes vocales, des infections, des troubles neurologiques, des traumatismes ou des comportements vocaux nocifs [6].

1.3.1 Troubles organiques

- Troubles structurels : Ces troubles découlent d'origines d'une anomalie physique. Ils impliquent une perturbation dans les mécanismes physiques de la production de la voix, impliquant souvent un tissu ou des fluides des cordes vocales.
- Troubles neurogènes : Ces troubles sont causés par un problème du système nerveux lors de son interaction avec l'appareil phonatoire.

1.3.2 Troubles fonctionnels

Signifient que la structure physique est normale, mais que le mécanisme vocal est utilisé de manière inappropriée ou inefficace, ceci est causé par un mauvais fonctionnement musculaire.

1.3.3 Troubles psychogènes

Une mauvaise qualité vocale peut devenir une manifestation symbolique ou extérieure de conflits psychologiques non résolus. Dans de tels cas, la voix peut être affectée sans cause traumatique ou infectieuse, sans altérations des cordes vocales et sans dommage apparent au larynx. La toux reste sonore, et il est possible que des difficultés psychologiques se soient converties en un trouble fonctionnel.

Souvent, différents types de troubles interagissent dans la même pathologie, ce qui rend la classification du type de trouble quelque peu difficile. Dans ce qui suit, nous allons explorer les différentes méthodes utilisées pour détecter le type de pathologie vocale.

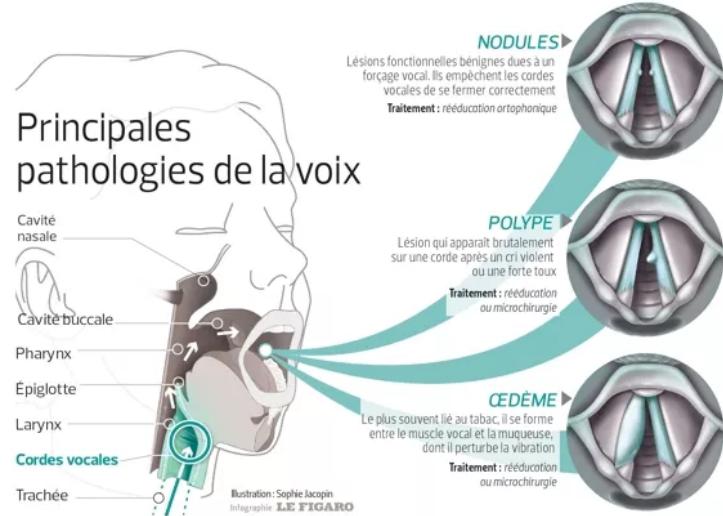


FIGURE 1.8 – Troubles de la voix [7]

1.4 Évaluation des voix acoustiques

L'évaluation des voix acoustiques est une composante cruciale de l'analyse vocale, visant à évaluer divers aspects de la qualité vocale, de la clarté et de la performance vocale. Cette évaluation peut être réalisée à l'aide de méthodes objectives et subjectives pour obtenir une compréhension complète des caractéristiques vocales d'un individu [8].

1.4.1 Méthodes objectives

Les méthodes objectives d'évaluation des voix acoustiques impliquent l'utilisation d'outils et de logiciels spécialisés pour mesurer et analyser des paramètres acoustiques spécifiques de la voix. Ces paramètres peuvent inclure la fréquence fondamentale, l'intensité vocale, la durée des phonèmes, la variabilité de la fréquence, et d'autres mesures objectives de la qualité et de la performance vocale. Des techniques telles que l'analyse spectrale, l'analyse de la périodicité, et l'analyse de la modulation peuvent être utilisées pour obtenir des mesures précises et quantifiables des caractéristiques acoustiques de la voix. Les avantages des méthodes objectives incluent leur objectivité et leur reproductibilité, mais elles peuvent parfois ne pas capturer toutes les nuances subjectives de la qualité vocale [9].

1.4.2 Méthodes subjectives

Les méthodes subjectives d'évaluation des voix acoustiques impliquent l'évaluation par des auditeurs humains formés ou non formés, qui évaluent la qualité vocale en fonction de leur perception personnelle. Ces évaluations peuvent être réalisées à l'aide d'échelles d'évaluation standardisées, de questionnaires ou d'évaluations qualitatives ou quantitatives basées sur des critères spécifiques. Les auditeurs peuvent évaluer des aspects tels que la clarté, la richesse tonale, l'articulation, l'expressivité et d'autres caractéristiques subjectives de la voix. Les méthodes subjectives offrent une perspective humaine importante sur la qualité vocale, mais elles peuvent être sujettes à des biais individuels et à des variations inter-évaluateurs [10].

Conclusion

En conclusion, ce chapitre a fourni une vue d'ensemble complète de l'étude de la voix humaine.

Nous avons exploré en détail l'appareil phonatoire, les caractéristiques acoustiques de la voix et les différents types de troubles vocaux.

De plus, nous avons examiné les méthodes d'évaluation des voix acoustiques, qu'elles soient objectives ou subjectives.

Cette exploration nous a permis de mieux comprendre l'anatomie et la physiologie de la voix, ainsi que les défis rencontrés dans l'évaluation et la gestion des troubles vocaux.

En résumé, ce chapitre constitue une base solide pour approfondir nos connaissances sur ce sujet crucial dans les domaines de la santé vocale et de la communication humaine.

CHAPITRE 2

DÉTECTION AUTOMATIQUE DES ÉMOTIONS AVEC LA VOIX

Introduction

La reconnaissance des émotions à partir de la voix est un domaine en pleine croissance qui vise à rendre les systèmes informatiques capables de comprendre et de réagir aux émotions humaines. Cette technologie trouve des applications dans les assistants virtuels, les services clients automatisés et la surveillance de la santé mentale.

Dans cette section, nous explorerons les étapes essentielles pour développer des systèmes de reconnaissance des émotions vocales. Nous commencerons par les bases de données utilisées pour l'entraînement des modèles. Ensuite, nous aborderons le pré-traitement des données, l'extraction des caractéristiques acoustiques telles que les MFCC, RMSE et ZRC, ainsi que les techniques de visualisation des émotions, notamment le signal d'onde, le spectrogramme et le chroma.

Nous examinerons également les techniques de classification couramment utilisées, telles que les forêts aléatoires et les réseaux de neurones convolutionnels. Enfin, nous discuterons des méthodes d'évaluation des modèles et présenterons des études de cas récentes qui démontrent l'efficacité de ces techniques.

Cet analyse fournit une vue d'ensemble des méthodes et des techniques nécessaires pour la reconnaissance des émotions vocales, mettant en lumière les avancées et les applications potentielles dans ce domaine.

2.1 Base de données des voix émotionnels

La première étape essentielle dans le processus de reconnaissance des émotions à partir de la voix est la construction d'une base de données contenant des enregistrements vocaux classifiés selon les différentes émotions exprimées, telles que la joie, la tristesse, la colère, la surprise, la peur, la neutralité, entre autres. Cette base de données doit être suffisamment riche et diversifiée pour inclure des variations dans la tonalité, l'intensité, le rythme, et les caractéristiques prosodiques qui sont propres à chaque émotion, et qui peuvent également varier selon le locuteur, le genre, l'âge, et la langue. La création d'une telle base de données nécessite la collecte de nombreux échantillons vocaux provenant de différents locuteurs dans diverses situations émotionnelles, ce qui peut être un processus long et coûteux.

2.2 Pré-traitement de données

Le pré-traitement des données est une étape cruciale dans l'analyse des voix émotionnelles. Il permet de nettoyer et de transformer les données brutes en un format plus approprié pour l'extraction de caractéristiques et la classification. Voici les principales étapes de pré-traitement des données vocales :

1- Normalisation et échantillonnage :

La première étape consiste souvent à normaliser le niveau sonore des enregistrements afin d'assurer une amplitude uniforme à travers tous les fichiers audio. Cela peut être réalisé en ajustant les niveaux de volume pour qu'ils soient comparables. L'échantillonnage consiste à convertir les enregistrements audio à une fréquence d'échantillonnage standard, généralement 16 kHz ou 44.1 kHz, afin de faciliter l'analyse et la comparaison des données.

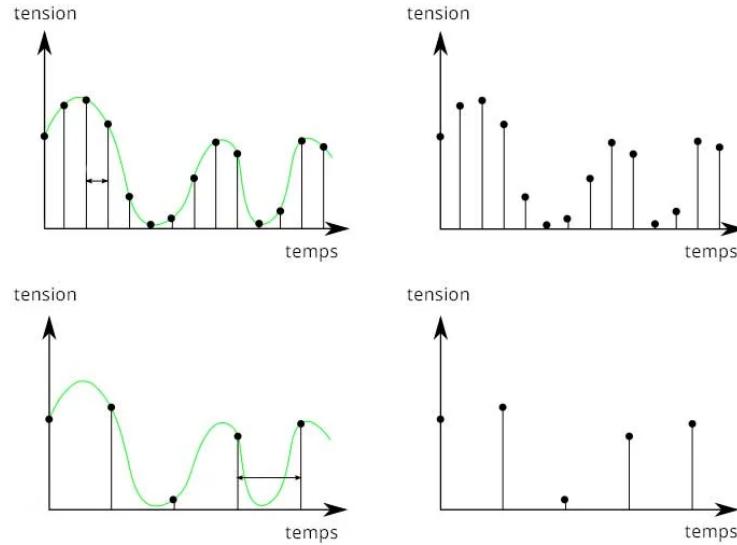


FIGURE 2.1 – Normalisation et échantillonnage [11]

2- Suppression du bruit :

Les enregistrements vocaux peuvent contenir du bruit de fond indésirable qui peut affecter l'analyse. La suppression du bruit est réalisée en utilisant des techniques telles que les filtres passe-bas, passe-haut ou passe-bande, et les algorithmes de réduction de bruit spectral. Ces méthodes aident à isoler la voix des bruits ambients, améliorant ainsi la qualité des données pour l'extraction des caractéristiques.

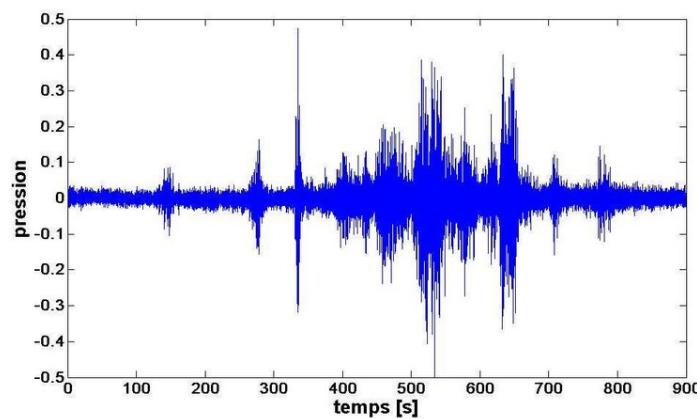


FIGURE 2.2 – Signal avec le bruit [12]

3- Segmentation :

La segmentation implique la division des enregistrements vocaux en segments plus petits, souvent à l'échelle de la phrase ou du mot. Cette étape est essentielle pour analyser les variations émotionnelles au sein de l'enregistrement.

La segmentation peut être réalisée automatiquement en détectant les pauses et les silences dans le discours.

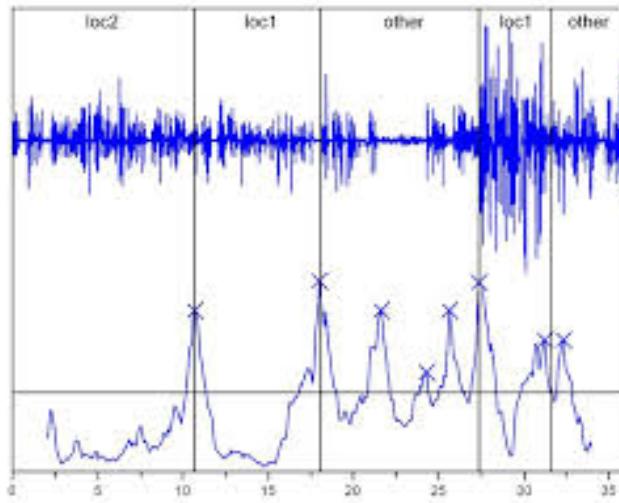


FIGURE 2.3 – Segmentation d'un audio [13]

4- Augmentation des données :

Pour compenser le manque de données ou pour équilibrer les classes, des techniques d'augmentation des données peuvent être utilisées. Cela inclut la variation de la vitesse de lecture, l'ajout de bruit, ou encore la modification des tonalités.

Ces techniques permettent d'augmenter la diversité des données d'entraînement et d'améliorer la robustesse des modèles.

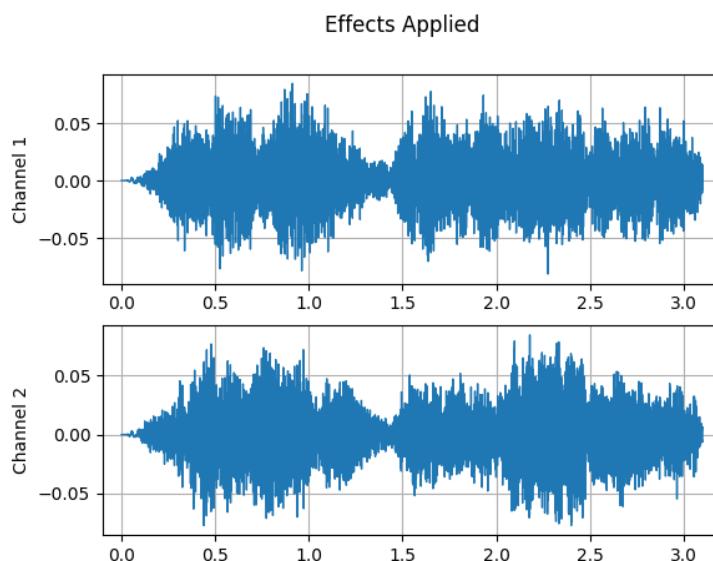


FIGURE 2.4 – Augmentation des données audio [14]

5- Normalisation des caractéristiques :

Enfin, les caractéristiques extraites sont souvent normalisées pour avoir une moyenne nulle et une variance unitaire. Cela aide à améliorer la performance des algorithmes de machine learning en garantissant que toutes les caractéristiques sont sur une échelle comparable.

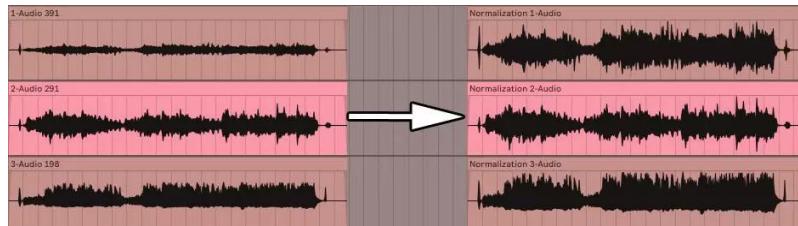


FIGURE 2.5 – Normalisation des caractéristiques audios [15]

2.3 Techniques d'analyse des caractéristiques des signaux audios

2.3.1 Signal d'onde

Le signal d'onde est la représentation temporelle des variations de pression acoustique dans le temps. Cette représentation est essentielle pour comprendre les caractéristiques fondamentales du signal audio. Voici les aspects clés concernant le signal d'onde :

- **Acquisition du signal** : Le signal d'onde est obtenu directement à partir des capteurs acoustiques lors de l'enregistrement audio. Il représente les variations de pression sonores mesurées au fil du temps.
- **Visualisation** : La représentation graphique du signal d'onde montre les variations de l'amplitude du signal en fonction du temps. Cette visualisation est utile pour l'analyse qualitative des propriétés du signal, comme les pics et les creux qui correspondent aux variations de volume et aux événements acoustiques.
- **Analyse temporelle** : Le signal d'onde permet d'étudier les propriétés temporelles du signal audio, comme la durée, les transitoires, et les structures rythmiques.

Pour une compréhension plus approfondie des signaux d'onde et de leur traitement [16].

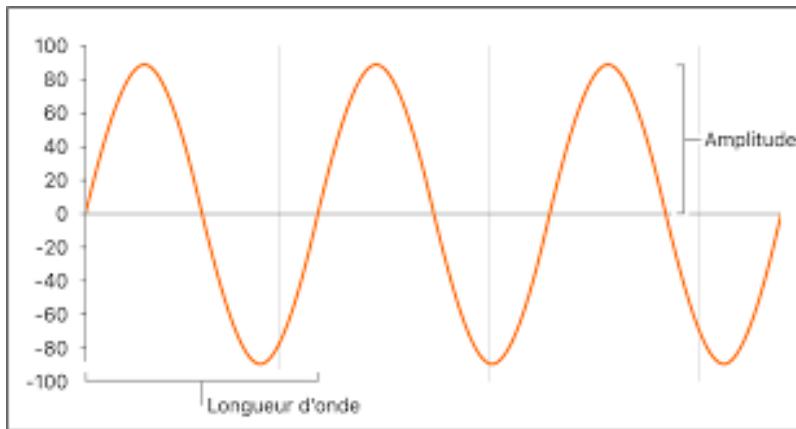


FIGURE 2.6 – Signal d’onde [16]

2.3.2 Spectrogrammes

Les spectrogrammes offrent une représentation visuelle des variations de fréquence dans le temps. Voici les étapes et les détails importants concernant les spectrogrammes :

- **Segmentation du signal** : Diviser le signal vocal en segments courts (fenêtres) similaires à ceux utilisés pour les MFCC, généralement de 20 à 40 ms avec chevauchement.
- **Transformée de Fourier** : Appliquer la transformée de Fourier sur chaque segment pour obtenir un spectre de puissance pour chaque fenêtre. Chaque spectre représente la distribution de l’énergie à travers différentes fréquences pour ce segment de temps.
- **Empilage des spectres** : Empiler les spectres successifs pour former une image où l’axe des x représente le temps, l’axe des y représente la fréquence, et les valeurs de couleur représentent l’amplitude de la fréquence à ce moment précis.
- **Interprétation des spectrogrammes** : Les motifs dans le spectrogramme peuvent révéler des caractéristiques spécifiques des émotions. Par exemple, des variations rapides et des pics peuvent indiquer des émotions comme la surprise ou la colère, tandis que des motifs plus réguliers peuvent être associés à des émotions neutres ou tristes.

Les spectrogrammes permettent ainsi de visualiser et d'analyser les variations temporelles et fréquentielles des signaux vocaux, fournissant des indices précieux sur les émotions.

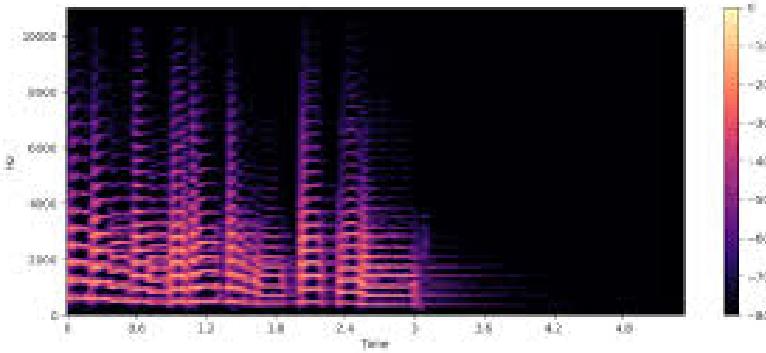


FIGURE 2.7 – Spectrogramme [16]

2.3.3 Chroma

Les caractéristiques chroma représentent l'intensité des 12 classes de hauteurs musicales (ou *pitch classes*) dans le signal audio. Voici comment les caractéristiques chroma sont calculées et utilisées :

- **Transformée de Fourier** : Convertir le signal audio en domaine fréquentiel en utilisant la transformée de Fourier, similaire aux étapes pour les spectrogrammes.
- **Regroupement par classes de pitch** : Regrouper les fréquences en 12 classes de pitch, correspondant aux 12 notes de la gamme musicale occidentale, indépendamment de l'octave.
- **Normalisation** : Normaliser les intensités des classes de pitch pour éliminer les variations dues à l'amplitude globale du signal audio.
- **Analyse des variations chroma** : Les caractéristiques chroma capturent les variations mélodiques et harmoniques. Par exemple, des émotions joyeuses peuvent présenter des variations chromatiques plus prononcées et dynamiques, tandis que des émotions tristes peuvent montrer des motifs plus monotones.

Les caractéristiques chroma sont particulièrement utiles pour l'analyse des aspects mélodiques et tonals de la voix, ce qui est crucial pour distinguer certaines émotions, comme illustré dans la **figure 2.8**.

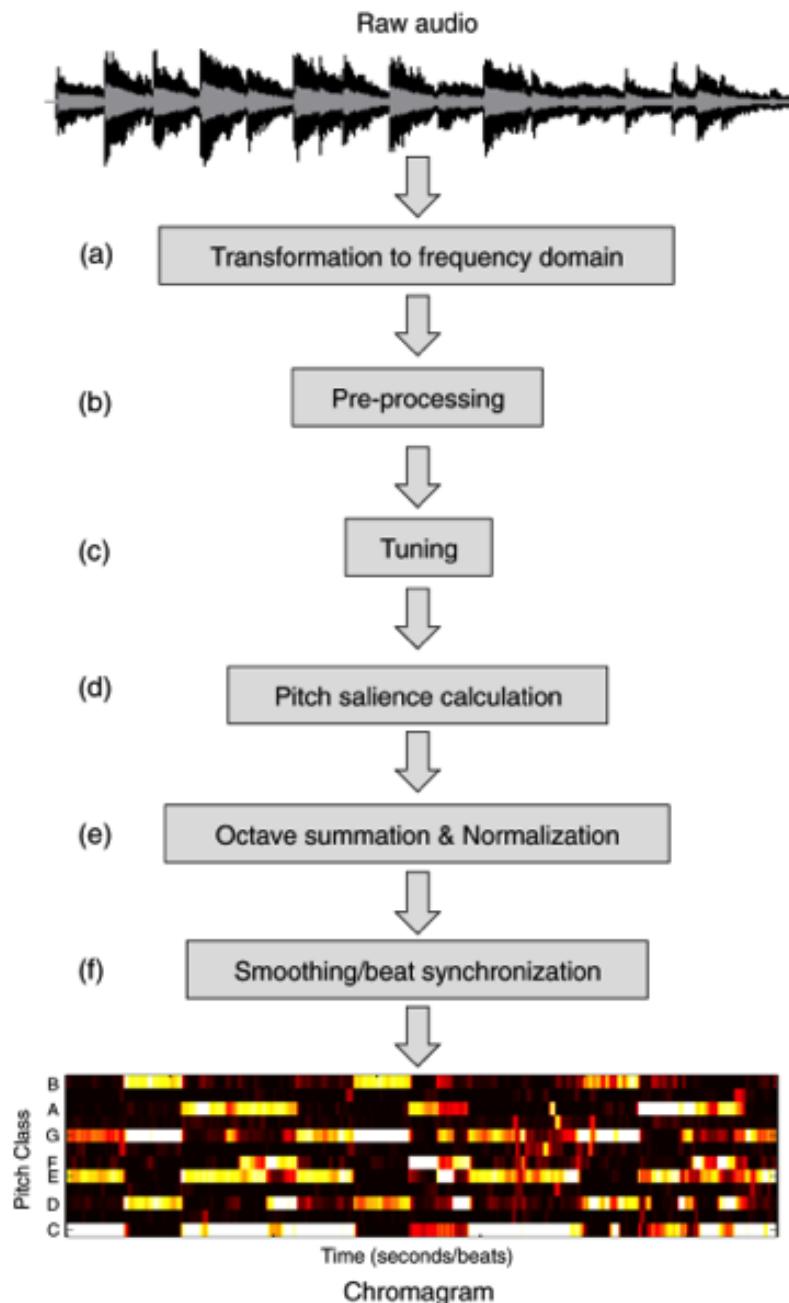


FIGURE 2.8 – Architecture de chroma [17]

2.4 Extraction des caractéristiques

L'extraction des caractéristiques est une étape cruciale dans le traitement des signaux vocaux pour la reconnaissance des émotions. Cette étape consiste à transformer les données audio brutes en un ensemble de variables ou de caractéristiques qui peuvent être utilisées par des algorithmes de machine learning pour identifier les émotions. Voici les principales techniques et caractéristiques utilisées pour l'extraction des caractéristiques dans les données vocales :

2.4.1 MFCC (Mel-Frequency Cepstral Coefficients)

Les coefficients cepstraux en fréquences mél (MFCC) sont parmi les caractéristiques les plus couramment utilisées pour l'analyse de la parole. Ils capturent les propriétés spectrales de la voix en imitant la perception humaine des fréquences sonores. Voici les étapes détaillées pour calculer les MFCC :

- **Pré-emphase** : Appliquer un filtre de pré-emphase pour amplifier les hautes fréquences et réduire les basses fréquences, améliorant ainsi la clarté des caractéristiques spectrales.

Formule :

$$y[n] = x[n] - \alpha x[n - 1]$$

où α est typiquement 0.95.

- **Fenêtrage** : Diviser le signal vocal en segments courts (ou fenêtres) de 20 à 40 ms avec un chevauchement (typiquement 50%) pour capter les propriétés stationnaires locales. Utiliser une fenêtre de Hamming pour réduire les effets de bord.

Formule de la fenêtre de Hamming :

$$w[n] = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N - 1}\right)$$

- **Transformée de Fourier (FFT)** : Appliquer la transformée de Fourier sur chaque segment pour obtenir le spectre de puissance, représentant la distribution de l'énergie sur différentes fréquences.

Formule :

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j \frac{2\pi k}{N} n}$$

- **Échelle en fréquences mél** : Appliquer une banque de filtres triangulaires sur une échelle en fréquences mél, qui est une échelle perceptuelle non linéaire alignée sur la perception humaine des sons.

Conversion de fréquence (Hz) à mél :

$$m = 2595 \log_{10} \left(1 + \frac{f}{700} \right)$$

- **Logarithme de l'énergie spectrale** : Calculer le logarithme de l'énergie dans chaque bande de fréquence mél pour imiter la perception humaine de l'intensité sonore.

Formule :

$$E_m = \log \left(\sum_{k=1}^N |X[k]|^2 \cdot H_m[k] \right)$$

où H_m est le filtre en mél.

- **Transformée en cosinus discrète (DCT)** : Appliquer la DCT pour obtenir les coefficients cepstraux finaux, représentant les caractéristiques spectrales compactes et efficaces.

Formule :

$$C[n] = \sum_{m=0}^{M-1} E_m \cos \left(\frac{\pi n(2m+1)}{2M} \right)$$

Les MFCC fournissent une représentation compacte et efficace des caractéristiques spectrales de la voix, essentielle pour la reconnaissance des émotions.

Comme illustré en **figure 2.9**, l'architecture de MFCC est importante pour l'analyse des signaux vocaux.

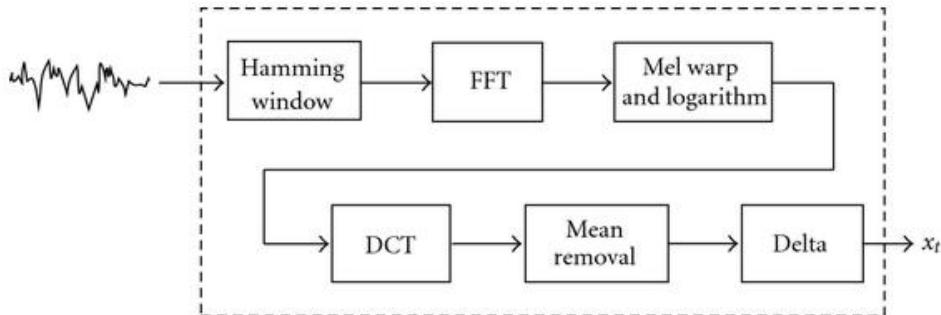


FIGURE 2.9 – Architecture de MFCC [18]

2.4.2 Root Mean Square Energy (RMSE)

L'énergie quadratique moyenne (RMSE) est une mesure qui donne une idée de la puissance moyenne du signal. Elle est souvent utilisée pour détecter la présence de parole ou d'autres caractéristiques sonores dans un segment audio [19].

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N x_i^2} \quad (2.1)$$

Où N est le nombre total d'échantillons dans le segment et x_i est l'amplitude du signal à l'indice i .

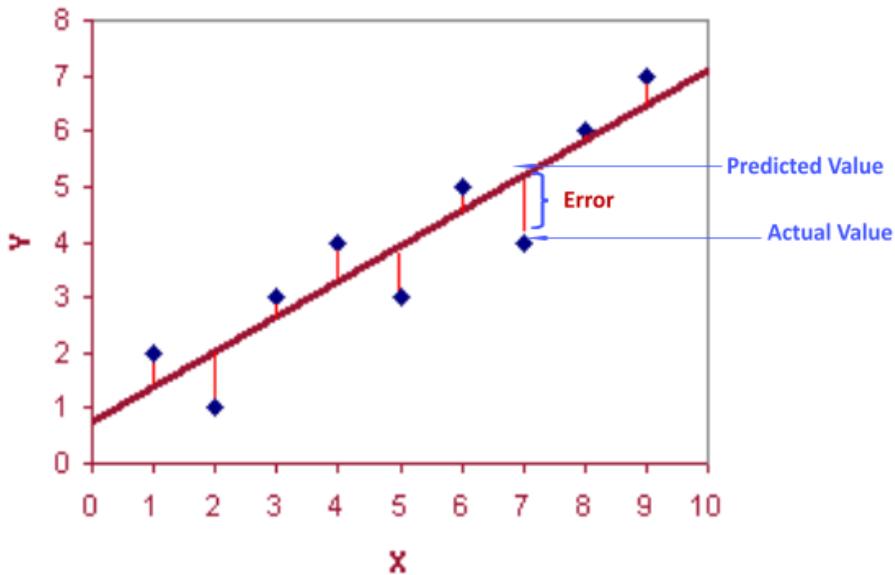


FIGURE 2.10 – Illustration de RMSE [20]

2.4.3 Zero Crossing Rate (ZCR)

Le taux de passage par zéro (ZCR) est le taux auquel le signal change de signe. Cette caractéristique est souvent utilisée pour analyser la nature percussive d'un son ou pour la segmentation des phonèmes dans la parole [21].

$$\text{ZCR} = \frac{1}{T-1} \sum_{t=1}^{T-1} |\text{sgn}(x_t) - \text{sgn}(x_{t-1})| \quad (2.2)$$

Où sgn est la fonction signe et T est le nombre total d'échantillons.

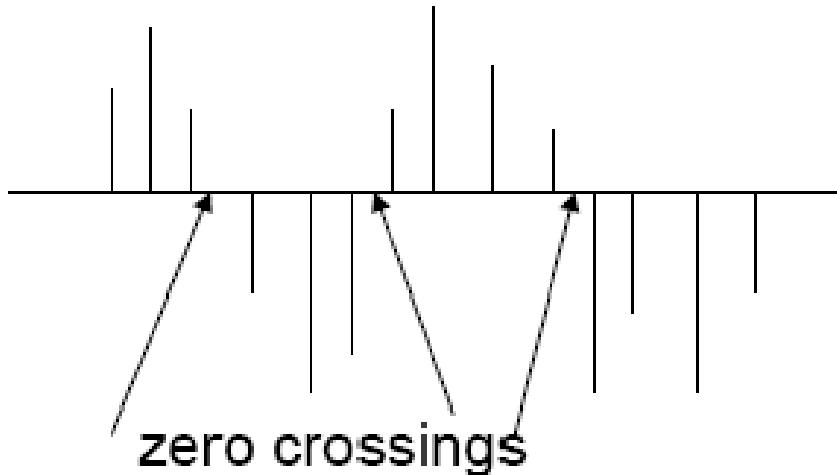


FIGURE 2.11 – Illustration de ZCR [22]

2.5 Classification

L'avant dernière étape est l'étape de la classification, cette dernière est le fait d'assigner une ou plusieurs classes à un élément, dans notre cas cet élément représente la voix. La classification se fait par des techniques d'apprentissage automatique. Dans cette partie, nous présentons l'apprentissage automatique.

2.5.1 Apprentissage automatique

L'apprentissage automatique est un domaine de l'intelligence artificielle qui permet aux systèmes informatiques d'apprendre et de s'améliorer à partir de données, sans être explicitement programmés. Il existe plusieurs types d'apprentissage automatique, mais les deux principaux sont l'apprentissage supervisé et l'apprentissage non supervisé. L'apprentissage supervisé consiste à entraîner un modèle sur des données étiquetées, afin qu'il puisse prédire avec précision la bonne sortie lorsqu'il reçoit de nouvelles données en entrée [23]. Les algorithmes d'apprentissage automatique sont généralement classés selon la tâche qu'ils réalisent, nous citons :

- Apprentissage supervisé
- Apprentissage non supervisé
- Apprentissage semi-supervisé
- Apprentissage par renforcement

2.5.1.1 Apprentissage supervisé

L'apprentissage supervisé est une technique d'apprentissage automatique dans laquelle un modèle est entraîné sur un ensemble de données étiquetées ou labellisées. Cela signifie que les données d'entrée sont associées à des sorties ou réponses correctes connues. L'objectif est que le modèle apprenne les relations entre les données d'entrée et les sorties désirées à partir des exemples fournis. Pendant la phase d'entraînement, le modèle ajuste ses paramètres internes pour minimiser l'erreur entre ses prédictions et les véritables sorties. Une fois l'entraînement terminé, le modèle peut être utilisé pour prédire les sorties sur de nouvelles données n'ayant pas été vues durant l'apprentissage. Les tâches d'apprentissage supervisé courantes incluent la classification (prédire une catégorie ou une étiquette) et la régression (prédire une valeur numérique) [23].

Les problèmes d'apprentissage supervisé peuvent être sous les formes suivantes :

- **Classification** : est un problème fondamental en apprentissage supervisé où l'objectif est d'affecter une étiquette de classe à des instances de données d'entrée. Cela implique d'entraîner un modèle sur un ensemble de données étiquetées afin qu'il puisse apprendre les motifs et les caractéristiques qui distinguent les différentes classes. Pendant la phase d'apprentissage, l'algorithme de classification ajuste ses paramètres internes pour minimiser le taux d'erreur de classification sur les exemples d'entraînement. Une fois le modèle entraîné, il peut prédire avec précision la classe appropriée pour de nouvelles instances qui n'ont pas été utilisées pendant l'apprentissage [24].

- **Régression** : est une technique d'apprentissage supervisé utilisée pour prédire une valeur numérique continue à partir d'un ensemble de variables d'entrée. Contrairement à la classification qui prédit des catégories, la régression vise à modéliser la relation entre les variables d'entrée et une variable de sortie numérique cible. Pendant l'entraînement, l'algorithme de régression apprend les liens et les tendances sous-jacentes dans les données étiquetées afin de minimiser l'erreur entre ses prédictions et les valeurs cibles réelles [25].

2.5.1.2 Apprentissage non supervisé

L'apprentissage non supervisé est une branche de l'apprentissage automatique où l'on cherche à détecter des motifs et des structures sous-jacentes à partir de données d'entrée non étiquetées. Contrairement à l'apprentissage supervisé qui utilise des exemples étiquetés, ici on ne dispose pas de sorties ou de réponses correctes

associées aux données. L'objectif est d'explorer les données de manière autonome afin de découvrir des regroupements naturels, des tendances intéressantes ou des représentations compactes des données [23].

Quelques tâches typiques sont :

- **Clustering** : consiste à regrouper les données d'entrée non étiquetées en différents clusters ou groupes, de telle sorte que les exemples au sein d'un même cluster soient similaires entre eux, tandis que ceux de clusters différents soient distincts. C'est une tâche d'exploration de données qui vise à révéler les structures et les regroupements naturels présents dans l'ensemble des données [26].
- **Réduction de dimensionnalité** : comportent un très grand nombre de dimensions ou de caractéristiques, il peut être intéressant d'appliquer des techniques de réduction de dimensionnalité. Celles-ci projettent les données dans un sous-espace de dimension inférieure, tout en conservant autant que possible la structure et les informations importantes [27].
- **Détection d'anomalies** : ou de valeurs aberrantes consiste à identifier au sein d'un ensemble de données les observations qui dévient significativement de la norme ou du comportement attendu. En d'autres termes, détecter les points de données rares et inhabituels ne correspondant pas aux motifs généraux [28].

2.5.1.3 Apprentissage semi-supervisé

L'apprentissage semi-supervisé est une combinaison des approches supervisée et non supervisée. Il tire parti à la fois d'un petit ensemble de données étiquetées et d'une grande quantité de données non étiquetées. L'idée principale est d'utiliser les données étiquetées pour obtenir un modèle de base, puis d'exploiter les données non étiquetées abondantes pour affiner et améliorer les performances de ce modèle initial [29].

2.5.1.4 Apprentissage par renforcement

L'apprentissage par renforcement est une approche d'apprentissage automatique où un agent apprend à prendre des décisions optimales dans un environnement donné, de manière à maximiser une récompense cumulative sur le long terme. Contrairement à l'apprentissage supervisé qui utilise des exemples étiquetés, l'agent n'a accès qu'à un signal de récompense associé à chaque action qu'il entreprend [30].

2.5.2 Forêt aléatoire

Les forêts aléatoires, ou random forests, sont une technique d'apprentissage automatique supervisé utilisée principalement pour la classification et la régression. Elles constituent une amélioration des arbres de décision en utilisant une méthode d'ensemble pour augmenter la précision et réduire le surapprentissage [31].

Principe de fonctionnement des forêts aléatoires :

Une forêt aléatoire est constituée de nombreux arbres de décision indépendants. Chaque arbre est construit à partir d'un échantillon aléatoire du jeu de données d'entraînement et utilise une sélection aléatoire de caractéristiques pour chaque division de nœud, comme indiquer dans la **figure 2.12**. Voici les étapes principales :

2.5.2.1 Bootstrap Aggregating (Bagging)

- **Échantillonnage** : Pour chaque arbre de la forêt, un sous-échantillon du jeu de données est tiré avec remise. Si D est le jeu de données d'origine de taille N , alors chaque sous-échantillon D_i est également de taille N mais est obtenu par échantillonnage avec remise de D .
- **Entraînement** : Un arbre de décision est formé sur ce sous-échantillon. La fonction de décision pour un arbre T_i est $h_i(x)$.

2.5.2.2 Sélection aléatoire des caractéristiques

- **Divisions de nœud** : À chaque division de nœud dans un arbre, un sous-ensemble aléatoire de caractéristiques est considéré pour la meilleure division. Si le jeu de données a p caractéristiques, alors pour chaque division de nœud, m caractéristiques ($m < p$) sont sélectionnées aléatoirement, et la meilleure division est choisie parmi ces m caractéristiques.

2.5.2.3 Agrégation des prédictions

- **Classification** : Pour une nouvelle donnée x , chaque arbre de la forêt fait une prédiction $h_i(x)$, et la classe finale \hat{y} est déterminée par vote majoritaire des arbres :

$$\hat{y} = \text{mode}\{h_i(x)\}$$

- **Régression** : La prédiction finale \hat{y} est la moyenne des prédictions des arbres :

$$\hat{y} = \frac{1}{B} \sum_{i=1}^B h_i(x)$$

où B est le nombre total d'arbres dans la forêt.

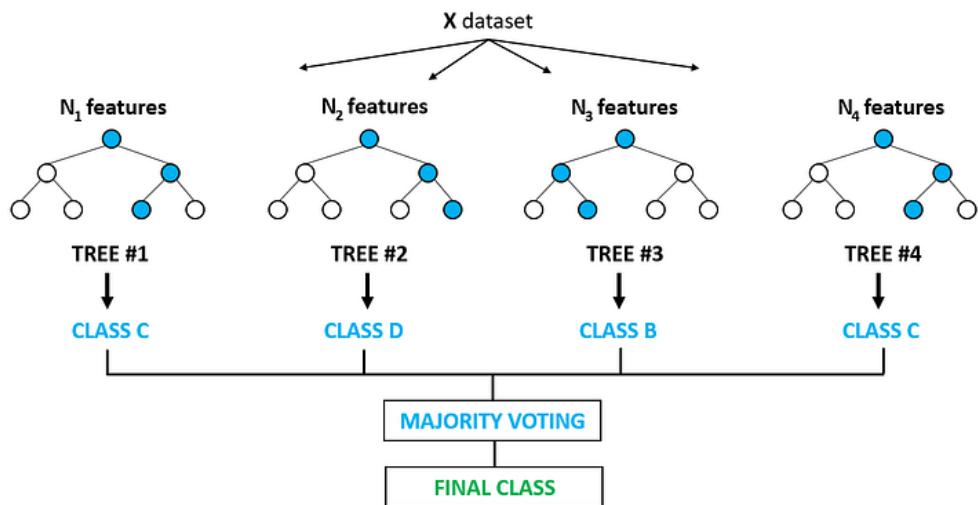


FIGURE 2.12 – Architecture de forêt aléatoire [31]

2.5.3 k-Nearest Neighbors (k-NN)

Le k-Nearest Neighbors (k-NN) est un algorithme d'apprentissage automatique non paramétrique utilisé pour les problèmes de classification et de régression. Il est particulièrement simple et intuitif, se basant sur la proximité des points de données pour faire des prédictions, comme dans la **figure 2.13**.

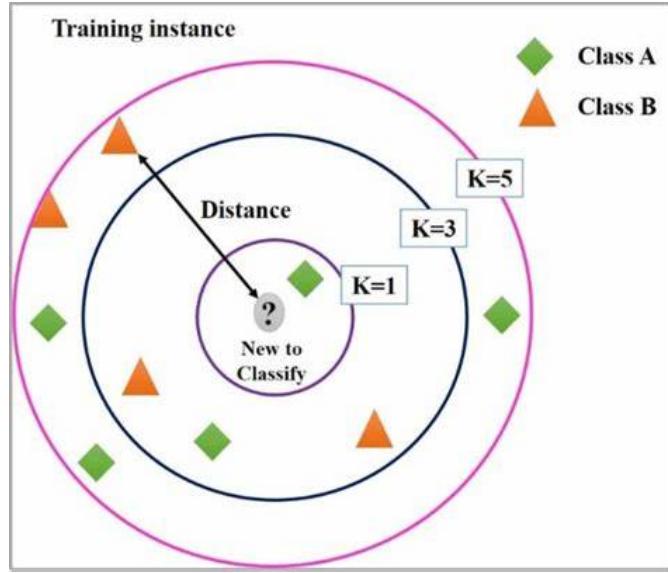


FIGURE 2.13 – Architecture de KNN [32]

Principe de fonctionnement de KNN :

L'algorithme k-NN fonctionne selon les étapes suivantes [32] :

2.5.3.1 Stockage des données

L'algorithme k-NN ne construit pas de modèle explicite. Au lieu de cela, il stocke simplement tous les exemples d'entraînement.

2.5.3.2 Calcul des distances

Pour prédire la classe ou la valeur d'un nouvel échantillon, l'algorithme calcule la distance entre cet échantillon et tous les échantillons d'entraînement. La distance la plus couramment utilisée est la distance euclidienne, définie par :

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

où x et y sont des points de données dans un espace n -dimensionnel.

2.5.3.3 Sélection des voisins

Les k points d'entraînement les plus proches (voisins) sont sélectionnés en fonction des distances calculées.

2.5.3.4 Vote ou moyenne

- **Classification** : La classe de l'échantillon est déterminée par un vote majoritaire des k voisins. La classe la plus fréquente parmi les voisins est attribuée à l'échantillon.
- **Régression** : La valeur de l'échantillon est déterminée par la moyenne des valeurs des k voisins les plus proches.

2.5.3.5 Choix du paramètre k

Le paramètre k représente le nombre de voisins à considérer pour faire une prédiction. Un petit k peut rendre le modèle sensible au bruit dans les données, tandis qu'un k trop grand peut inclure trop de points dans la prédiction, diluant les décisions locales. Une méthode courante pour choisir k consiste à utiliser la validation croisée.

2.5.4 Apprentissage profond

L'apprentissage profond, ou deep learning en anglais, est une sous-catégorie de l'apprentissage automatique. Il utilise des réseaux de neurones artificiels composés de multiples couches pour apprendre à partir de données massives, comme dans la **figure 2.14**.

2.5.4.1 Réseaux de Neurones Artificiels (ANN)

Les réseaux de neurones artificiels (ANN) sont des modèles computationnels inspirés du cerveau humain, utilisés dans l'apprentissage automatique pour résoudre divers problèmes complexes. Ils sont composés de neurones artificiels organisés en couches et sont capables d'apprendre des représentations à partir de données d'entrée pour faire des prédictions ou des classifications. Voici une description détaillée de la structure et du fonctionnement des ANN [33].

Structure des ANN

Un réseau de neurones artificiels typique est constitué de trois types de couches :

Couche d'entrée (Input Layer)

- Reçoit les données d'entrée brutes.
- Chaque neurone dans cette couche représente une caractéristique de l'ensemble de données.

Couches cachées (Hidden Layers)

- Effectuent des transformations et des calculs complexes sur les données d'entrée.
- Un ANN peut avoir plusieurs couches cachées, chacune permettant au réseau d'apprendre des caractéristiques de plus en plus abstraites.

Couche de sortie (Output Layer)

- Produit la sortie finale du réseau.
- Le nombre de neurones dans cette couche correspond au nombre de classes pour un problème de classification ou à une seule valeur pour un problème de régression.

Fonctionnement des ANN

Propagation Avant (Forward Propagation)

- Les données d'entrée traversent le réseau de la couche d'entrée à la couche de sortie.
- Chaque neurone effectue une somme pondérée de ses entrées suivie d'une fonction d'activation pour produire une sortie.

Formule de calcul pour un neurone j dans une couche l :

$$z_j^{(l)} = \sum_{i=1}^n w_{ij}^{(l)} a_i^{(l-1)} + b_j^{(l)}$$
$$a_j^{(l)} = \sigma(z_j^{(l)})$$

où $w_{ij}^{(l)}$ est le poids entre le neurone i de la couche $l-1$ et le neurone j de la couche l , $b_j^{(l)}$ est le biais du neurone j dans la couche l , et σ est la fonction d'activation (par exemple, ReLU, sigmoïde, tanh).

Rétropropagation (Backpropagation)

- Utilisée pour entraîner le réseau en ajustant les poids et les biais pour minimiser l'erreur de prédiction.
- Calcule les gradients de la fonction de perte par rapport aux poids du réseau en utilisant la règle de la chaîne.

Formule de mise à jour des poids :

$$w_{ij}^{(l)} \leftarrow w_{ij}^{(l)} - \eta \frac{\partial L}{\partial w_{ij}^{(l)}}$$

où η est le taux d'apprentissage et $\frac{\partial L}{\partial w_{ij}^{(l)}}$ est le gradient de la perte L par rapport au poids $w_{ij}^{(l)}$.

Types de fonctions d'activation

Les fonctions d'activation introduisent la non-linéarité dans le réseau, permettant d'apprendre des modèles complexes. Les plus courantes sont :

Sigmoïde

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Utilisée dans les couches de sortie pour des problèmes de classification binaire.

Tanh

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

Utilisée dans les couches cachées pour centrer les sorties autour de zéro.

ReLU (Rectified Linear Unit)

$$\text{ReLU}(x) = \max(0, x)$$

Largement utilisée dans les couches cachées en raison de ses avantages en termes de convergence rapide.

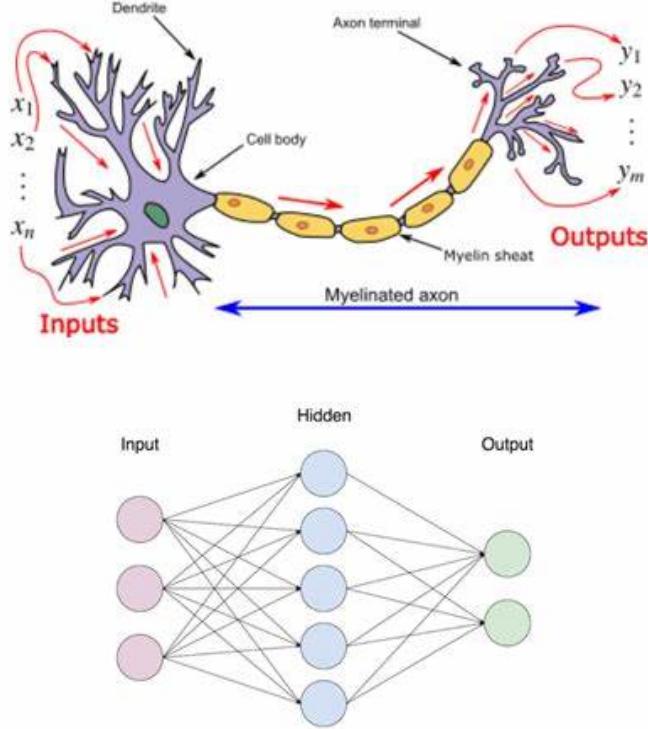


FIGURE 2.14 – Architecture de ANN [33]

2.5.4.2 Réseaux de neurones convolutionnels (CNN)

Les réseaux de neurones convolutionnels (CNN) sont une classe spécialisée de réseaux de neurones profonds pour le traitement de données à structure de grille comme les images. Un CNN est composé d'une succession de couches convolutionnelles qui appliquent des opérations de convolution sur les données d'entrée pour extraire des caractéristiques locales [33].

La couche convolutionnelle calcule la sortie \mathbf{y} à partir de l'entrée \mathbf{x} comme suit :

$$\mathbf{y} = f \left(\sum_{i=1}^D \mathbf{w}_i * \mathbf{x}_i + \mathbf{b} \right)$$

où \mathbf{w}_i sont les D noyaux de convolution appris, $*$ représente l'opération de convolution, \mathbf{b} le biais, et f une fonction d'activation non linéaire comme ReLU.

Les couches de pooling réduisent ensuite la dimension spatiale avec une opération comme le max-pooling :

$$\mathbf{y}_{\text{pool}} = \max(i, j) \in \mathcal{R} \mathbf{y}_{i,j}$$

où \mathcal{R} représente une région de la carte de caractéristiques \mathbf{y} .

Les couches convolutionnelles et de pooling s'enchaînent pour extraire des caractéristiques de plus en plus abstraites et invariantes aux transformations locales.

Le CNN se termine par des couches fully-connected classiques pour la tâche finale comme :

$$\hat{\mathbf{y}} = \text{softmax}(\mathbf{W}^T \mathbf{x} + \mathbf{b})$$

où \mathbf{x} sont les caractéristiques extraites et $\hat{\mathbf{y}}$ les scores de prédiction.

L'entraînement d'un CNN consiste à optimiser les paramètres \mathbf{w} , \mathbf{W} et \mathbf{b} par rétropropagation du gradient pour minimiser une fonction de perte sur un grand ensemble de données étiquetées.

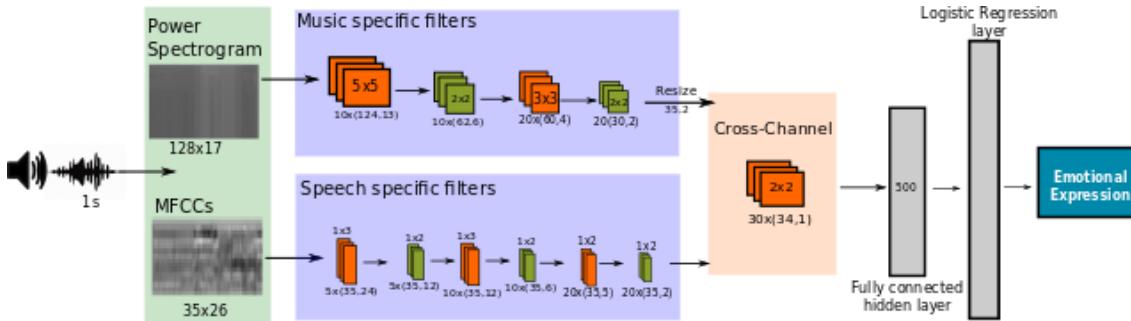


FIGURE 2.15 – Architecture de CNN [33]

2.6 Méthodes d'évaluation

La dernière étape cruciale pour un système de détection des pathologies vocales est l'évaluation du modèle de classification. En effet, un modèle n'est pas forcément toujours fiable, c'est pourquoi il est obligatoire d'évaluer ses performances afin de s'assurer qu'il contribuera correctement à prédire la cible pour de nouvelles données. Puisque ces nouvelles données auront une cible inconnue que nous cherchons à prédire, l'évaluation se fait sur des données labellisées dont nous connaissons déjà les cibles réelles.

Plusieurs métriques et approches existent pour évaluer un modèle de classification. Une première approche consiste à utiliser une partie des données pour l'entraînement du modèle, et le reste pour tester ses prédictions. On peut alors calculer la matrice de confusion qui résume les prédictions correctes et incorrectes, ainsi que diverses métriques dérivées comme la précision, le rappel et le score F1 [34].

2.6.1 Matrice de confusion

La matrice de confusion est un outil d'évaluation essentiel pour les problèmes de classification en apprentissage supervisé. Elle permet de visualiser les performances

d'un modèle de classification en résumant les prédictions correctes et incorrectes pour chaque classe. Pour un problème à deux classes (positif et négatif), la matrice de confusion est une matrice 2x2 définie comme :

[Vrais Positifs (VP) Faux Négatifs (FN) Faux Positifs (FP) Vrais Négatifs (VN)]

Où les entrées représentent :

- VP : nombre d'exemples positifs correctement classés
- FN : nombre d'exemples positifs incorrectement classés comme négatifs
- FP : nombre d'exemples négatifs incorrectement classés comme positifs
- VN : nombre d'exemples négatifs correctement classés

À partir de cette matrice, on peut calculer diverses métriques pour évaluer les performances du classifieur :

$$\text{Précision} = \frac{\text{VP}}{\text{VP} + \text{FP}} \quad (2.3)$$

$$\text{Rappel} = \frac{\text{VP}}{\text{VP} + \text{FN}} \quad (2.4)$$

$$\text{Score F1} = 2 \cdot \frac{\text{Précision} \cdot \text{Rappel}}{\text{Précision} + \text{Rappel}} \quad (2.5)$$

La précision mesure la capacité du modèle à ne pas se tromper sur les exemples positifs, tandis que le rappel mesure sa capacité à détecter tous les exemples positifs. Le score F1 est une moyenne harmonique combinant précision et rappel. Pour les problèmes multi-classes, on génère une matrice de confusion où chaque ligne représente les instances d'une classe réelle et chaque colonne les instances d'une classe prédite.

		Reality	
Confusion matrix		Negative : 0	Positive : 1
Prediction	Negative : 0	True Negative : TN	False Negative : FN
	Positive : 1	False Positive : FP	True Positive : TP

FIGURE 2.16 – Matrice de confusion [35]

2.6.2 Courbe ROC

La courbe ROC (Receiver Operating Characteristic) est un outil graphique puissant permettant d'évaluer les performances d'un modèle de classification binaire.

Elle représente le compromis entre les taux de vrais positifs (sensibilité) et de faux positifs ($1 - \text{spécificité}$) pour différents seuils de décision du modèle.

Pour construire la courbe ROC, on trace la sensibilité en fonction de ($1 - \text{spécificité}$), où :

$$\text{Sensibilité} = \frac{\text{Vrais Positifs}}{\text{Vrais Positifs} + \text{Faux Négatifs}} \quad (2.6)$$

$$\text{Spécificité} = \frac{\text{Vrais Négatifs}}{\text{Vrais Négatifs} + \text{Faux Positifs}} \quad (2.7)$$

Un classifieur parfait aurait une courbe ROC passant par le point $(0, 1)$, représentant 100% de sensibilité et 0% de faux positifs. Un classifieur aléatoire correspondrait à la droite diagonale.

L'aire sous la courbe ROC (AUC) est une métrique résumant les performances du classifieur. Une AUC de 1 représente un modèle parfait, tandis qu'une AUC de 0,5 correspond à un classifieur aléatoire.

La courbe ROC est particulièrement utile lorsque les classes sont déséquilibrées ou que les coûts des différents types d'erreurs de classification sont différents. Elle permet de sélectionner le seuil optimal pour le modèle en fonction des contraintes de l'application.

De plus, comparer les courbes ROC de différents modèles permet d'identifier celui ayant les meilleures performances de classification, indépendamment des distributions de classe ou des seuils de décision.

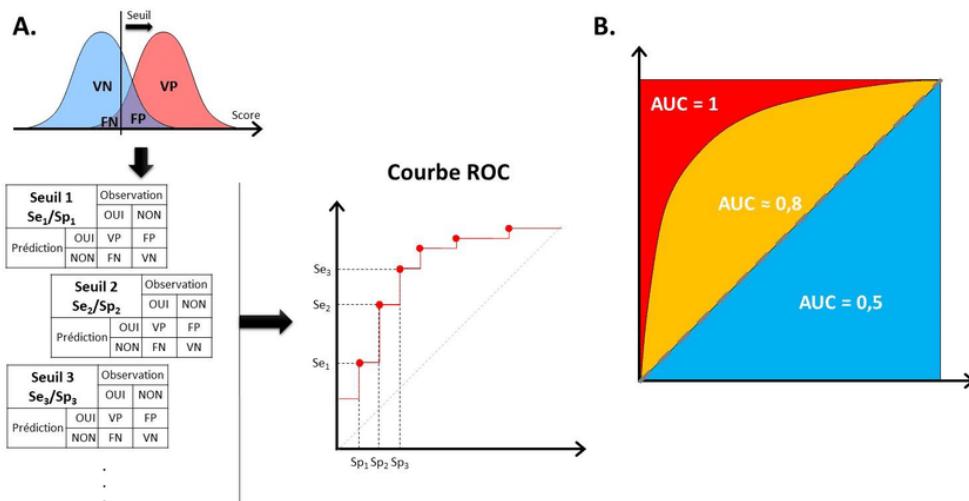


FIGURE 2.17 – Courbe ROC [36]

2.6.3 Validation croisée

La validation croisée, également appelée "cross-validation" en anglais, est une technique statistique couramment utilisée pour évaluer les performances de prédiction d'un modèle d'apprentissage automatique et estimer sa capacité à bien généraliser sur de nouvelles données. Le principe est le suivant : au lieu d'utiliser un seul échantillon de données pour l'entraînement et un autre pour le test, on divise de manière aléatoire l'ensemble des données d'apprentissage en k sous-ensembles (communément $k=10$ pour la "stratified 10-fold cross-validation"). On entraîne ensuite le modèle k fois de façon itérative, en prenant à chaque fois un sous-ensemble différent comme ensemble de validation, tandis que les $(k-1)$ autres sous-ensembles servent à l'entraînement. À la fin de la procédure, on calcule la moyenne des scores d'erreur obtenus sur les k itérations pour obtenir une estimation fiable de l'erreur de généralisation du modèle. Cette approche permet de limiter le risque de surapprentissage (overfitting) ainsi que la variance élevée associée à un unique échantillonnage aléatoire des données d'entraînement et de test.

La validation croisée est une technique standard dans la phase d'évaluation et de sélection d'un modèle, notamment pour le choix du meilleur paramétrage ou du meilleur algorithme.

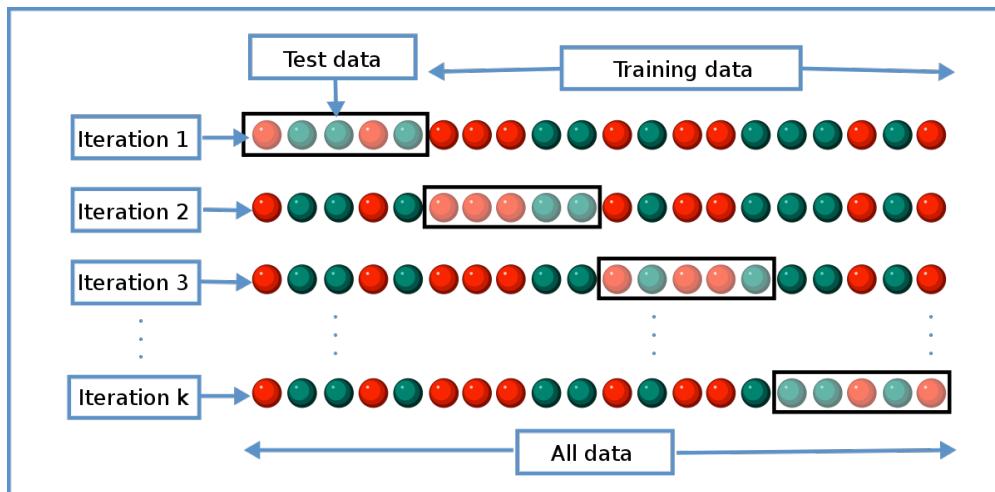


FIGURE 2.18 – Validation croisée [37]

2.6.4 Calcul de confiance

Pour évaluer la fiabilité de la détection des émotions par la voix, nous utilisons une classification basée sur un score de confiance calculé. Soit C le pourcentage de confiance calculé pour la détection de l'émotion dans la voix. Le niveau de confiance de cette détection est alors défini par la relation suivante [38] :

$$\text{Niveau de confiance} = \begin{cases} \text{Élevé (high),} & \text{si } C \geq 80 \\ \text{Moyen (medium),} & \text{si } 50 \leq C < 80 \\ \text{Faible (low),} & \text{si } C < 50 \end{cases}$$

Cette classification permet d'évaluer rapidement la fiabilité de la détection des émotions vocales. Un niveau de confiance élevé indique une forte probabilité que l'émotion détectée soit correcte, tandis qu'un niveau faible suggère une détection moins fiable, potentiellement due à des facteurs tels que la qualité de l'enregistrement, l'ambiguïté de l'expression émotionnelle, ou des limitations du modèle de détection.

2.7 Exemples des travaux en détection automatique des émotions avec la voix

Dans le domaine de la reconnaissance des émotions par la parole, plusieurs études ont exploré l'utilisation de techniques de machine learning et de deep learning pour analyser des données issues de bases de données variées telles que RAVDESS, TESS, Emo-DB, et SAVEE. Ces travaux utilisent des modèles tels que les réseaux de neurones convolutifs (CNN), les machines à vecteurs de support (SVM) et les réseaux multicouches de perceptrons (MLP), avec des résultats prometteurs quant à l'efficacité de ces méthodes pour cette tâche complexe.

- **Recognizing emotion from Speech using Machine learning and Deep learning** : cet article [39] explore la reconnaissance des émotions dans la voix en utilisant des algorithmes de machine learning et de deep learning. Utilisant les bases de données RAVDESS et TESS, les auteurs mettent en œuvre des modèles de CNN, SVM, et MLP. Les résultats montrent une précision de 82%, démontrant l'efficacité de l'utilisation de ces techniques pour cette tâche complexe.

- Speech Emotion Recognition Using Deep Learning Techniques : A Review : cet article [40] propose une revue des techniques de deep learning appliquées à la reconnaissance des émotions dans la parole. En se concentrant sur les bases de données RAVDESS, Emo-DB, et SAVEE, il met en évidence les architectures CNN et LSTM comme étant particulièrement efficaces, atteignant une précision de 85%. Cette revue est disponible sur IEEE Xplore et offre un aperçu complet des avancées dans ce domaine.

- Human Speech Emotion Recognition : L'article [41] présente un modèle de reconnaissance des émotions dans la parole utilisant les bases de données EMO-DB et RAVDESS. Les auteurs utilisent des caractéristiques spectrales et des réseaux de neurones convolutifs et récurrents (CNN et RNN), obtenant une précision de 83%. Ce travail est accessible sur ResearchGate et constitue une ressource précieuse pour les chercheurs dans ce domaine.

- Speech Emotion Recognition Based on Two-Stream Deep Learning Model : Dans cet article [42], les auteurs introduisent un modèle de reconnaissance des émotions à double flux, utilisant des caractéristiques audios et des réseaux de neurones récurrents (CNN et GRU). En se basant sur les bases de données IEMOCAP et RAVDESS, ils atteignent une précision de 87%. L'article est publié sur MDPI et propose une approche innovante pour améliorer la reconnaissance des émotions.

- Identification of emotions from speech using Deep Learning : cet article [43] explore la reconnaissance des émotions en utilisant des techniques de deep learning appliquées aux bases de données RAVDESS et EMO-DB. Les auteurs utilisent des modèles CNN, LSTM, et GRU pour atteindre une précision de 88%. Cette étude est accessible sur IEEE Xplore et démontre des résultats prometteurs dans le domaine de la reconnaissance des émotions.

Conclusion

Ce chapitre a exploré les différentes étapes et techniques utilisées pour la reconnaissance des émotions à partir de la voix. Nous avons d'abord examiné l'importance des bases de données vocales émotionnelles, qui fournissent des enregistrements variés et de haute qualité, essentiels pour l'entraînement des modèles. Ensuite, nous avons détaillé les étapes de pré-traitement des données, y compris la normalisation, la suppression du bruit, et la segmentation.

L'extraction des caractéristiques s'est concentrée sur les MFCC (Mel-Frequency Cepstral Coefficients), le RMSE (Root Mean Square Energy), et le ZCR (Zero Crossing Rate), toutes cruciales pour capturer les propriétés acoustiques de la voix. Pour la visualisation des émotions, nous avons exploré les techniques du signal d'onde, du spectrogramme, et du chroma, qui offrent des représentations graphiques des caractéristiques vocales liées aux émotions.

Nous avons également exploré diverses techniques de classification, notamment les forêts aléatoires, k-Nearest Neighbors (k-NN), réseaux de neurones artificiels (ANN) et réseaux convolutionnels (CNN), chacune présentant des avantages spécifiques pour la reconnaissance des émotions.

L'évaluation des modèles a été discutée en termes d'utilisation de la matrice de confusion, de la courbe ROC et de la validation croisée pour mesurer la performance des modèles. Enfin, plusieurs études de cas ont démontré l'efficacité des techniques de machine learning et de deep learning pour la reconnaissance des émotions.

CHAPITRE 3

CONTEXTE EXPÉRIMENTAL ET RÉSULTAT

Introduction

Dans ce chapitre, nous explorons en détail les technologies, outils et méthodes utilisés pour développer notre application de détection des émotions par la voix, ainsi que les résultats obtenus. Nous avons utilisé **Bootstrap** pour le front-end afin de créer des interfaces utilisateur modernes et réactives, et **Flask** pour le back-end en raison de sa simplicité et flexibilité. Les interfaces utilisateur incluent des pages d'accueil avant et après authentification, des pages pour enregistrer ou importer un audio, ainsi qu'une page d'affichage des résultats, offrant ainsi une expérience utilisateur fluide.

Pour le développement des modèles de machine learning, nous avons exploité des bibliothèques Python telles que **TensorFlow** pour la création de réseaux de neurones et **Librosa** pour l'analyse des fichiers audio. Nous avons utilisé plusieurs bases de données, dont CREMA-D, RAVDESS, SAVEE, et TESS, pour entraîner et évaluer nos modèles, en appliquant des étapes de prétraitement comme la normalisation du volume et l'élimination des silences pour optimiser les données audio. Les caractéristiques telles que les MFCCs, RMSE, et ZCR ont été extraites pour entraîner nos modèles.

Nous avons expérimenté différentes techniques d'apprentissage, comprenant des méthodes d'apprentissage automatique (comme k-nearest neighbors (k-NN) et les forêts aléatoires) et d'apprentissage profond (comme les réseaux de neurones artificiels (ANN) et les réseaux de neurones convolutifs (CNN)). Les résultats obtenus avec ces méthodes, y compris les matrices de confusion et les courbes ROC, sont analysés en détail pour évaluer leur efficacité dans la classification des émotions.

3.1 Environnement de travail

3.1.1 Technologies utilisées pour le développement du site web de détection des émotions

3.1.1.1 Front-end de l'application

- **Bootstrap :**

Bootstrap est une bibliothèque front-end open-source initialement développée par Twitter. Elle fournit des outils pour créer des interfaces utilisateur modernes et réactives. Bootstrap comprend une collection de composants CSS et JavaScript tels que des grilles, des boutons, des formulaires, des modals et des carrousels, qui permettent de concevoir des sites web de manière rapide et efficace [44].



FIGURE 3.1 – Bootstrap [44]

3.1.1.2 Back-end de l'application

- **Flask :**

Flask est un micro-framework web écrit en Python. Il est conçu pour être simple, flexible et léger, ce qui en fait un choix populaire pour le développement de petites applications web et de prototypes. Flask est livré avec un serveur web intégré et un débogueur, et il supporte les extensions qui peuvent ajouter des fonctionnalités comme la gestion de formulaires, la validation des données, la gestion des sessions, etc [45].



FIGURE 3.2 – Flask [45]

3.1.2 Outils et bibliothèques utilisés pour la création du modèle de détection des émotions

3.1.2.1 Outils

- Kaggle :

Kaggle est une plateforme en ligne dédiée à la science des données et à l'apprentissage automatique. Elle permet aux utilisateurs de partager des ensembles de données, de collaborer sur des projets de data science, et de participer à des compétitions d'apprentissage automatique pour résoudre des défis réels. Kaggle fournit également un environnement de développement intégré (IDE) basé sur le cloud, ce qui facilite le travail collaboratif et l'expérimentation avec différents modèles et algorithmes [46].



FIGURE 3.3 – Kaggle [46]

3.1.2.2 Language de programmation

- Python :

Python est un langage de programmation interprété, de haut niveau et à usage général. Créé par Guido van Rossum et publié pour la première fois en 1991, Python met l'accent sur la lisibilité du code avec une syntaxe qui permet aux programmeurs d'exprimer des concepts en moins de lignes de code qu'en utilisant des langages comme C++ ou Java. Il supporte plusieurs paradigmes de programmation, notamment la programmation procédurale, orientée objet et fonctionnelle. Python est largement utilisé dans diverses applications, notamment le développement web, l'analyse de données, l'automatisation, l'intelligence artificielle et le développement de logiciels [47].



FIGURE 3.4 – Python [47]

3.1.2.3 Bibliothèques utilisés

- Tensorflow :

TensorFlow est une bibliothèque open-source de Google dédiée à l'apprentissage automatique et à l'intelligence artificielle. Initialement développée par l'équipe Google Brain, TensorFlow permet de créer et de déployer des modèles de machine learning, en particulier des réseaux de neurones profonds. Elle offre une flexibilité et une extensibilité considérables, ce qui en fait un outil privilégié pour les chercheurs et les développeurs dans le domaine de l'intelligence artificielle. TensorFlow est utilisé pour une variété d'applications, y compris la reconnaissance d'images, la traduction automatique, et bien d'autres tâches de traitement de données complexes [48].



FIGURE 3.5 – Tensorflow [48]

- Librosa :

Librosa est une bibliothèque Python open-source spécialisée dans l'analyse et le traitement de fichiers audio. Elle est largement utilisée dans les domaines de l'apprentissage automatique et du traitement du signal pour extraire des caractéristiques audio, telles que les MFCC (Mel-frequency cepstral coefficients), le chromatogramme, le spectrogramme et bien d'autres. librosa offre une interface simple et intuitive pour charger, transformer et visualiser des données audio, facilitant ainsi le développement et l'expérimentation de modèles de machine learning basés sur des données audio [49].



FIGURE 3.6 – Librosa [49]

3.2 Conception de l'application pour la détection des émotions

3.2.1 Diagramme de classe

Ce diagramme dans la **figure 3.7** représente l'architecture du système sous forme de classes UML. Il montre les différentes composantes de l'application : Utilisateur, Authentification, FichierAudio, Microphone, AnalyseurEmotions et Resultat. Chaque classe est détaillée avec ses attributs et méthodes. Les relations entre les classes sont indiquées par des flèches, montrant par exemple que l'Utilisateur utilise l'Authentification, importe un FichierAudio ou utilise le Microphone. L'AnalyseurEmotions reçoit l'audio soit du FichierAudio, soit du Microphone, et produit un Resultat. Ce diagramme offre une vue d'ensemble de la structure interne de l'application et de l'interaction entre ses différents composants.

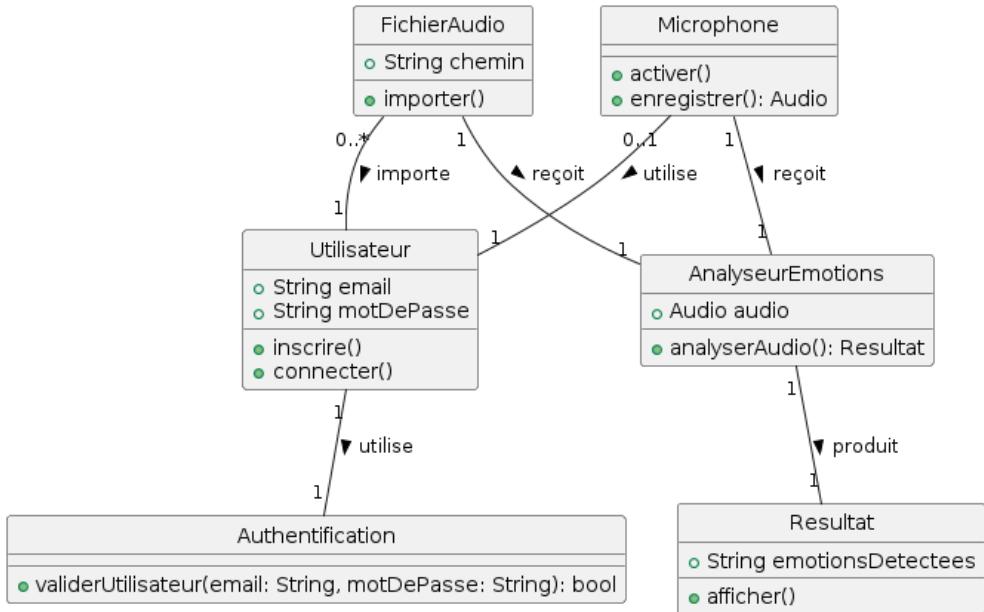


FIGURE 3.7 – Diagramme de classe

3.2.2 Diagramme de cas d'utilisation

Ce diagramme dans la **figure 3.8** présente le flux d'utilisation d'une application de détection des émotions. Le processus commence avec l'utilisateur qui s'inscrit, puis se connecte à l'application. Après l'authentification, l'utilisateur a deux options : soit allumer le microphone de son PC, soit importer un fichier audio. Ces deux chemins mènent à la même finalité : l'affichage du résultat de la détection des émotions. Le schéma illustre clairement la séquence linéaire des actions, chaque étape étant liée à la suivante par une flèche pointillée accompagnée du mot "suit" ou "authentifié".

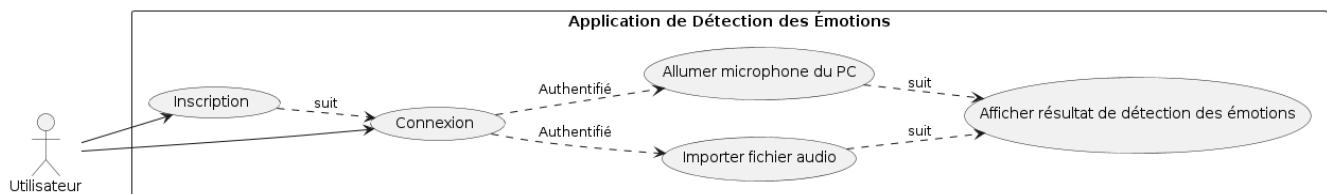


FIGURE 3.8 – Diagramme de cas d'utilisation

Nous présentons un modèle de classification des émotions provoquées par des discours en utilisant des approches d'apprentissage automatique (ML) et d'apprentissage profond (DL). Pour le ML, nous avons utilisé les algorithmes des k-plus proches voisins (K-Nearest Neighbors, KNN) et la forêt aléatoire (Random Forest). Pour le DL, nous avons travaillé avec des réseaux de neurones artificiels (Artificial Neural Networks, ANN) et des réseaux de neurones convolutifs (Convolutional Neural Networks, CNN). Ces modèles sont basés sur des caractéristiques acoustiques telles que le Mel Frequency Cepstral Coefficient (MFCC), le Root Mean Square Energy (RMSE) et le Zero-Crossing Rate (ZCR). Le modèle a été formé pour classer sept émotions différentes : neutre, heureux, triste, en colère, craintif, dégoût, et surprise, comme illustré dans la **figure 3.9**.

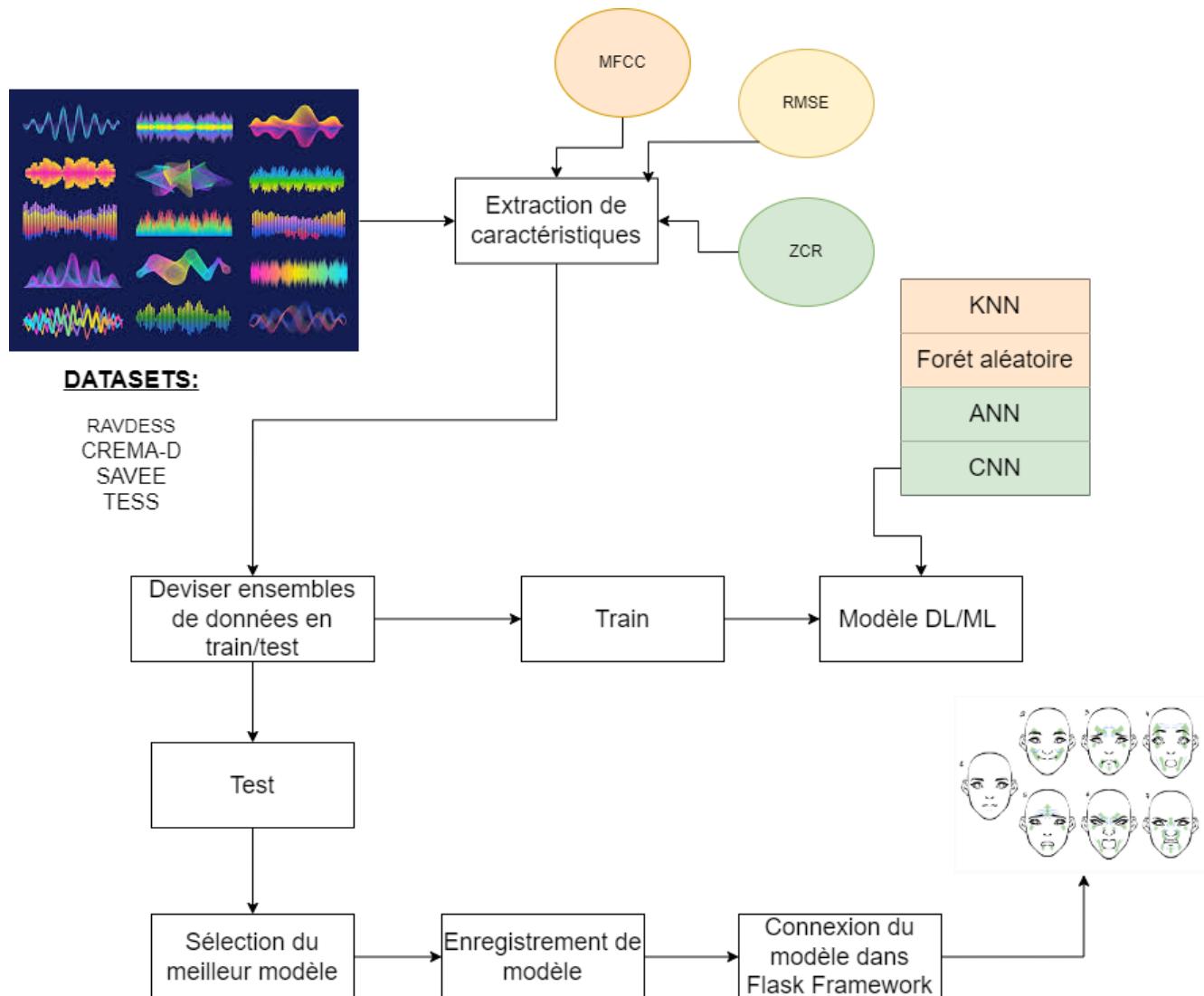


FIGURE 3.9 – Architecture générale de l'application

3.3 Bases de données utilisées

Pour l'analyse des émotions exprimées dans la parole, nous utilisons quatre bases de données qui offrent une grande diversité d'enregistrements. Voici une description détaillée de chaque ensemble de données, y compris le genre des participants, l'âge, le nombre d'enregistrements audio par émotion, la fréquence d'échantillonnage et la durée des enregistrements :

- **L'ensemble de données CREMA-D [50]** (Crowd-sourced Emotional Multimodal Actors Dataset) est un ensemble multimodal comprenant des enregistrements audio-visuels d'acteurs performants diverses expressions émotionnelles. Ce corpus inclut 91 acteurs (48 hommes et 43 femmes) âgés de 20 à 74 ans, couvrant six émotions : colère (1 271), peur (1 271), bonheur (1 271), tristesse (1 271), dégoût (1 271), et neutre (1 087). Chaque acteur exprime ces émotions à travers 12 phrases, ce qui donne un total de 7 442 clips audio avec une fréquence d'échantillonnage de 16 kHz et une durée moyenne d'environ 2 secondes par clip.
- **L'ensemble de données RAVDESS [51]** (Ryerson Audio-Visual Database of Emotional Speech and Song) se compose d'enregistrements d'acteurs (12 hommes et 12 femmes) qui expriment huit émotions différentes, notamment la colère (192), la peur (192), le bonheur (192), la tristesse (192), la surprise (192), le dégoût (192), le calme (288), et une émotion neutre (288). L'âge des acteurs varie de 20 à 50 ans. L'ensemble comprend 1 440 clips de discours et 1 008 clips de chansons avec une fréquence d'échantillonnage de 48 kHz. La durée des enregistrements de discours est en moyenne de 4 secondes, tandis que les enregistrements de chansons durent environ 12 secondes.
- **L'ensemble de données SAVEE [52]** (Surrey Audio-Visual Expressed Emotion) contient des enregistrements audio-visuels de quatre hommes britanniques exprimant sept émotions : colère (60), peur (60), bonheur (60), tristesse (60), surprise (60), dégoût (60), et neutre (120). Chaque acteur est âgé de 27 à 31 ans. L'ensemble de données compte un total de 480 fichiers audio enregistrés avec une fréquence d'échantillonnage de 44,1 kHz et une durée moyenne d'environ 3 à 4 secondes par clip.

- L’ensemble de données TESS [53] (Toronto Emotional Speech Set) est constitué d’enregistrements audio où deux actrices (âgées de 26 et 64 ans) prononcent des phrases neutres de contenu tout en exprimant sept émotions : colère (400), peur (400), bonheur (400), tristesse (400), surprise (400), dégoût (400), et neutre (400). L’ensemble de données comprend 2 800 enregistrements audio avec une fréquence d’échantillonnage de 22,05 kHz et une durée moyenne de 3 à 4 secondes par enregistrement.

Ces ensembles de données couvrent une large gamme de genres, d’âges, d’émotions, de fréquences d’échantillonnage et de durées, offrant une base solide pour affiner les modèles de reconnaissance émotionnelle.

Après avoir regroupé les quatre bases de données [RAVDESS, CREMA-D, TESS et SAVEE], les totaux des enregistrements par émotion ont été calculés. L’émotion “dégoût” compte un total de 1 923 enregistrements, tout comme les émotions “peur”, “tristesse”, “bonheur” et “colère”, qui ont également 1 923 enregistrements chacune. L’émotion “neutre” est légèrement inférieure avec 1 895 enregistrements, tandis que “surprise” a 652 enregistrements.

La visualisation dans la **figure 3.10** montre le nombre total d’enregistrements pour chaque émotion après le regroupement de ces bases de données, offrant une perspective claire sur la distribution des émotions représentées dans l’ensemble consolidé. Cette distribution homogène, à l’exception de “surprise”, montre un équilibre dans les échantillons pour chaque émotion, ce qui est bénéfique pour la formation de modèles de reconnaissance émotionnelle, réduisant les biais potentiels dus à des déséquilibres de classe.

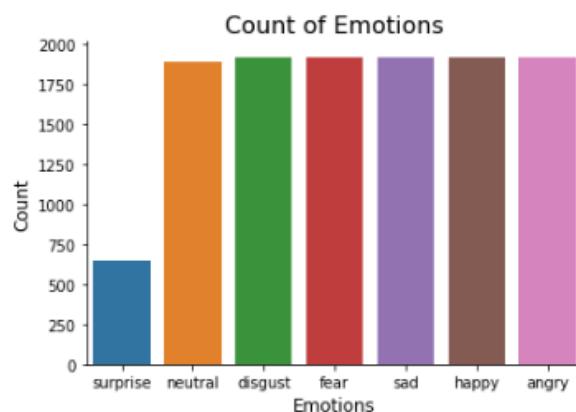


FIGURE 3.10 – Distribution des enregistrements d’émotions après regroupement

3.4 Représentation des caractéristiques pour chaque émotion

3.4.1 Représentation des caractéristiques de l'émotion "Neutral"

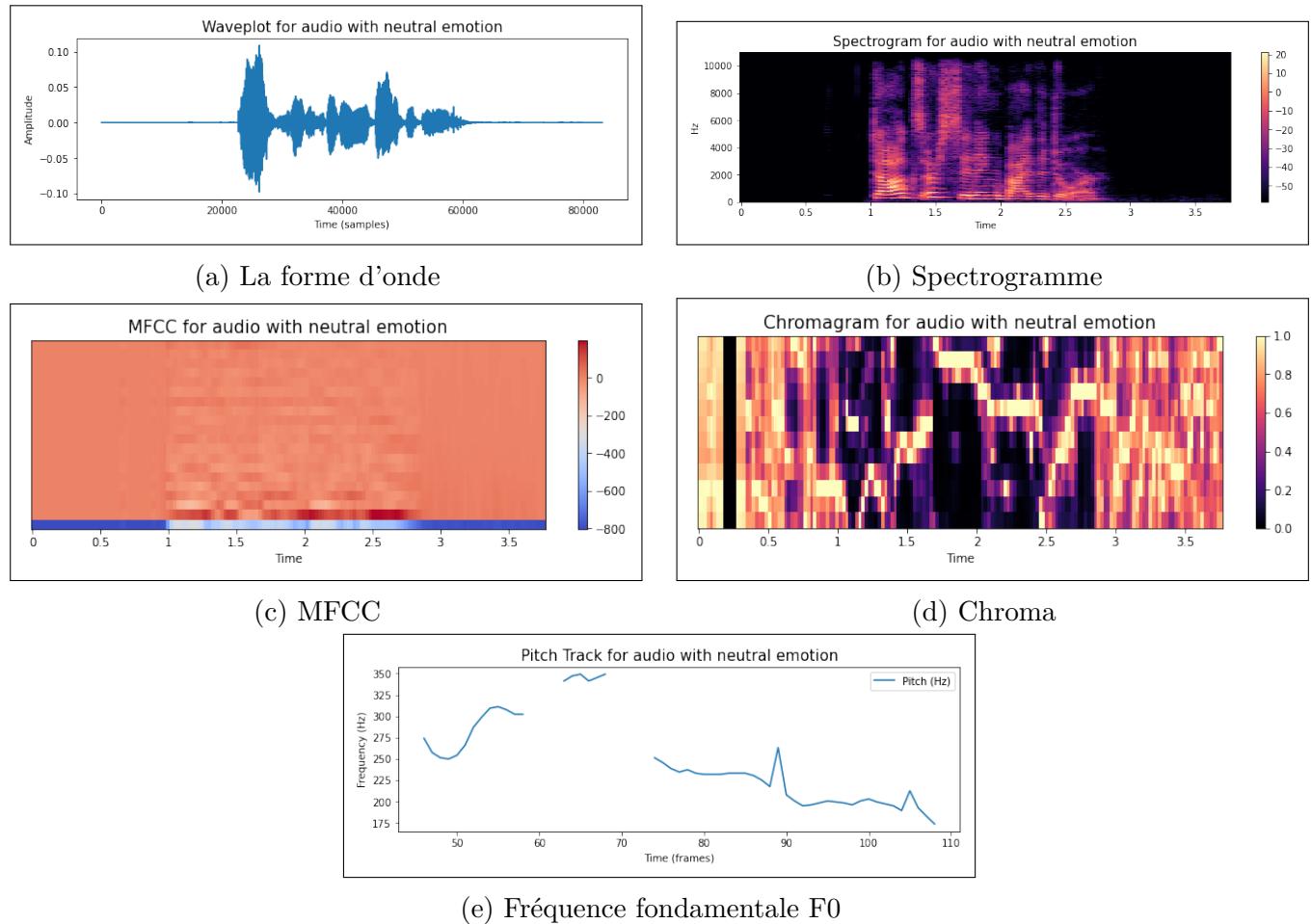


FIGURE 3.11 – Caractéristiques d'émotion "Neutral"

3.4.2 Représentation des caractéristiques de l'émotion "Happy"

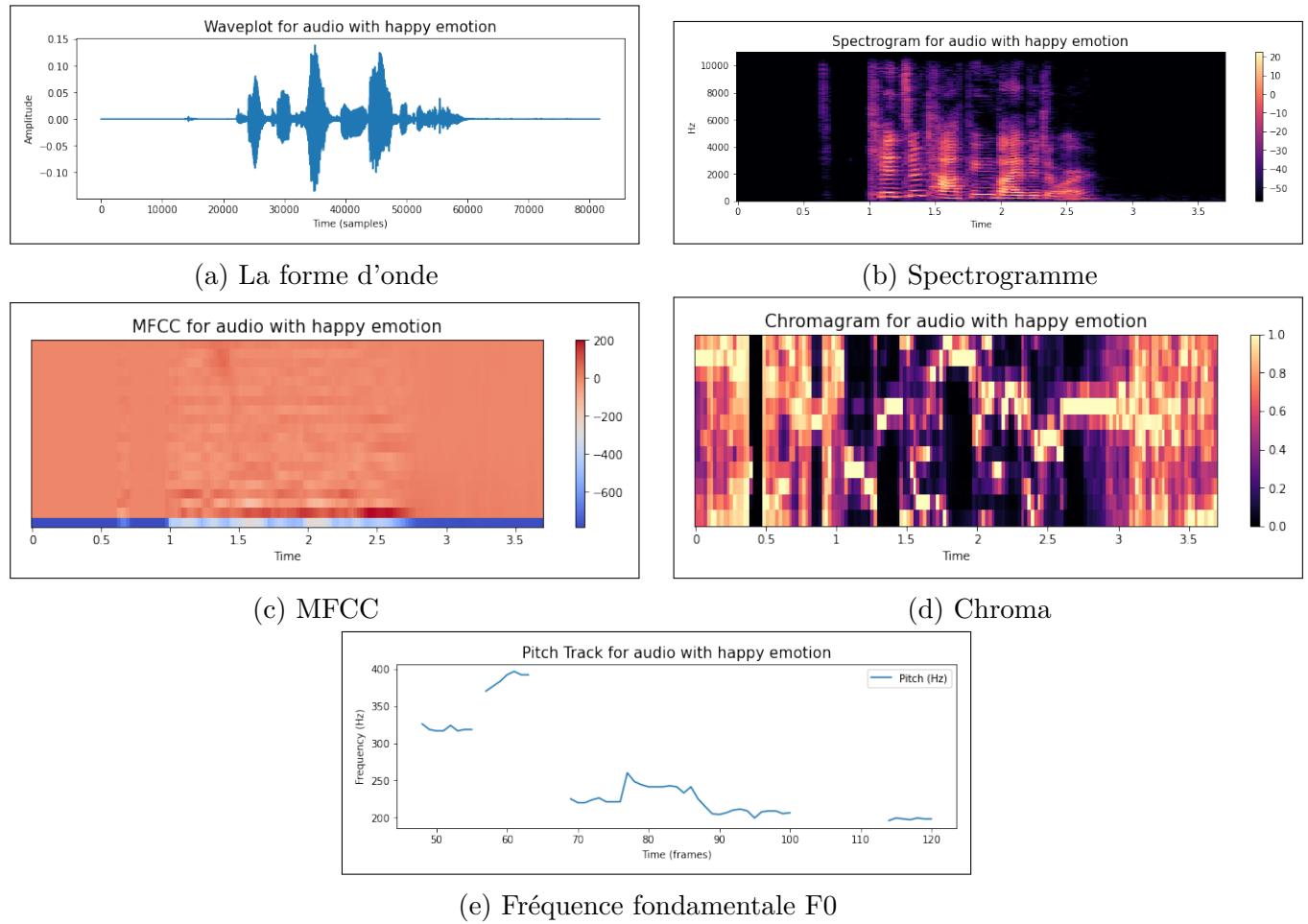


FIGURE 3.12 – Caractéristiques d'émotion "Happy"

3.4.3 Représentation des caractéristiques de l'émotion "Sad"

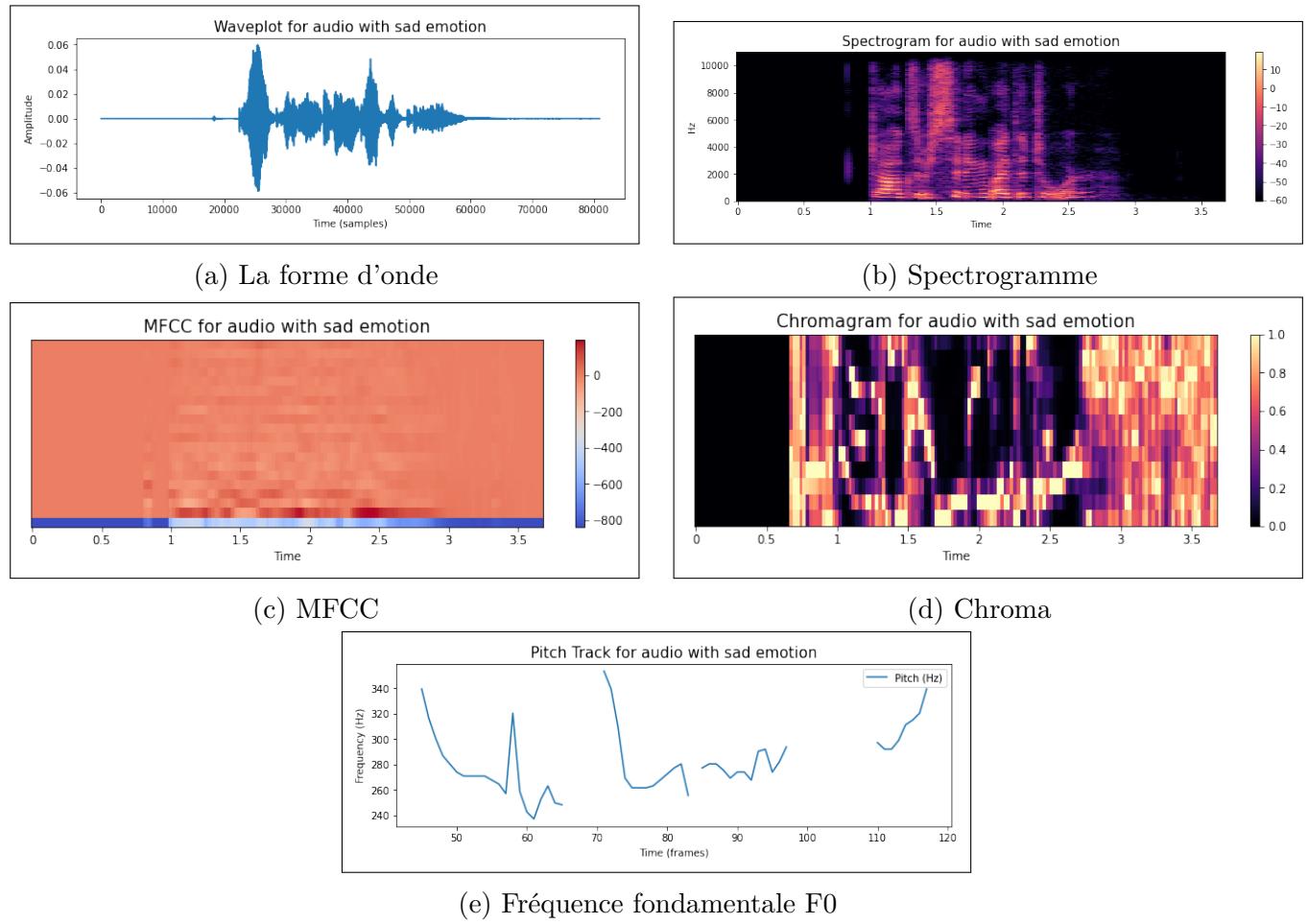


FIGURE 3.13 – Caractéristiques d'émotion "Sad"

3.4.4 Représentation des caractéristiques de l'émotion "Angry"

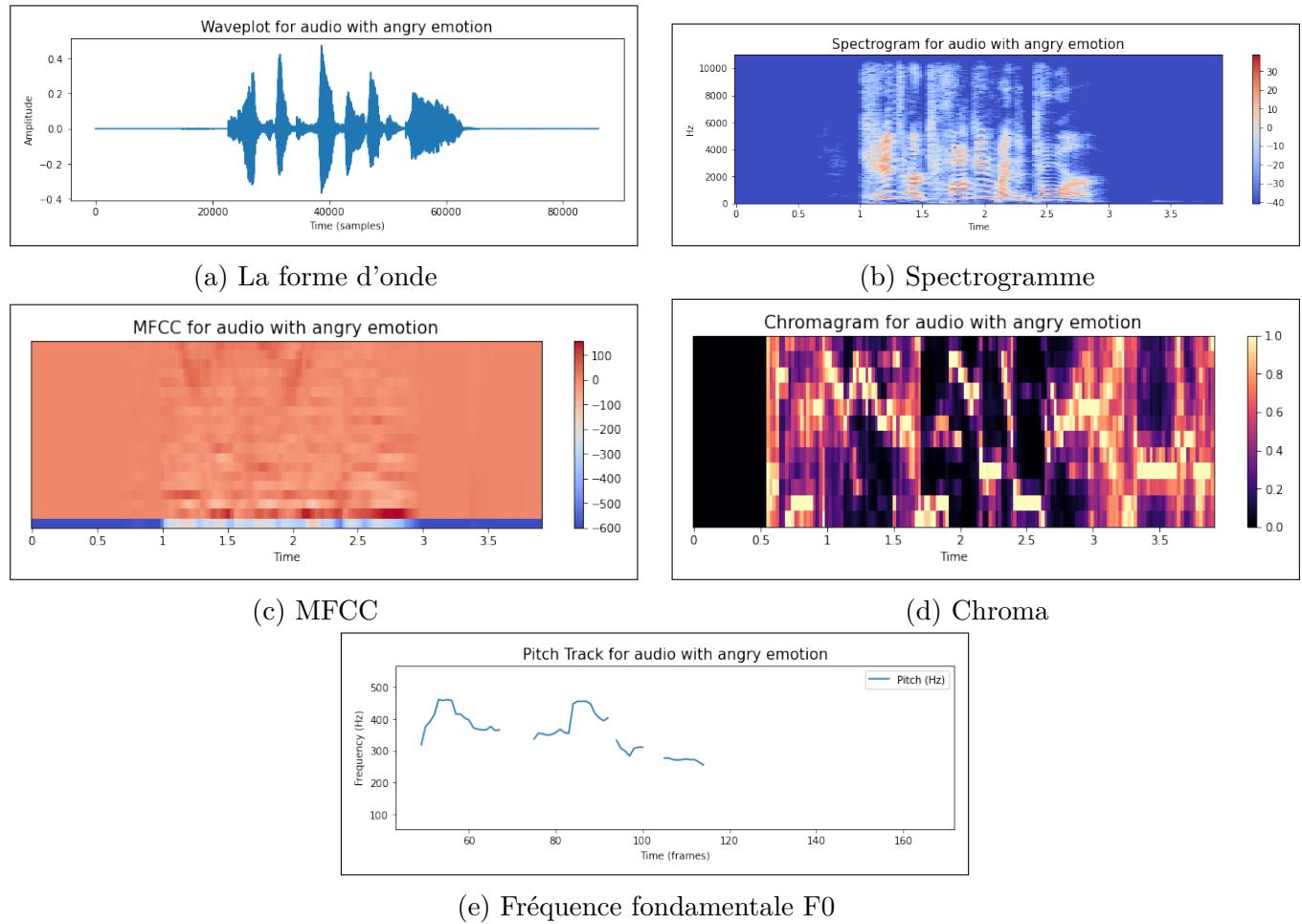


FIGURE 3.14 – Caractéristiques d'émotion "Angry"

3.4.5 Représentation des caractéristiques de l'émotion "Fear"

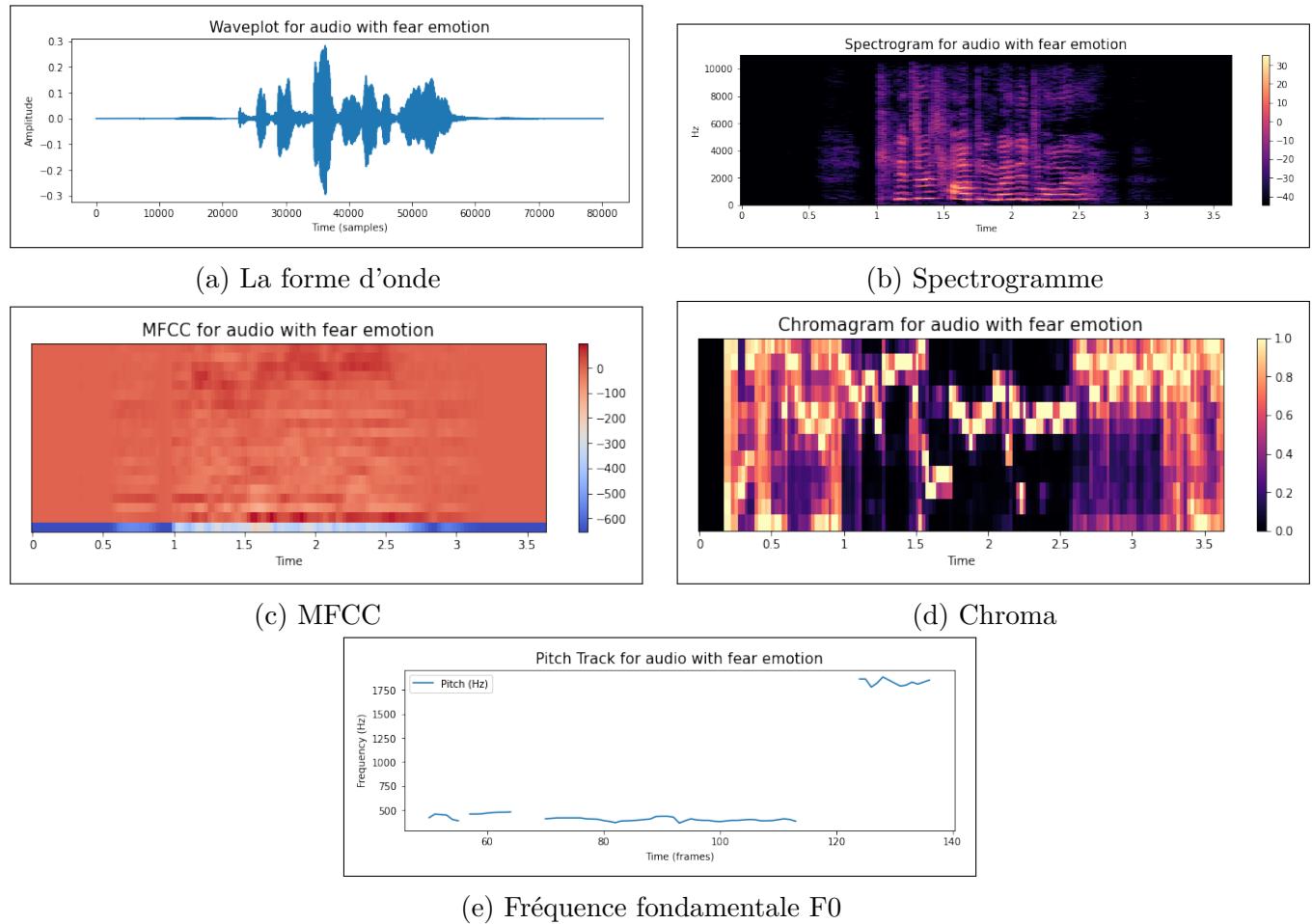


FIGURE 3.15 – Caractéristiques d'émotion "Fear"

3.4.6 Représentation des caractéristiques de l'émotion "Disgust"

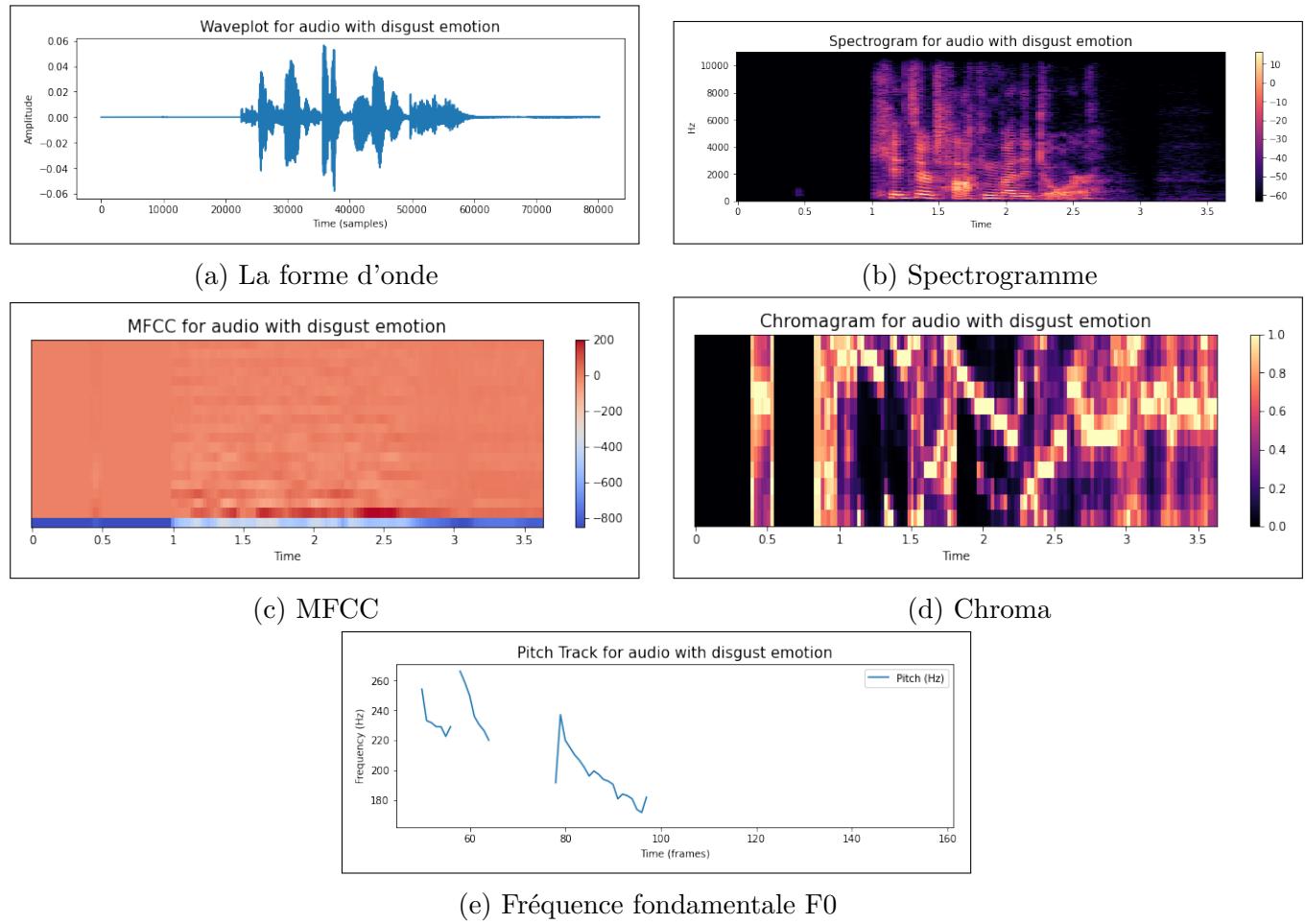


FIGURE 3.16 – Caractéristiques d'émotion "Disgust"

3.4.7 Représentation des caractéristiques de l'émotion "Surprise"

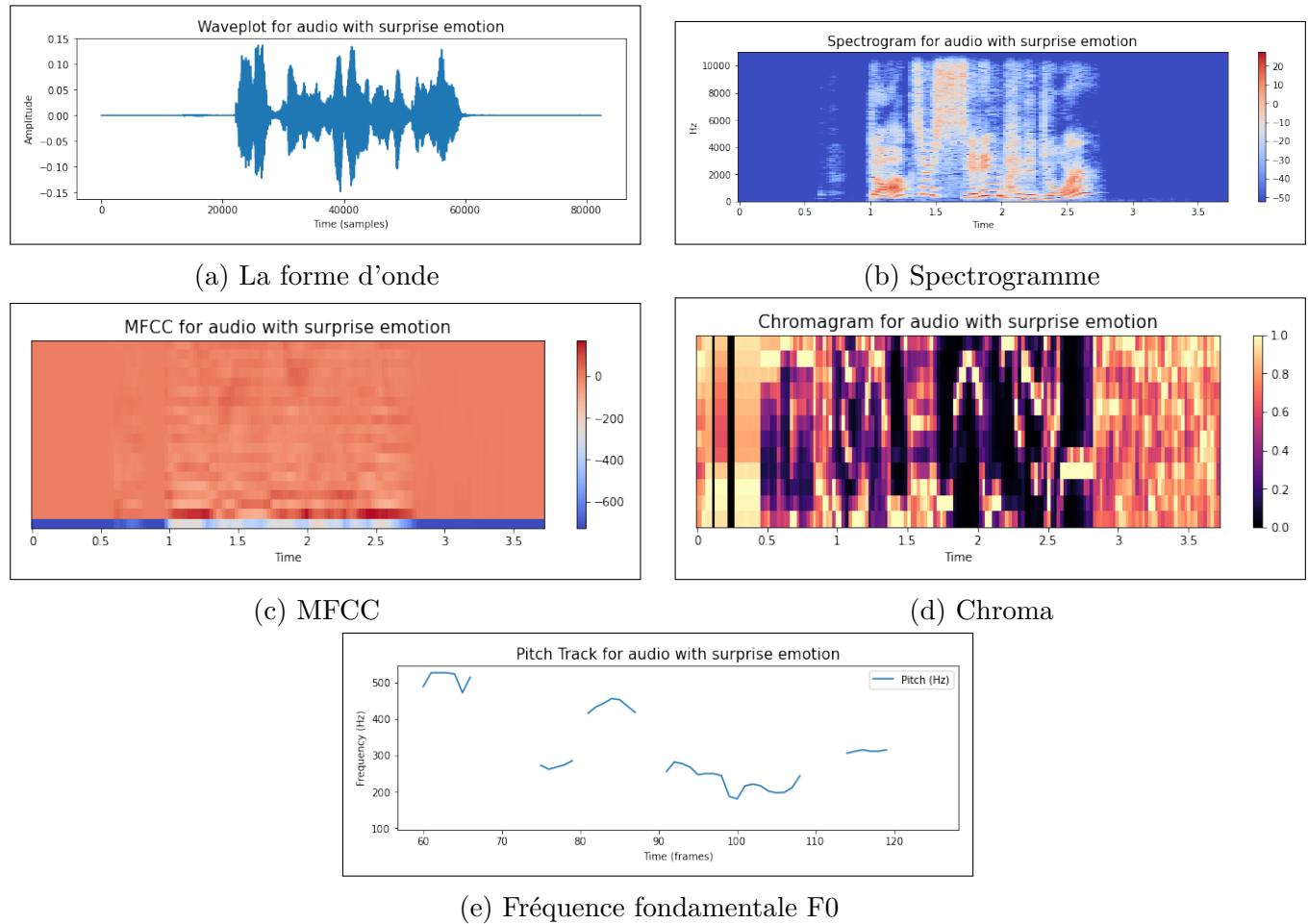


FIGURE 3.17 – Caractéristiques d'émotion "Surprise"

3.5 Augmentation de données

L'augmentation de données est une technique couramment utilisée pour enrichir les ensembles de données audio en créant de nouvelles variations des enregistrements existants, améliorant ainsi la robustesse et la performance des modèles d'apprentissage automatique. Plusieurs méthodes d'augmentation de données sont appliquées pour manipuler les enregistrements audio de manière subtile tout en préservant leurs caractéristiques émotionnelles :

- **Ajout de bruit** : Cette méthode consiste à ajouter un bruit aléatoire aux enregistrements audio existants. Le bruit est amplifié en fonction de l'amplitude maximale du signal, ce qui permet de simuler des conditions d'enregistrement réelles où le bruit de fond est présent.
- **Étirement temporel** : L'étirement temporel modifie la vitesse de lecture de l'enregistrement audio sans changer sa hauteur. Cela peut être réalisé en ralentissant ou en accélérant le signal, ce qui permet de simuler des variations naturelles dans la vitesse de la parole.
- **Décalage** : Le décalage des enregistrements consiste à décaler le signal audio de manière aléatoire sur l'axe du temps. Cette technique aide à rendre le modèle plus robuste aux variations de début et de fin d'enregistrements.
- **Modification de la hauteur** : Cette méthode change la hauteur du signal audio tout en conservant la vitesse. En ajustant la hauteur des enregistrements, on peut simuler des voix plus graves ou plus aiguës, ajoutant ainsi une diversité supplémentaire à l'ensemble de données.

Ces techniques d'augmentation permettent d'élargir l'ensemble de données d'entraînement, améliorant la capacité du modèle à généraliser et à reconnaître les émotions dans des contextes variés.

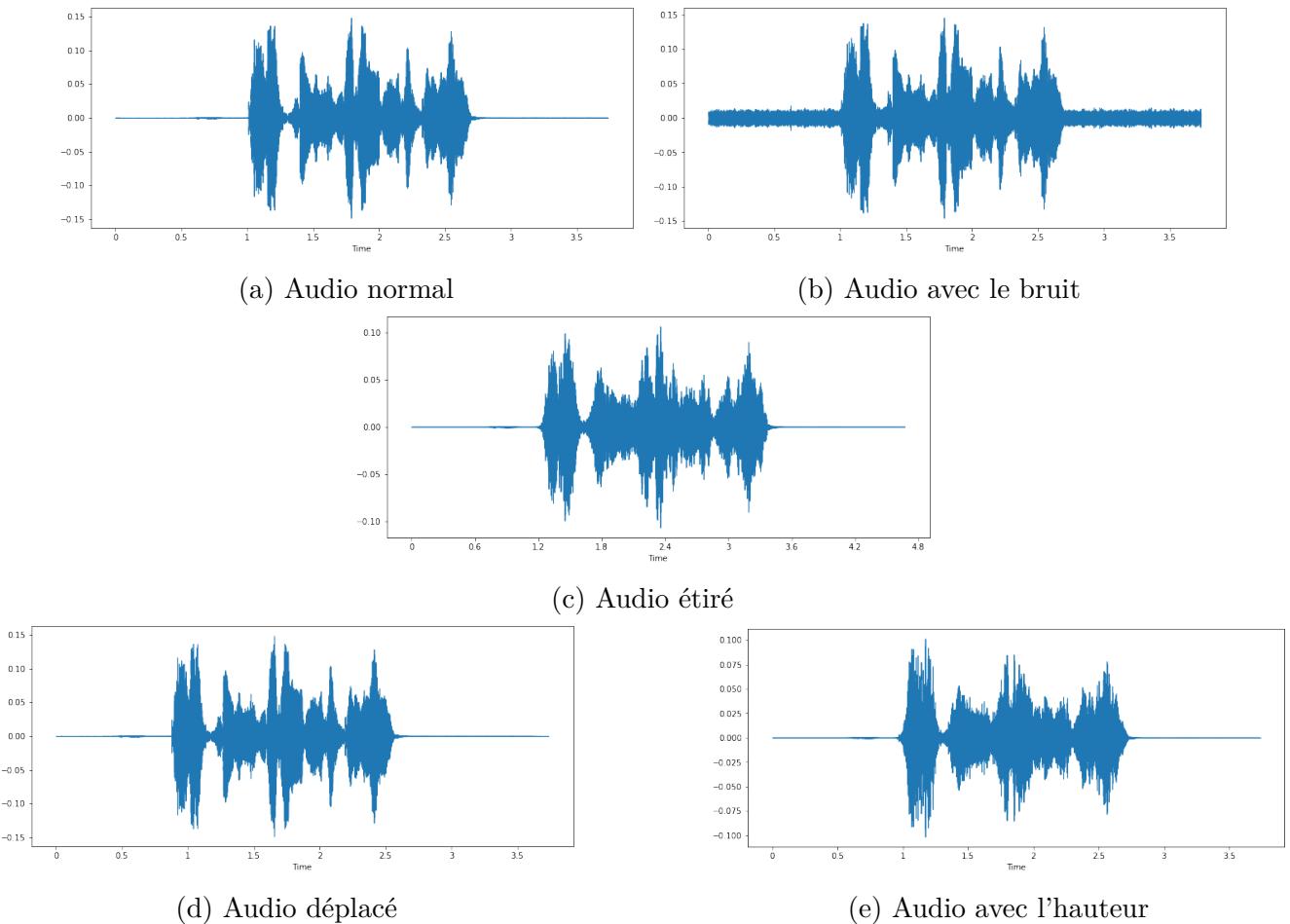


FIGURE 3.18 – Augmentation de donnée audio

3.6 Expérimentations et résultats

3.6.1 Techniques d'apprentissage automatique utilisées

3.6.1.1 Classification par KNN

- Résultat de performance

Les résultats de performance pour la classification des émotions avec KNN ($k=1$) sont présentés dans le **tableau 3.1**. Ces résultats illustrent la précision, le rappel et le score F1 pour chaque classe émotionnelle évaluée, ainsi que la précision globale du modèle.

Émotion	Précision	Rappel	F1-score	Support
Angry	0.91	0.85	0.88	1484
Disgust	0.86	0.83	0.85	1558
Fear	0.81	0.85	0.83	1505
Happy	0.86	0.83	0.84	1619
Neutral	0.85	0.86	0.86	1558
Sad	0.79	0.87	0.83	1478
Surprise	0.96	0.91	0.94	528
Accuracy : 0.85				
Macro avg : 0.86, 0.86, 0.86, 9730				
Weighted avg : 0.85, 0.85, 0.85, 9730				

TABLE 3.1 – Résultats de classification des émotions avec KNN

- Matrice de confusion

La matrice de confusion obtenue par la classification des émotions en utilisant KNN ($k=1$) est illustrée dans la **figure 3.19**.

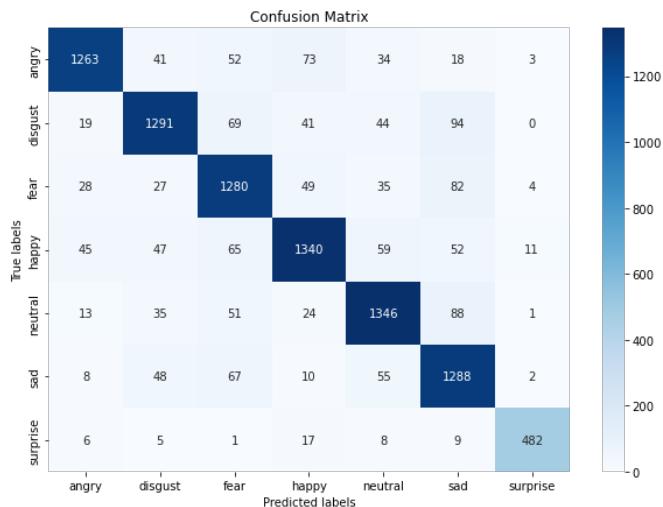


FIGURE 3.19 – Matrice de confusion résultant de classification par KNN ($k=1$) entre les classes des émotions

- Courbe ROC

La courbe ROC obtenue est illustré dans la **figure 3.20**.

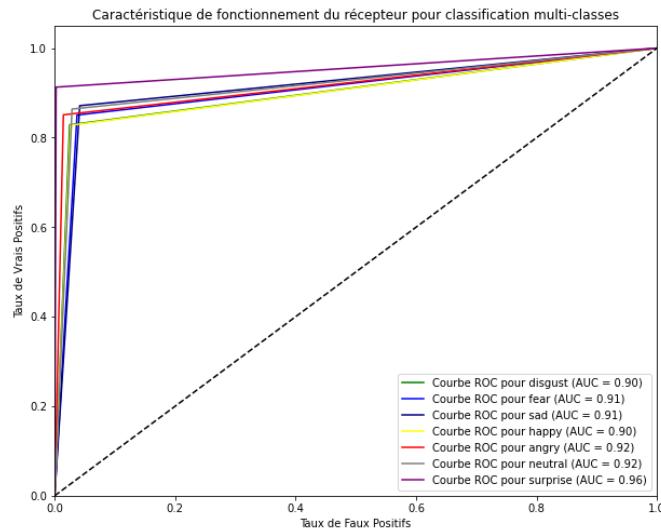


FIGURE 3.20 – Courbe ROC résultant de classification par KNN ($k=1$) entre les classes des émotions

3.6.1.2 Classification par forêt aléatoire

- Résultat de performance

La performance de classification avec forêt aléatoire pour chaque classe émotionnelle est présentée dans le **tableau 3.2**. Ces résultats montrent l'efficacité du modèle à classifier différentes émotions basées sur les métriques de précision, rappel, et score F1.

Émotion	Précision	Rappel	Score F1	Support
Angry	0.74	0.85	0.79	1484
Disgust	0.69	0.74	0.71	1558
Fear	0.86	0.62	0.72	1505
Happy	0.77	0.72	0.75	1619
Neutral	0.75	0.79	0.77	1558
Sad	0.70	0.79	0.74	1478
Surprise	0.88	0.78	0.83	528
Accuracy	0.75			
Macro Avg	0.77 Précision, 0.75 Rappel, 0.76 Score F1			
Weighted Avg	0.76 Précision, 0.75 Rappel, 0.75 Score F1			

TABLE 3.2 – Performance de classification des émotions avec forêt aléatoire

- Matrice de confusion

La matrice de confusion obtenue par la classification des émotions en utilisant forêt aléatoire est illustrée dans la **figure 3.21**.

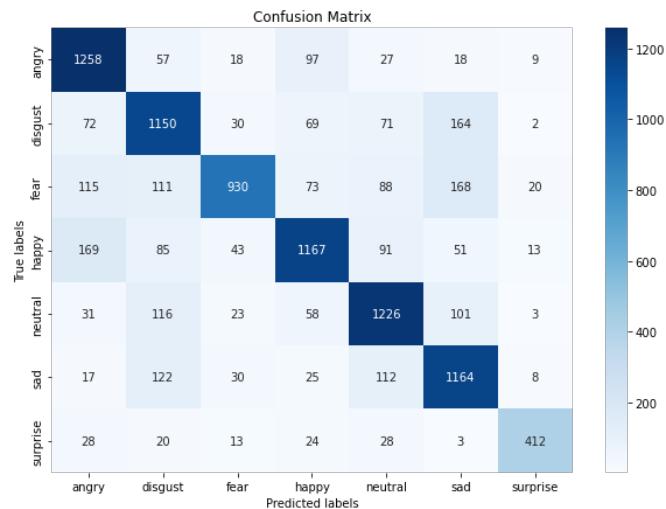


FIGURE 3.21 – Matrice de confusion résultant de classification par forêt aléatoire entre les classes des émotions

- Courbe ROC

La courbe ROC obtenue est illustré dans la **figure 3.22**.

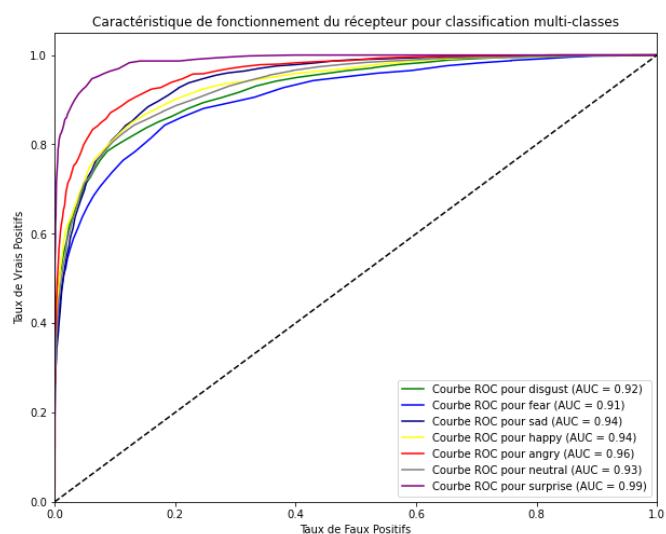


FIGURE 3.22 – Courbe ROC résultant de classification par forêt aléatoire entre les classes des émotions

3.6.2 Techniques d'apprentissage profond utilisées

3.6.2.1 Classification par ANN

- Modèle

Nous avons conçu un modèle de réseau de neurones artificiels (ANN) pour la classification, utilisant la normalisation par lots et l'activation LeakyReLU. Le **tableau 3.3** décrit l'architecture détaillée du modèle.

Couche	Neurones / Activation	Paramètres supplémentaires
Dense	256	Input shape = (x_train.shape[1],)
LeakyReLU	-	alpha = 0.01
Batch Normalization	-	-
Dropout	-	0.3
Dense	256	-
LeakyReLU	-	alpha = 0.01
Batch Normalization	-	-
Dropout	-	0.3
Dense	256	-
LeakyReLU	-	alpha = 0.01
Batch Normalization	-	-
Dropout	-	0.3
Dense	7	Activation = softmax

TABLE 3.3 – Architecture du modèle ANN avec Batch Normalization

Le modèle commence par une couche dense de 256 neurones, suivie d'une activation LeakyReLU avec un alpha de 0.01. Cette configuration est répétée trois fois avec des couches de normalisation par lots et de dropout à 0.3 pour réduire le surajustement. La dernière couche dense compte 7 neurones correspondant aux classes de sortie et utilise l'activation 'softmax' pour la classification multi-classe.

- Résultat de performance

Comme montré dans le **tableau 3.4**, les performances du modèle de réseau de neurones artificiels (ANN) sont évaluées en termes de précision, de rappel et de score F1 pour différentes classes. Le modèle a été entraîné avec un taux d'apprentissage (`learning_rate`) de 0.001, sur 200 époques (`epochs`) avec une taille de lot (`batch_size`) de 32 et une fraction de validation (`validation_split`) de 0.2. Le tableau suivant détaille les performances de classification pour chaque classe ainsi que les statistiques globales, y compris la précision globale, la moyenne macro et la moyenne pondérée des scores.

Classe	Précision	Rappel	F1-Score	Support
0	0.88	0.88	0.88	1484
1	0.84	0.76	0.80	1558
2	0.83	0.77	0.80	1505
3	0.81	0.82	0.81	1619
4	0.81	0.86	0.84	1558
5	0.77	0.85	0.81	1478
6	0.92	0.92	0.92	528
Accuracy	0.83			
Macro Avg	0.84			
Weighted Avg	0.83			

TABLE 3.4 – Performance du modèle ANN sur l'ensemble de test

- Matrice de confusion

La matrice de confusion obtenue par la classification des émotions en utilisant ANN est illustrée dans la **figure 3.23**.

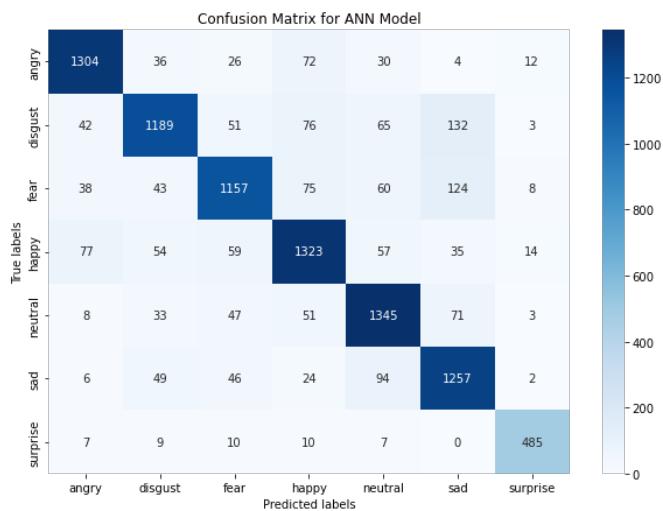


FIGURE 3.23 – Matrice de confusion résultant de classification par ANN entre les classes des émotions

- Courbe ROC

La courbe ROC obtenue est illustré dans la **figure 3.24**.

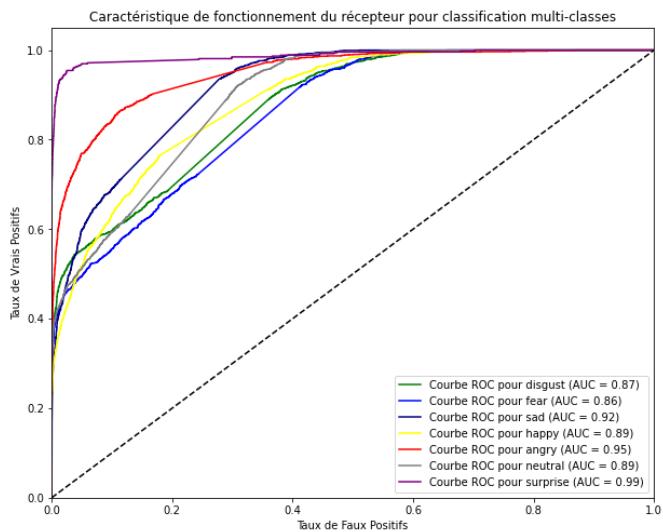


FIGURE 3.24 – Courbe ROC résultant de classification par ANN entre les classes des émotions

3.6.2.2 Classification par CNN

- Modèle

Le **tableau 3.5** présente un modèle de réseau de neurones convolutif (CNN) utilisé pour une tâche de classification. Ce modèle est de type séquentiel, signifiant que chaque couche est empilée l'une sur l'autre dans un ordre précis sans boucles ou connexions de saut. Les composants clés du modèle incluent :

- **Conv1D** : Deux couches de convolution sont employées, avec la première disposant de 256 filtres et la seconde de 128. Ces couches sont conçues pour extraire des caractéristiques spatiales à partir de l'entrée.
- **Batch Normalization** : Chaque couche Conv1D est suivie d'une couche de normalisation par lots, qui accélère la convergence du modèle en normalisant les activations.
- **MaxPooling1D** : Les couches de pooling maximal sont utilisées après chaque couche de convolution pour réduire les dimensions spatiales des caractéristiques entrantes, aidant ainsi à rendre le modèle plus robuste aux variations de position des caractéristiques dans l'entrée.

- **Dropout** : Une couche de dropout est implémentée pour réduire le risque de surajustement en désactivant aléatoirement un certain pourcentage de neurones lors de l’entraînement.
- **Flatten** : Cette couche transforme les données multi-dimensionnelles en un vecteur unidimensionnel pour permettre le traitement par les couches denses suivantes.
- **Dense** : Deux couches denses sont utilisées, avec la dernière couche dense ayant 7 unités, indiquant que le modèle effectue une classification dans 7 classes distinctes.

Le modèle a un total de 19,568,775 paramètres, dont 19,567,495 sont entraînables, signifiant qu’ils sont ajustés au cours du processus d’apprentissage, et 1,280 paramètres qui restent non entraînables. Ce modèle peut être utilisé pour des tâches complexes de classification où la compréhension des caractéristiques spatiales est cruciale, telles que la classification de séquences de temps ou de textes lorsque représentées sous une forme appropriée.

Model : "sequential"		
Layer (type)	Output Shape	Param #
conv1d (Conv1D)	(None, 2376, 256)	1536
batch_normalization (BatchNo	(None, 2376, 256)	1024
max_pooling1d (MaxPooling1D)	(None, 1188, 256)	0
conv1d_1 (Conv1D)	(None, 1188, 128)	98432
batch_normalization_1 (Batch	(None, 1188, 128)	512
max_pooling1d_1 (MaxPooling1	(None, 594, 128)	0
dropout (Dropout)	(None, 594, 128)	0
flatten (Flatten)	(None, 76032)	0
dense (Dense)	(None, 256)	19464448
batch_normalization_2 (Batch	(None, 256)	1024
dense_1 (Dense)	(None, 7)	1799
Total params : 19,568,775		
Trainable params : 19,567,495		
Non-trainable params : 1,280		

TABLE 3.5 – Architecture du modèle CNN

- Résultat de performance

Les performances du modèle CNN pour la classification des émotions sont résumées dans le **tableau 3.6**. Ce modèle montre une haute précision et une bonne capacité à généraliser, comme indiqué par les scores de rappel, de précision et F1 pour chaque classe d'émotions. Le modèle a été entraîné sur 30 époques (`epochs`) avec une taille de lot (`batch_size`) de 64.

Émotion	Précision	Rappel	F1-score	Support
Angry	0.93	0.92	0.92	1484
Disgust	0.92	0.87	0.89	1558
Fear	0.91	0.89	0.90	1505
Happy	0.91	0.88	0.90	1619
Neutral	0.89	0.94	0.92	1558
Sad	0.88	0.94	0.91	1478
Surprise	0.94	0.95	0.94	528
Global Performance Metrics				
Accuracy			0.91	
Macro Avg	0.91	0.91	0.91	9730
Weighted Avg	0.91	0.91	0.91	9730

TABLE 3.6 – Performance du modèle CNN sur la classification des émotions

- Courbes d'erreur et de précision obtenues

Dans le but de visualiser les résultats, nous présentons pour chaque classification les courbes d'erreur et de précision obtenues, ainsi que la matrice de confusion et courbe ROC.

La **figure 3.25** présente la courbe de précision obtenue pour la classification entre les classes des émotions. La courbe représentant l'erreur durant l'apprentissage est représentée par la **figure 3.25**. Après l'apprentissage, les poids du modèle de la meilleure epoch sont restaurés. Nous avons mesuré la précision de notre modèle, en utilisant la base de données de test. La précision obtenue est de 90.9%.

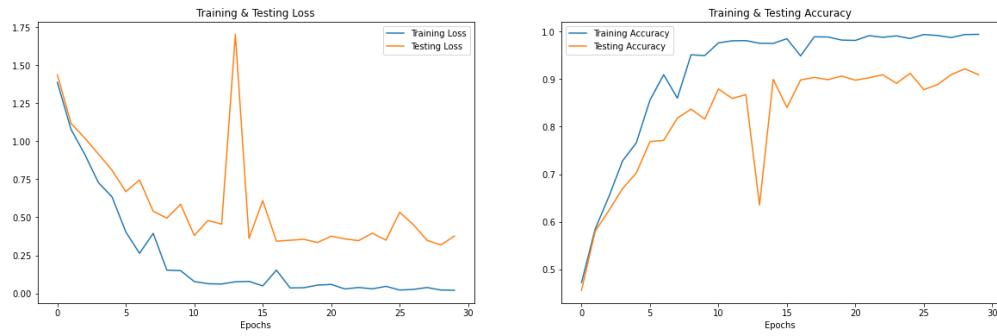


FIGURE 3.25 – Courbe de précision & de perte

- Matrice de confusion

La matrice de confusion obtenue par la classification des émotions en utilisant CNN est illustrée dans la **figure 3.26**.

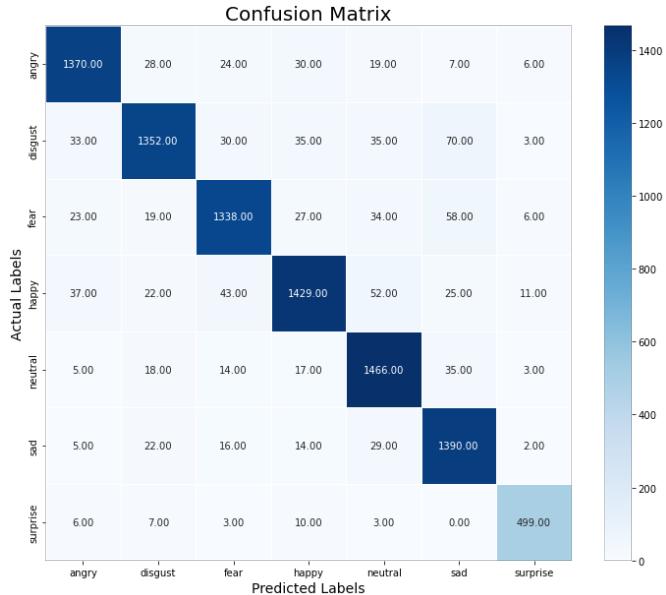


FIGURE 3.26 – Matrice de confusion résultant de classification par CNN entre les classes des émotions

- Courbe ROC

La courbe ROC obtenue est illustré dans la **figure 3.27**.

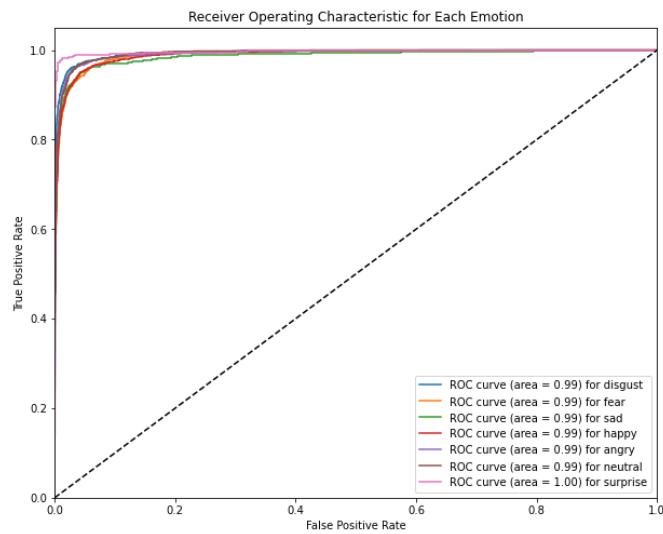


FIGURE 3.27 – Courbe ROC résultant de classification par CNN entre les classes des émotions

- Comparaison des étiquettes prédites et réelles

Cette **figure 3.28** illustre un extrait des résultats obtenus par un modèle de classification des émotions. Il montre les étiquettes prédites par le modèle en comparaison avec les étiquettes réelles, permettant d'évaluer visuellement la précision des prédictions pour diverses émotions.

	Predicted Labels	Actual Labels
0	angry	angry
1	angry	angry
2	disgust	disgust
3	happy	happy
4	fear	fear
5	happy	happy
6	happy	happy
7	fear	fear
8	fear	fear
9	surprise	surprise
10	angry	angry
11	fear	fear
12	neutral	neutral
13	sad	sad
14	disgust	disgust
15	happy	happy
16	disgust	disgust
17	sad	sad
18	fear	fear
19	disgust	disgust

FIGURE 3.28 – Comparaison des étiquettes prédites et réelles pour un modèle de classification des émotions avec CNN

Les performances des modèles de classification des émotions varient de manière significative selon l’algorithme utilisé. Chaque approche, qu’il s’agisse de KNN, de forêt aléatoire, d’ANN ou de CNN, présente des avantages et des inconvénients spécifiques. Notamment, les modèles CNN se distinguent par leur capacité à atteindre une précision supérieure, comme le montrent les métriques de performance et les courbes ROC. En raison de ses meilleures performances par rapport aux autres algorithmes, le modèle CNN a été intégré dans l’interface graphique de l’application. Pour plus de détails sur les performances des algorithmes, voir le **tableau 3.7** et **tableau 3.8**.

TABLE 3.7 – Comparaison des performances des algorithmes de classification des émotions

Algorithme	Précision globale	Rappel global	F1-score global	Accuracy global
KNN	0.86	0.86	0.86	0.85
Forêt aléatoire	0.76	0.75	0.75	0.75
ANN	0.84	0.84	0.84	0.83
CNN	0.91	0.91	0.91	0.91

TABLE 3.8 – Comparaison des valeurs ROC des algorithmes de classification des émotions

Algorithm	ROC global
KNN	0.92
Forêt aléatoire	0.94
ANN	0.91
CNN	0.99

3.7 Interfaces utilisateurs pour une application de détection des émotions par la voix

3.7.1 Page d'accueil avant authentification

La **figure 3.29** présente la page d'accueil de notre site web de détection des émotions par la voix. Cette interface conviviale est conçue pour accueillir les utilisateurs et leur offrir une introduction claire aux fonctionnalités de la plateforme. En haut de la page, un menu de navigation permet aux utilisateurs de se connecter ou de s'inscrire facilement. Le corps principal de la page d'accueil comporte un message de bienvenue mettant en avant les capacités de notre technologie avancée de reconnaissance des émotions vocales. Deux sections distinctes invitent les nouveaux utilisateurs à créer un compte pour accéder à des analyses personnalisées et permettent aux membres existants de se connecter pour explorer leurs résultats précédents. L'interface est simple et efficace, visant à offrir une expérience utilisateur fluide et intuitive. Cette conception reflète notre engagement à rendre la technologie de détection des émotions accessible et utile pour tous.

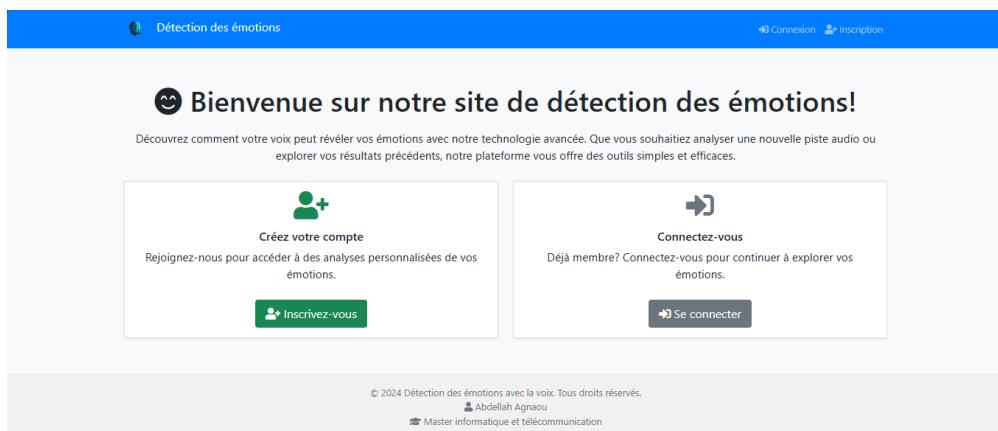


FIGURE 3.29 – Page d'accueil avant authentification

3.7.1.1 Page pour inscription

La **figure 3.30** illustre la page d’inscription de notre plateforme de détection des émotions par la voix. Cette interface épurée permet aux nouveaux utilisateurs de créer facilement un compte en fournissant un nom d’utilisateur, une adresse email et un mot de passe. Le design est simple et fonctionnel, avec des champs de saisie clairement étiquetés et des boutons d’action distincts pour s’inscrire ou retourner à la page d’accueil. La navigation en haut de la page reste accessible, permettant aux utilisateurs de passer rapidement de la page de connexion à celle d’inscription. En bas de la page, une note de copyright et des informations de contact rappellent les droits réservés et l’identité du développeur. Cette conception vise à offrir une expérience utilisateur fluide et intuitive dès le premier contact avec notre service.

The screenshot shows a registration form titled "Inscription". At the top, there's a blue header bar with the text "Détection des émotions" and links for "Connexion" and "Inscription". Below the header, the main form has three input fields: "Nom d'utilisateur", "Email", and "Mot de passe". Each field has a small icon indicating its purpose (user icon for name, envelope for email, lock for password). Below the fields are two buttons: "S'inscrire" (in blue) and "Retour à l'accueil" (in grey). At the bottom of the page, there's a footer with copyright information: "© 2024 Détection des émotions avec la voix. Tous droits réservés.", "Abdellah Agraou", and "Master informatique et télécommunication".

FIGURE 3.30 – Page pour inscription

3.7.1.2 Page pour connexion

La **figure 3.31** montre la page de connexion de notre site web de détection des émotions par la voix. Cette page permet aux utilisateurs enregistrés d'accéder à leur compte en entrant leur nom d'utilisateur et leur mot de passe. L'interface est minimaliste et fonctionnelle, offrant une expérience utilisateur intuitive avec des champs de saisie bien étiquetés et des boutons d'action clairs pour se connecter ou retourner à la page d'accueil. La navigation en haut de la page reste accessible, facilitant la transition entre les pages de connexion et d'inscription. Un message de bienvenue encourage les utilisateurs à explorer la puissance de notre technologie de détection des émotions vocales. En bas de la page, les informations de copyright et de contact assurent les utilisateurs de la protection de leurs données et de la disponibilité d'assistance en cas de besoin. Cette conception vise à garantir une expérience de connexion fluide et sécurisée.

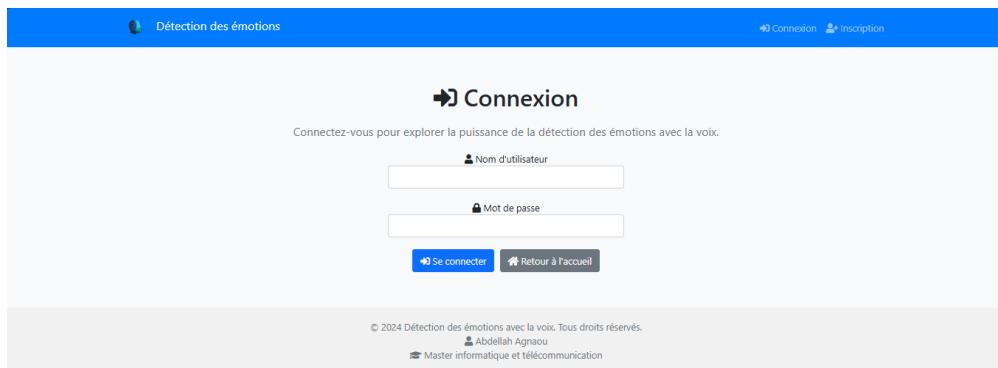


FIGURE 3.31 – Page pour connexion

3.7.2 Page d'accueil après authentification

La **figure 3.32** présente la page d'accueil après la connexion de l'utilisateur sur notre site web de détection des émotions par la voix. Une fois connecté, l'utilisateur est accueilli par un message personnalisé, ce qui renforce le sentiment d'inclusion et d'engagement. Trois options principales sont offertes : accéder aux analyses et statistiques, enregistrer un nouvel audio pour une analyse en temps réel, ou importer un fichier audio existant pour détecter les émotions. Chaque option est clairement présentée avec des icônes et des boutons d'action distincts pour faciliter la navigation. En haut de la page, une option de déconnexion est facilement accessible, assurant une expérience utilisateur fluide et sécurisée. Les informations de copyright et de contact sont affichées en bas de la page, garantissant la transparence et la disponibilité du support.

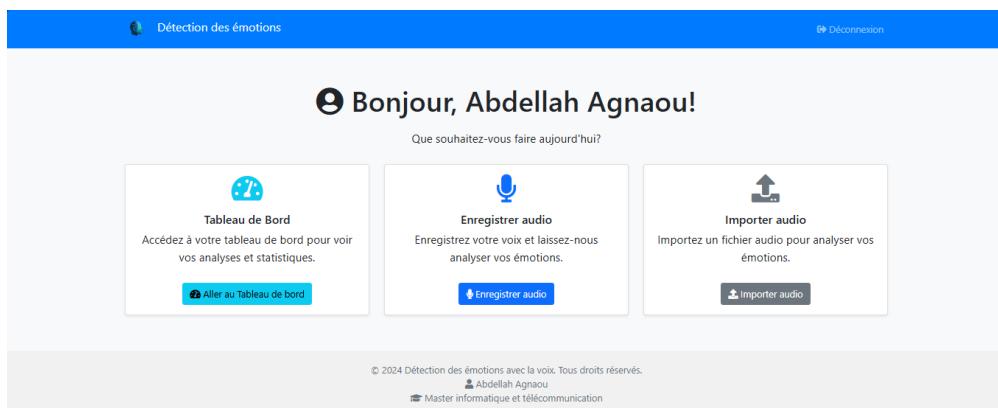


FIGURE 3.32 – Page d'accueil après authentification

3.7.3 Page pour enregistrer un audio

La **figure 3.33** illustre la page d'enregistrement audio de notre site web de détection des émotions par la voix. Cette interface permet aux utilisateurs d'enregistrer leur voix pour analyser les émotions détectées. Cinq options principales sont disponibles : commencer l'enregistrement, arrêter l'enregistrement, enregistrer l'audio sur le serveur, réinitialiser l'enregistrement, et retourner à la page d'accueil. Chaque option est représentée par un bouton clairement identifiable et accompagné d'instructions succinctes pour guider l'utilisateur à travers le processus. La disposition intuitive et les instructions claires visent à simplifier l'expérience utilisateur, garantissant une interaction fluide et efficace avec la plateforme. En haut de la page, les options de navigation permettent aux utilisateurs d'accéder facilement au tableau de bord ou de se déconnecter. Cette conception assure une utilisation sans effort de la fonctionnalité d'enregistrement, essentielle pour l'analyse précise des émotions vocales.

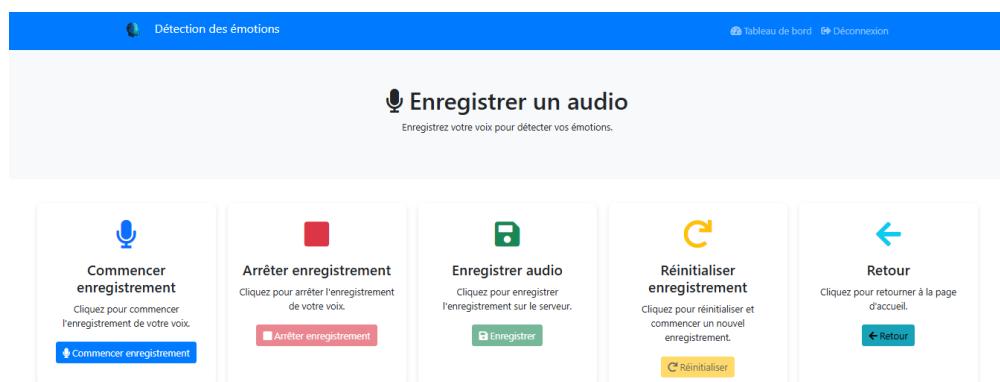


FIGURE 3.33 – Page d'enregistrement d'un audio

3.7.4 Page pour importer un audio

La **figure 3.34** présente la page d'importation audio de notre site web de détection des émotions par la voix. Cette interface permet aux utilisateurs de télécharger un fichier audio pour analyser les émotions détectées dans leur enregistrement. La page est conçue de manière intuitive avec un champ pour sélectionner le fichier audio à importer, suivi de trois boutons d'action distincts : Retour, Importer audio, et Réinitialiser. Le bouton "Retour" permet de revenir à la page précédente, le bouton "Importer audio" lance le processus de téléchargement et d'analyse du fichier, tandis que le bouton "Réinitialiser" permet de réinitialiser la sélection pour commencer un nouvel import. En haut de la page, les utilisateurs peuvent accéder

au tableau de bord ou se déconnecter via les options de navigation. Les informations de copyright et de contact sont affichées en bas de la page, offrant une transparence et une assistance en cas de besoin. Cette conception assure une utilisation facile et efficace de la fonctionnalité d'importation audio, essentielle pour l'analyse des émotions vocales.

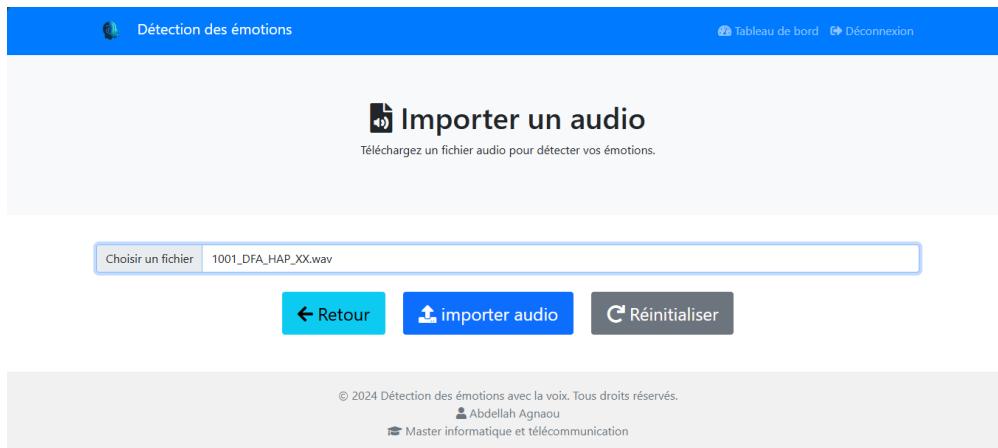


FIGURE 3.34 – Page d'importation d'un audio

3.7.5 Affichage de résultat

La **figure 3.35** montre la page de résultats d'analyse de notre site web de détection des émotions par la voix. Cette interface affiche les résultats de l'analyse émotionnelle de l'enregistrement vocal de l'utilisateur. En haut de la page, l'émotion détectée est présentée de manière claire avec un emoji correspondant et le nom de l'émotion détectée. Trois visualisations principales sont fournies pour une analyse détaillée : la forme d'onde, le spectrogramme, et les coefficients cepstraux en fréquences de Mel (MFCC). Ces visualisations permettent aux utilisateurs de voir comment les caractéristiques de leur voix correspondent à l'émotion détectée. Un bouton "Retour" en bas de la page permet de revenir à l'interface précédente. Les informations de copyright et de contact sont également affichées en bas de la page, garantissant la transparence et la disponibilité du support en cas de besoin. Cette conception assure une interprétation claire et complète des résultats d'analyse émotionnelle, offrant une expérience utilisateur informative et enrichissante.

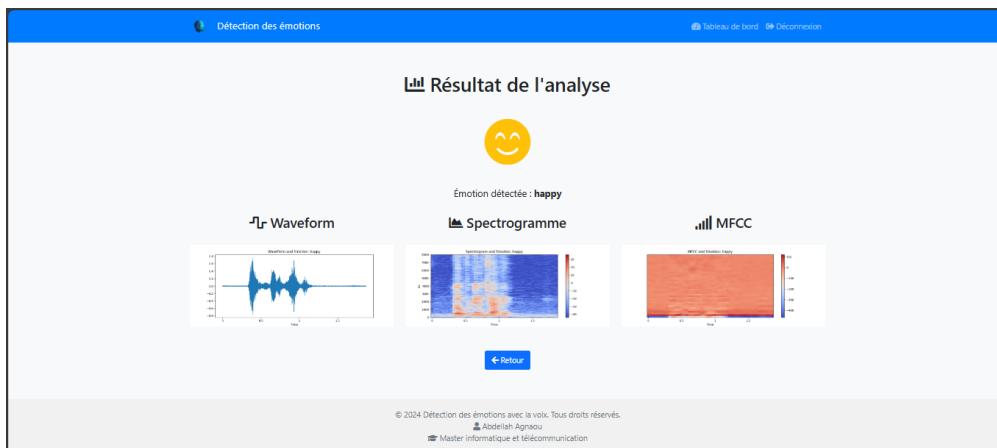


FIGURE 3.35 – Affichage de résultat

Conclusion

Ce chapitre a fourni une vue d'ensemble complète des technologies, outils, méthodes et résultats expérimentaux obtenus dans le cadre du développement de notre application de détection des émotions par la voix. En combinant **Bootstrap** pour le front-end et **Flask** pour le back-end, nous avons développé une application web intuitive avec des interfaces utilisateur adaptées pour l'enregistrement et l'importation d'audio, ainsi que pour l'affichage des résultats de détection des émotions.

Les bibliothèques Python telles que **TensorFlow** et **Librosa** ont joué un rôle central dans la création des modèles de machine learning et dans l'analyse des fichiers audio. Les bases de données CREMA-D, RAVDESS, SAVEE, et TESS ont été fondamentales pour l'entraînement et l'évaluation des modèles, et les étapes de prétraitement ont permis d'améliorer la qualité des données. L'extraction des caractéristiques audio, notamment les MFCCs, RMSE et ZCR, a été cruciale pour le développement de modèles robustes.

Après avoir évalué plusieurs techniques d'apprentissage automatique et d'apprentissage profond, nous avons conclu que le modèle CNN surpassait les autres algorithmes en termes de précision globale. C'est pourquoi le modèle CNN a été choisi pour être intégré dans l'interface graphique de l'application, permettant ainsi une détection des émotions par la voix plus efficace et précise.

CONCLUSION GÉNÉRALE ET PERSPECTIVE

La reconnaissance automatique des émotions représente un domaine en plein essor, marqué par de multiples défis liés aux variations inter et intra-individuelles des caractéristiques émotionnelles. Les signaux émotionnels, captés par la voix et les expressions faciales, sont souvent perturbés par le bruit de fond et les conditions environnementales, ce qui complique la tâche des systèmes de reconnaissance. Grâce à l'intégration de plusieurs modalités et à l'usage avancé de techniques d'apprentissage automatique et profond, ce rapport a exploré diverses approches pour améliorer la précision et la robustesse de ces systèmes.

Les résultats obtenus montrent clairement l'importance d'une approche multimodale et sophistiquée pour la reconnaissance des émotions. Les différentes étapes de traitement, depuis la présélection des données vocales jusqu'à la classification via des algorithmes avancés, constituent une chaîne complexe qui doit être optimisée à chaque niveau pour obtenir des performances précises. La variabilité des expressions émotionnelles demeure un obstacle majeur. Les modulations vocales dues aux différents états émotionnels ou aux variations individuelles des locuteurs requièrent des systèmes adaptatifs capables d'apprendre ces variations de manière dynamique. Par ailleurs, les interférences dues au bruit de fond et aux conditions environnementales soulignent la nécessité de techniques robustes de prétraitement et de nettoyage des données vocales. Des méthodes telles que les filtres adaptatifs ou l'apprentissage de représentations robustes peuvent être envisagées pour améliorer cette dimension. De plus, la combinaison de plusieurs modalités (voix, expressions faciales, gestes, etc.) pour la reconnaissance des émotions peut conduire à des résultats plus précis et plus fiables. Cependant, l'intégration de ces données hétérogènes nécessite des modèles sophistiqués et une gestion efficace des informations multi-sources.

Pour les recherches futures, plusieurs pistes peuvent être envisagées. Le développement d'algorithmes d'apprentissage plus performants et mieux adaptés à la reconnaissance des émotions peut permettre de surmonter les défis actuels. Les réseaux de neurones plus profonds, les modèles d'apprentissage par transfert et les techniques d'apprentissage par renforcement posent des avenues prometteuses. Constituer et exploiter des bases de données plus riches et diversifiées en termes de langues, de cultures et de contextes socio-économiques permettra d'entraîner des modèles plus généralisables et moins biaisés. Explorer des techniques innovantes pour la fusion des différentes modalités d'entrée peut aider à créer des modèles plus robustes capables de mieux comprendre et interpréter les états émotionnels. Enfin, l'application de ces avancées technologiques dans des contextes réels, tels que la santé mentale, les call centers, ou les interactions homme-machine, demeure une perspective fascinante pour évaluer l'utilité pratique et l'impact de ces systèmes. En conclusion, bien que des challenges importants subsistent, les progrès continus dans le domaine de la reconnaissance des émotions promettent des innovations significatives, ouvrant ainsi la voie à des interactions homme-machine plus naturelles et empathiques.

BIBLIOGRAPHIE ET WEBOGRAPHIE

- [1] “The Voice Mechanism - THE VOICE FOUNDATION”. In : (). Philadelphia, New York, Los Angeles, Cleveland, Boston, Paris, Lebanon, Brazil, China, Japan, India, Mexico. URL : <https://voicefoundation.org/health-science/voice-disorders/anatomy-physiology-of-voice-production/the-voice-mechanism/>.
- [2] “Cordes vocales : anatomie, pathologies, traitements”. In : (). URL : <https://www.passeportsante.net/fr/parties-corps/Fiche.aspx?doc=cordes-vocales>.
- [3] “Capsule outil : La voix et l'appareil de phonation”. In : (). URL : https://lecerveau.mcgill.ca/flash/capsules/outil_bleu21.html.
- [4] “LE SYSTEME PHONOLOGIQUE DU FRANÇAIS”. In : (). URL : https://mortain.circonscription.ac-normandie.fr/IMG/pdf/2._systeme_francais.pdf.
- [5] “Comment fonctionne la voix ?” In : (). Anatomie et physiologie de la voix, 26 mai. URL : <https://stephaniedumouch.com/articles/comment-fonctionne-la-voix>.
- [6] “Les troubles de la voix : les signes et les causes”. In : (). Service d'ORL - HUG, 2021.
- [7] Par Martine Lochouarn et SERVICE INFOGRAPHIE. “Les troubles de la voix mieux pris en charge”. In : (). Publié le 12/02/2016 à 17 :00. URL : <https://sante.lefigaro.fr/actualite/2016/02/12/24618-troubles-voix-mieux-pris-charge>.
- [8] “Dimension esthétique des voix normales et dysphoniques : Approches perceptive et acoustique”. In : ().
- [9] “Bilan clinique de la voix”. In : (). Service d'ORL - HUG, 2021.
- [10] “EVALUATION CLINIQUE DES SIGNAUX ACOUSTIQUES VOCaux”. In : (). Phona-nium, 2019.
- [11] TONES AND VIBES. *Fréquence d'échantillonnage et débit binaire*. URL : <https://tonesandvibes.com/fr/frequence-echantillonnage-debit-binaire/>.
- [12] RESEARCHGATE. *Spectre lisse du signal brut et du bruit de fond*. URL : https://www.researchgate.net/figure/spectre-lisse-du-signal-brut-et-du-bruit-de-fond-Il-est-maintenant-possible-a-laide_fig4_47723644.

- [13] RESEARCHGATE. *Segmentation automatique de l'enregistrement audio en environnement bruyant de cafétéria*. URL : https://www.researchgate.net/figure/Automatic-segmentation-of-the-audio-recorded-in-the-cafeteria-noisy-environment-by_fig2_323155847.
- [14] PYTORCH. *Audio Data Augmentation Tutorial*. URL : https://pytorch.org/audio/2.0.1/tutorials/audio_data_augmentation_tutorial.html.
- [15] AMPED STUDIO. *Normalisation audio*. URL : <https://ampedstudio.com/fr/normalisation-audio/>.
- [16] HUGGING FACE. *Audio Data - Hugging Face Audio Course*. URL : https://huggingface.co/learn/audio-course/fr/chapter1/audio_data.
- [17] Allen LU. *Audio Chroma Feature Vector*. Disponible sur le blog de Allen Lu, consulté le 7 juin 2024. 2018. URL : <https://allenlu2007.wordpress.com/2018/07/04/audio-chroma-feature-vector/>.
- [18] MFCC Feature Extraction Block Diagram. Disponible sur ResearchGate, consulté le 7 juin 2024. 2023. URL : <https://www.researchgate.net/publication/50283099>.
- [19] Alan V. OPPENHEIM et Ronald W. SCHAFER. *Discrete-Time Signal Processing*. Prentice Hall, 1999, p. 101-130. ISBN : 978-0-13-754920-7. URL : <https://www.pearson.com/us/higher-education/program/Oppenheim-Discrete-Time-Signal-Processing-3rd-Edition/PGM334830.html>.
- [20] MEDIUM - GREAT LEARNING. *RMSE : What Does It Mean ?* URL : <https://medium.com/@mygreatlearning/rmse-what-does-it-mean-2d446c0b1d0e>.
- [21] Richard G. LYONS. *Understanding Digital Signal Processing*. 3^e éd. Prentice Hall, 2011, p. 275-300. ISBN : 978-0137027415. URL : <https://www.pearson.com/store/p/understanding-digital-signal-processing/P100000648257>.
- [22] RESEARCHGATE. *Definition of Zero Crossings Rate*. URL : https://www.researchgate.net/figure/Definition-of-zero-crossings-rate_fig2_259823741.
- [23] Kevin P. MURPHY. *Machine Learning : A Probabilistic Perspective*. MIT Press, 2012. ISBN : 0262018020, 9780262018029.
- [24] Gareth JAMES et al. *An Introduction to Statistical Learning*. Springer, 2013. ISBN : 9781461471370.
- [25] Christopher M. BISHOP. *Pattern Recognition and Machine Learning*. Springer, 2006. ISBN : 0387310738, 9780387310732.
- [26] Trevor HASTIE, Robert TIBSHIRANI et Jerome FRIEDMAN. *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. Springer, 2009. ISBN : 9780387848570.
- [27] Ian T. JOLLIFFE. *Principal Component Analysis*. Springer, 2002. ISBN : 9780387954424.
- [28] Varun CHANDOLA, Arindam BANERJEE et Vipin KUMAR. “Anomaly Detection : A Survey”. In : *ACM Computing Surveys (CSUR)* 41.3 (2009), p. 1-58. DOI : [10.1145/1541880.1541882](https://doi.org/10.1145/1541880.1541882).
- [29] Olivier CHAPELLE, Bernhard SCHÖLKOPF et Alexander ZIEN. *Semi-Supervised Learning*. MIT Press, 2009. ISBN : 0262033585, 9780262033589.
- [30] Richard S. SUTTON et Andrew G. BARTO. *Reinforcement Learning : An Introduction*. MIT Press, 2018. ISBN : 9780262039246.
- [31] Leo BREIMAN. “Random Forests”. In : *Machine Learning* 45.1 (2001), p. 5-32.

- [32] Naomi S. ALTMAN. "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression". In : *The American Statistician* 46.3 (1992), p. 175-185.
- [33] Ian GOODFELLOW, Yoshua BENGIO et Aaron COURVILLE. *Deep Learning*. MIT Press, 2016.
- [34] Trevor HASTIE, Robert TIBSHIRANI et Jerome FRIEDMAN. *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. 2nd. Springer, 2009, p. 219-259. ISBN : 978-0387848570. DOI : 10.1007/978-0-387-84858-7.
- [35] KOBIA. *Classification Metrics et Matrice de Confusion*. Consulté le 22 juillet 2024. 2024. URL : <https://kobia.fr/classification-metrics-matrice-de-confusion/>.
- [36] RESEARCHGATE. *Principe des courbes ROC*. Consulté le 22 juillet 2024. 2024. URL : https://www.researchgate.net/figure/Principe-des-courbes-ROC-A-Pour-chaque-seuil-possible-la-sensibilite-Se-et-la_fig29_337533229.
- [37] Kevin DEGILA. *La validation croisée en machine learning : tout ce qu'il faut savoir*. Consulté le 22 juillet 2024. 2024. URL : <https://www.kevindegila.com/fr-blog/la-validation-croisee-en-machine-learning-tout-ce-quil-faut-savoir/>.
- [38] DATATAB. *Intervalle de Confiance*. Consulté le 22 juillet 2024. 2024. URL : <https://datatab.fr/tutorial/confidence-interval>.
- [39] Aditya KARMOKAR. "Recognizing emotion from Speech using Machine learning and Deep learning". In : *Medium* (2020). URL : <https://medium.com/@adityakarmokar/recognizing-emotion-from-speech-using-machine-learning-and-deep-learning-30a6f4b8b0e5>.
- [40] Tao ZHANG, Ming WANG et Lei ZHANG. "Speech Emotion Recognition Using Deep Learning Techniques : A Review". In : *IEEE Access* 9 (2021), p. 100345-100365. DOI : 10.1109/ACCESS.2021.3059359.
- [41] Björn SCHULLER, Anton BATLINER et Stefan STEIDL. "Human Speech Emotion Recognition". In : *ResearchGate* (2015). URL : https://www.researchgate.net/publication/299185942_Human_speech_emotion_recognition.
- [42] Yang LIU, Qiang ZHANG et Wei WANG. "Speech Emotion Recognition Based on Two-Stream Deep Learning Model". In : *MDPI Sensors* 20.12 (2020), p. 3361. DOI : 10.3390/s20123361.
- [43] K. Sreenivasa RAO, Shashidhar G. KOOLAGUDI et M.A. WANI. "Identification of emotions from speech using Deep Learning". In : *IEEE Xplore* 8 (2021), p. 147973-147993. DOI : 10.1109/ACCESS.2020.3015184.
- [44] Mark OTTO et Jacob THORNTON. *Bootstrap*. <https://getbootstrap.com/>. Open-source front-end framework initially developed by Twitter. 2011.
- [45] Armin RONACHER. *Flask*. <https://flask.palletsprojects.com/>. Micro-framework for web development in Python. 2010.
- [46] KAGGLE. *Kaggle*. <https://www.kaggle.com/>. Platform for data science competitions and collaboration. 2010.
- [47] Guido van ROSSUM. *Python*. <https://www.python.org/>. Interpreted, high-level, general-purpose programming language. 1991.
- [48] Martín ABADI et al. *TensorFlow : Large-Scale Machine Learning on Heterogeneous Distributed Systems*. <https://www.tensorflow.org/>. Open-source library for machine learning and artificial intelligence. 2015.

- [49] Brian MC FEE et al. *librosa : Audio and Music Signal Analysis in Python*. <https://librosa.org/>. Open-source library for audio and music signal analysis in Python. 2015.
- [50] H. CAO et al. “CREMA-D : Crowd-sourced emotional multimodal actors dataset”. In : *IEEE Transactions on Affective Computing* 5.4 (2014), p. 377-390. DOI : 10.1109/TAFFC.2014.2336244. URL : <https://ieeexplore.ieee.org/document/6873295>.
- [51] Steven R. LIVINGSTONE et Frank A. RUSSO. “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)”. In : *PLOS ONE* 13.5 (2018), e0196391. DOI : 10.1371/journal.pone.0196391. URL : <https://doi.org/10.1371/journal.pone.0196391>.
- [52] Author or ORGANIZATION. *Surrey Audio-Visual Expressed Emotion (SAVEE) Dataset*.
- [53] Author or ORGANIZATION. *Toronto Emotional Speech Set (TESS)*.

- Codes

1- Caractéristiques audios

```
1 import matplotlib.pyplot as plt
2 import librosa
3 import librosa.display
4
5 def create_waveplot(data, sr, e):
6     plt.figure(figsize=(10, 3))
7     plt.title('Waveplot for audio with {} emotion'.
8             format(e), size=15)
9     plt.plot(data)
10    plt.xlabel('Time (samples)')
11    plt.ylabel('Amplitude')
12    plt.show()
13
14 def create_spectrogram(data, sr, e):
15     X = librosa.stft(data)
16     Xdb = librosa.amplitude_to_db(abs(X))
17     plt.figure(figsize=(12, 3))
18     plt.title('Spectrogram for audio with {} emotion'.
              format(e), size=15)
19     librosa.display.specshow(Xdb, sr=sr, x_axis='time',
20                             y_axis='hz')
```

```

19 plt.colorbar()
20 plt.show()
21
22 def create_mfcc(data, sr, e):
23     mfccs = librosa.feature.mfcc(y=data, sr=sr)
24     plt.figure(figsize=(10, 3))
25     plt.title('MFCC for audio with {} emotion'.format(e),
26               size=15)
27     librosa.display.specshow(mfccs, sr=sr, x_axis='time')
28     plt.colorbar()
29     plt.show()
30
31 def create_chroma(data, sr, e):
32     chromagram = librosa.feature.chroma_stft(y=data, sr=
33                                                 sr)
34     plt.figure(figsize=(10, 3))
35     plt.title('Chromagram for audio with {} emotion' .
36               format(e), size=15)
37     librosa.display.specshow(chromagram, sr=sr, x_axis='
38                                 time')
39     plt.colorbar()
40     plt.show()
41
42 def create_pitch_track(data, sr, e):
43     pitches, voiced_flag, voiced_probs = librosa.pyin(
44         data, fmin=librosa.note_to_hz('C2'), fmax=librosa.
45         note_to_hz('C7'), sr=sr)
46     plt.figure(figsize=(10, 3))
47     plt.title(f'Pitch Track for audio with {e} emotion',
48               size=15)
49     plt.plot(pitches, label='Pitch (Hz)')
50     plt.xlabel('Time (frames)')
51     plt.ylabel('Frequency (Hz)')
52     plt.legend()
53     plt.show()

```

2- Augmentation de donnée audio

```
1 import numpy as np
2 import librosa
3 # NOISE
4 def noise(data):
5     noise_amp = 0.035*np.random.uniform()*np.amax(data)
6     data = data + noise_amp*np.random.normal(size=data.
7         shape[0])
8     return data
9 # STRETCH
10 def stretch(data, rate=0.8):
11     return librosa.effects.time_stretch(data, rate)
12 # SHIFT
13 def shift(data):
14     shift_range = int(np.random.uniform(low=-5, high=5)
15     *1000)
16     return np.roll(data, shift_range)
17 # PITCH
18 def pitch(data, sampling_rate, pitch_factor=0.7):
19     return librosa.effects.pitch_shift(data,
20         sampling_rate, pitch_factor)
```