République Algérienne Démocratique et Populaire
الجمهورية الجزائرية الديمقراطية الشعبية
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
وزارة التعليم العالي و البحث العلمي
École nationale Polytechnique

Département électronique

Mémoire de projet de fin d'études

Pour l'obtention du diplôme d'Ingénieur d'État en Électronique

Développement et implémentation d'un système de reconnaissance des émotions à partir de la parole et des expressions faciales

Maroua AISSA & Romaila AIT MESBAH

Sous la direction de Mme. Nesrine Bouadjenek Dr. ENP, Alger

Présenté et soutenu publiquement le 22/06/2023 auprès des membres du jury :

Président	M. Rachid	ZERGUI	Prof.	ENP, Alger
Promotrice	Mme. Nesrine	Bouadjenek	Dr.	ENP, Alger
Examinateur	Mme. Nour El-Houda	BENALIA	Dr.	ENP, Alger

ENP 2023

10, Avenue des Frères Oudek, Hassen Badi, BP. 182, 16200 El Harrach, Alger, Algérie.
www.enp.edu.dz

République Algérienne Démocratique et Populaire
الجمهورية الجزائرية الديمقراطية الشعبية
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
وزارة التعليم العالي و البحث العلمي
École nationale Polytechnique

Département électronique

Mémoire de projet de fin d'études

Pour l'obtention du diplôme d'Ingénieur d'État en Électronique

Développement et implémentation d'un système de reconnaissance des émotions à partir de la parole et des expressions faciales

Maroua AISSA & Romaila AIT MESBAH

Sous la direction de Mme. Nesrine Bouadjenek Dr. ENP, Alger

Présenté et soutenu publiquement le 22/06/2023 auprès des membres du jury :

Président	M. Rachid	ZERGUI	Prof.	ENP, Alger
Promotrice	Mme. Nesrine	Bouadjenek	Dr.	ENP, Alger
Examinateur	Mme. Nour El-Houda	BENALIA	Dr.	ENP, Alger

ENP 2023

10, Avenue des Frères Oudek, Hassen Badi, BP. 182, 16200 El Harrach, Alger, Algérie.
www.enp.edu.dz

ملخص

تعد العواطف مكونًا رئيسيًا للتواصل البشري ، مما يضفي ثراءً فريدًا وفروقًا دقيقة في تفاعلاتنا. وقد جعل هذا التعرف التلقائي على العواطف مجالًا جذابًا للبحث ، مما يوفر إمكانيات تطبيق واسعة ، ولا سيما في مجال التفاعل بين الإنسان والحاسوب. في هذا العمل ، نقترح نظامًا للتعرف على المشاعر يعتمد على المعلومات السمعية والبصرية ، واستغلال تقنيات التعلم الآلي المختلفة. تم تصميم هذا النظام ليتم تنفيذه على لوحة إلكترونية ، من أجل عمل تنبؤات في الوقت الفعلي.

كلمات مفتاحية : .CNN التعرف التلقائي على العواطف ، معلومات سمعية وبصرية

Abstract

Emotions are a key component of human communication, bringing a unique richness and nuance to our interactions. This has made automatic emotion recognition an attractive area of research, offering vast application possibilities, particularly in the field of human-computer interaction. In this work, we propose an emotion recognition system based on audio-visual information, exploring various machine learning techniques. This system is designed to be implemented on an electronic board, in particular the Raspberry Pi, in order to make predictions in real time.

Keywords : Automatic emotion recognition, Audio-visual information, SVM, CNN, MFCCs, Raspberry Pi.

Résumé

Les émotions représentent une composante clé de la communication humaine, apportant une richesse et une nuance unique à nos interactions. Ce qui a rendu de la reconnaissance automatique des émotions un domaine de recherche attirant, offrant de vastes possibilités d'application, notamment dans le domaine de l'interaction homme-machine. Dans ce travail, nous proposons un système de reconnaissance des émotions basé sur des informations audio-visuelles, en exploitant les différentes techniques d'apprentissage automatique. Ce Système est conçu pour être implémenté sur une carte électronique, en particulier la Raspberry Pi, afin d'effectuer des prédictions en temps réel.

Mots clés : Reconnaissance automatique des émotions, Information audio-visuelle, SVM, CNN, MFCCs, Raspberry Pi .

Dédicace

À mes parents, mes sœurs Asma et Fella et mon frère Billel, à ma grand-père Ahmed et à tous les membres de ma famille, vous avez été mes piliers, mon soutien inconditionnel tout au long de mon parcours d'études. Votre amour, vos encouragements et vos sacrifices ont été la fondation solide sur laquelle j'ai construit mon chemin vers la réussite. Cette dédicace est une humble reconnaissance de votre présence constante dans ma vie, de votre soutien et de votre foi en moi.

À mes professeurs, depuis mes premiers jours à l'école primaire jusqu'à aujourd'hui, vous avez été les gardiens de la connaissance, les guides qui m'ont éclairé sur les sentiers complexes de l'apprentissage. Je vous suis reconnaissante pour chaque leçon enseignée, chaque encouragement prodigué et chaque moment d'inspiration partagé.

À ma promotion exceptionnelle *Eln Match*, nous avons été un groupe soudé, une équipe qui s'est soutenue mutuellement tout au long de ces trois ans. Ensemble, nous avons surmonté les défis, échangé des idées, et nous nous sommes encouragés à donner le meilleur de nous-mêmes. Je dédie ce projet à notre amitié et à notre complicité qui ont rendu ces années d'études inoubliables.

Enfin, à toi Romaila, ma plus belle rencontre. Ensemble, nous avons partagé des moments de rires, de doutes et de victoires, créant des souvenirs inoubliables qui resteront gravés dans mon cœur. Je suis reconnaissante d'avoir rencontré une âme aussi exceptionnelle que la tienne. Merci d'avoir été cette étincelle qui a animé notre projet de fin d'études. Notre collaboration restera gravée dans ma mémoire, et je suis impatiente de continuer à partager de nouveaux horizons avec toi au-delà de nos études.

- Maroua

Dédicace

À mes chers parents, qui ont été mes plus grands soutiens depuis que j'ai connu le jour. Votre amour inconditionnel, votre soutien indéfectible et vos sacrifices sans fin ont été les piliers qui m'ont permis d'atteindre cette étape importante de ma vie.Les mots ne seront jamais suffisants pour exprimer ma gratitude envers vous. Je vous aime infiniment.

À la mémoire de ma chère grand-mère, ma plus grande admiratrice. Tu as toujours été une source d'inspiration pour moi, ton amour, tes conseils et tes souvenirs resteront gravés dans mon cœur pour toujours. J'espère que j'ai réussi à te rendre fière Mami.

À mes amies , Ikram , Manel et Ichrak qui ont toujours été à mes côtés.

À ma promotion extraordinaire "Eln Match", avec laquelle j'ai partagé une aventure pleine de hauts et de bas, de larmes et de fous rires. Vous avez rendu ces trois années de spécialité inoubliables, et je suis reconnaissante d'avoir pu faire partie d'une si belle famille. Que nos chemins se croisent à nouveau dans l'avenir, et que nous continuions à grandir et à prospérer ensemble, en tant que deuxième famille indéfectible.

À ma meilleure rencontre à Polytech, *Maroua*, ma binôme qui est avant tout une amie très chère à mon cœur. Tu as toujours été là pour moi, prête à m'écouter, à me soutenir et à m'encourager. Sans ta persévérance, ton courage et ton savoir faire, ce travail n'aurait pas vu le jour. Merci d'avoir rendu de cette collaboration une magnifique expérience, ***Iyi ki sen***.

- Romaila

Remerciements

En ce moment particulièrement important de notre vie académique, nous tenons à exprimer notre profonde gratitude et reconnaissance à nos chères familles, qui ont été une source inépuisable de soutien et d'encouragement, dès le début de notre parcours, nous assurant le meilleur cadre possible pour étudier et nous épanouir.

Nous tenons à exprimer notre sincère gratitude à notre promotrice, Dr. Nesrine BOUADJENEK, qui nous a soutenu tout au long de ce travail. Nous la remercions pour sa patience, sa motivation, sa disponibilité et son dévouement sans pareil. Nous n'aurions pas pu espérer un meilleur superviseur et mentor.

Nous remercions également les membres du jury qui ont accepté de consacrer leur temps à examiner ce travail : Monsieur Rachid ZERGUI et Madame Nour El Houda BENALIA, nos Professeur à l'École Nationale Polytechnique.

Bien sur, nous tenons à remercier chaleureusement nos amis et nos camarades de classe, *ELN Match*, en particulier : Faten , Anes et Riad qui ont partagé ce voyage avec nous. Leurs encouragements, leur collaboration et leur soutien mutuel ont rendu cette expérience d'apprentissage encore plus significative et agréable.

Romaila & Maroua.

Table des matières

Liste des tableaux

Table des figures

Liste des abréviations

UA *Unité d'action.*

FACS *Facial Action Coding System.*

CNN *Convolutional neural network.*

MFCC *Mel-Frequency Cepstral Coefficients.*

RGB *Rouge Green Bleu.*

RAE *Reconnaissance automatique des emotions.*

RFE *Reconnaissance faciale des emotions.*

RVE *Reconnaissance vocale des emotions.*

HMM *Hidden Markov Model.*

SVM *Support Vector Machine.*

ReLu *Fonction Unité Linéaire Rectifiée.*

ZCR *Zero crossing rate .*

HoG *Histogram of oriented gradient.*

SAVEE *Surrey Audio-Visual Expressed Emotion.*

TESS *Toronto Emotional Speech Set .*

RAVDESS *Ryerson Audio-Visual Database of Emotion and Song .*

CK+ *The Extended Cohn-Kanade.*

KDEF *Karolinska Directed Emotional Face.*

RBF *Radial Basis Function.*

BDD *Base de données.*

TfLite *Tensor Flow Lite*

Introduction générale

La communication est le pilier essentiel de nos interactions quotidiennes, qu'elles soient personnelles, professionnelles ou sociales. Elle est le moyen par lequel nous échangeons des idées, partageons des expériences et établissons des liens avec les autres. Cependant, au-delà des mots prononcés, les émotions jouent un rôle primordial dans notre capacité à transmettre et recevoir des messages avec précision et profondeur. Elles constituent une composante intrinsèque de la communication humaine, apportant une richesse et une nuance uniques à nos interactions. Elles se manifestent à travers une multitude de canaux, tels que les expressions faciales, les intonations de voix, les gestes et les postures corporelles. Ces signaux non verbaux transmettent des informations essentielles sur notre état émotionnel, nos intentions et nos attitudes, complétant ainsi le sens littéral de nos paroles.

La compréhension et l'interprétation de ces émotions sont donc une compétence clés pour les individus ainsi que pour les machines interactives et intelligentes qui assure une communication efficace et harmonieuse. Dans ce contexte, la reconnaissance automatique des émotions est un domaine de recherche en plein essor qui vise à analyser et interpréter les différents états émotionnels grâce aux avancés technologiques. Les progrès de l'intelligence artificielle, du traitement du signal et de l'apprentissage automatique ont permis le développement de systèmes sophistiqués capables de détecter et d'analyser les émotions à partir de différents types de données, tels que la parole, les images et les vidéos. Ces systèmes offrent de vastes possibilités d'application, allant de l'amélioration des interactions homme-machine à l'évaluation des réponses émotionnelles dans des domaines tels que la psychologie et les sciences sociales.

Dans ce projet, notre objectif principal est de développer un système de reconnaissance automatique des émotions en utilisant les informations audiovisuelles provenant des expressions faciales et de la parole. Ce système sera conçu pour être implémenté sur une carte électronique, en particulier la Raspberry Pi 4. Pour atteindre cet objectif, nous allons procéder par étapes. Tout d'abord, nous évaluerons différentes caractéristiques extraites à partir des données prétraitées à l'aide des techniques de prétraitement, associées à chaque modalité. Ensuite, nous les introduirons à des algorithmes de classification, notamment les réseaux de neurones convolutifs et les SVMs, afin de développer des modèles correspondant à chaque modalité.

Une fois que nous aurons obtenu le modèle final pour chaque modalité, nous les fusionnerons en utilisant différentes techniques de fusion basées sur les scores. Cette fusion nous permettra de combiner les informations des expressions faciales et de la parole pour obte-

nir un modèle final plus robuste et précis dans la reconnaissance des émotions. Ce modèle final sera implémenté sur une Raspberry Pi 4.

Ce rapport est constitué de quatre chapitres qui sont structuré comme suit :

Le premier chapitre : présente un aperçu général sur les systèmes reconnaissance automatique des émotions, en commençant par la définition de son élément de base qui est l'émotion arrivant à la présentation des différents travaux réalisés dans ce domaine jusqu'à présent, offrant une vision globale de l'état de l'art.

Le deuxième chapitre : fournie une description détaillé des étapes essentielles du processus de reconnaissance automatique des émotions, en mettant l'accent sur les techniques de prétraitement, l'extraction des caractéristiques et la classification à l'aide d'algorithmes d'apprentissage automatique et approfondi.

Le troisième chapitre : consiste à expliquer la méthodologie de travail adopté , tout en fournissant une analyse profondes des performances obtenus.

Le quatrième chapitre : Décris le processus de l'implémentation du modèle final de la fusion de la RFE et la RVE sur une Raspberry Pi 4.

Chapitre 1

Généralités sur la reconnaissance automatique des émotions : Défis et état de l'art

1.1 Introduction

Les émotions jouent un rôle essentiel dans la communication et le comportement humain, influençant nos processus de prise de décision, nos relations et notre bien-être général. Par conséquent, le développement de systèmes automatisés capables de reconnaître et de comprendre avec précision les émotions humaines est devenu un domaine de recherche fascinant qui suscite un intérêt croissant dans de nombreux domaines tels que la psychologie, les sciences cognitives, l'informatique et l'intelligence artificielle.

Dans ce chapitre, nous explorons en détail les concepts fondamentaux de la reconnaissance automatique des émotions. Nous commençons par présenter une vue d'ensemble des émotions, en abordant leur définition, leurs caractéristiques. Ensuite, nous examinons les différentes approches de classification des émotions, mettant en évidence les modèles les plus couramment utilisés.

Nous nous concentrons ensuite sur les modalités de communication des émotions, en mettant en évidence deux canaux majeurs : la parole et les expressions faciales.De plus, nous passerons en revue les travaux de recherche préalablement réalisés dans ce domaine, afin de comprendre les avancées technologiques et les défis actuels de la reconnaissance automatique des émotions.

1.2 Émotions

L'émotion est l'un des concepts les plus difficiles à définir en psychologie. Selon *Plutchik*[1] , plus de quatre-vingt-dix définitions de l'émotion ont été proposées au cours du vingtième siècle . C'est un état psychologique complexe qui implique une série de changements physiologiques et psychologiques en réponse à un stimulus. Elle peut inclure des sentiments, des pensées et des comportements qui sont liés à la perception d'une situation ou d'un événement.Les émotions jouent un rôle crucial dans les interactions et le comportement humain et peuvent avoir un impact profond sur notre vie quotidienne.
Elles peuvent être aperçues sous forme de plusieurs composantes inter-connectées :

- **La composante comportementale :** c'est l'ensemble de comportements observables qui accompagne l'émotion tel que : les expressions faciales et vocales.
- **La composante physiologique :** réaction corporelle tel que le changement du rythme cardiaque et la respiration , augmentation de la tension artérielle.
- **La composante cognitive :** Cette composante implique la manière dont nous interprétons, évaluons et attribuons du sens aux événements, ainsi que la manière dont nous anticipons les résultats futurs.

1.3 Classification des émotions

En raison de la complexité et de la subjectivité des émotions, les spécialistes n'ont pas pu parvenir à un consensus clair sur la classification des émotions tout comme sa définition. Par conséquent, plusieurs modèles de classification ont été développés, chacun apportant une perspective différente sur la compréhension des émotions. Parmi les modèles les plus couramment discutés :

1.3.1 Modèle discret

Ce modèle est basé sur une approche théorique qui suggère que les émotions peuvent être classées en catégories distinctes et discrètes, chacune ayant des caractéristiques spécifiques. Ainsi, nous distinguons deux principaux types :

1.3.1.1 Émotions primaires

Ce sont des états émotionnels universels et inter-culturels, étant reconnus et vécus de manière similaire dans différentes cultures dont le nombre varie selon les auteurs.

Selon le psychologue *Paul Ekman*, il existe 6 émotions de base : la joie, la tristesse, le dégoût, la peur, la colère et la surprise.Cependant, d'autres chercheurs, tels que *Robert Plutchik*, soutiennent l'existence de huit émotions primaires, en ajoutant l'anticipation et la confiance à cette liste.

Acteurs	Émotions basiques
EKMAN ET AL	Colère, dégoût, joie, tristesse, peur, surprise, neutralité
PLUTCHIK	Acceptation, colère, anticipation, dégoût, peur, joie, tristesse, surprise
FRIDJA	Intérêt, joie, désir, chagrin, émerveillement
TOMKINS	Colère, intérêt, mépris, dégoût, détresse, peur, joie, honte, surprise
Arnold	Courage, colère, aversion, désir, désespoir, tristesse, amour, espoir, abattement, haine, peur
McDougall	Peur, dégoût, exaltation, colère, émotion tendre, soumission, émerveillement

Tab. 1.1 : Émotions de base selon différents acteurs[2]

L'ensemble d'émotions le plus adopté dans les travaux de recherches est celui établit par *Eckman* :

- **Joie :** c'est une émotion associée à un sentiment de satisfaction et de plaisir qui se manifeste par les signes physiologiques suivants : sourire, ton de voix vif, rapide et aigu.
- **Peur :** elle est associée à une sensation d'anxiété ou de menace. Elle peut être provoquée par des événements tels que des dangers physiques, des situations inconnues ou des phobies : accélération du rythme cardiaque , ouverture des yeux, tremblement.
- **Tristesse :** C'est le sentiment provoqué par une défaite, une déception ou une perte. Cette émotion se caractérise par : une faible voix avec une intonation descendante, des sourcils sous forme oblique, les commissures des lèvres orientées vers le bas..

- **Colère :** C'est une émotion intense qui est généralement en réponse à une menace perçue, une injustice ou une frustration. Elle se manifeste par une voix forte avec un rythme rapide , un visage contracté (mâchoire serré, froncement des sourcils)

- **Dégoût :** Cette émotion tend à impliquer un sentiment de répulsion à l'égard du goût, de l'odeur, du toucher ou de la vue d'une chose désagréable. Elle est caractérisée par un visage crispé et une voix faible et nasale.

- **Surprise :** cette émotion est associée à une sensation d'étonnement ou de choc. Elle peut être provoquée par des événements inattendus ou surprenants.Elle se manifeste par un front contracté et des sourcils en mouvement vers le haut..

- **Neutralité :** L'émotion neutre est un état émotionnel caractérisé par une absence ou une faible intensité émotionnelle. Il peut être considéré comme une sorte de point de départ ou de référence pour les émotions positives ou négatives.

La figure ci-dessous illustre les expressions faciales associés à chaque émotions :

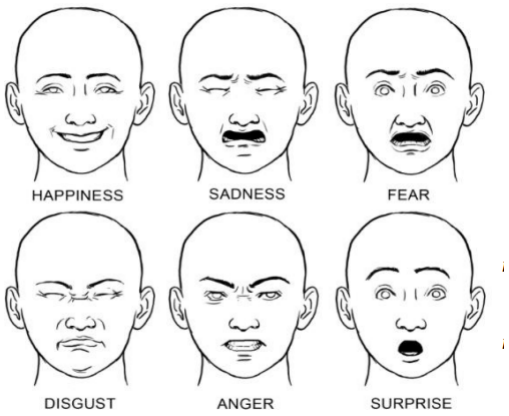


FIG. 1.1 : Les six émotions faciales [3]

1.3.1.2 Émotions secondaires

Ce sont des émotions complexes, qui peuvent être issues de la combinaison des émotions primaires.Elles peuvent être plus complexes et nuancées que les émotions primaires et peuvent varier selon les individus et les situations. Par exemple, la haine, la jalousie et la culpabilité...

1.3.2 Modèle multidimensionnel

Le modèle émotionnel dimensionnel est un modèle qui utilise certains facteurs connus comme 'dimension' pour caractériser les émotions, tels que la valence, l'excitation, le contrôle et la puissance. Cette approche affirme que les états affectifs ne sont pas indépendants les uns des autres ; ils sont plutôt reliés les uns aux autres de manière systématique. L'un des modèles multidimensionnels les plus courant c'est le modèle bi-dimensionnel qui répartit les émotions en fonction de deux axes :

1. **la valence :** Il s'agit de la dimension qui mesure la positivité ou la négativité d'une émotion. Les émotions positives, comme la joie, ont une valence positive, tandis que les émotions négatives, comme la colère, ont une valence négative.

2. **l'arousal :** Cette dimension définit l'intensité du ressenti de l'émotion. Le degré d'excitation varie de la somnolence ou de l'ennui vers l'excitation frénétique.

1.4 Reconnaissance automatique des émotions

La reconnaissance automatique des émotions(RAE) est un sujet de recherche en constante évolution qui vise à développer des systèmes capables de détecter et d'interpréter les émotions à partir de différents types de signaux. Cette dernière appartient à la famille de technologies souvent désignées par l'expression *affective computing*[4], un domaine de recherche pluridisciplinaire portant sur les capacités des ordinateurs à reconnaître et à interpréter les émotions humaines et les états affectifs, et qui s'appuie souvent sur les technologies de l'intelligence artificielle.

Le processus de la RAE est porté sur 3 étapes principales comme le montre la figure 1.2 :

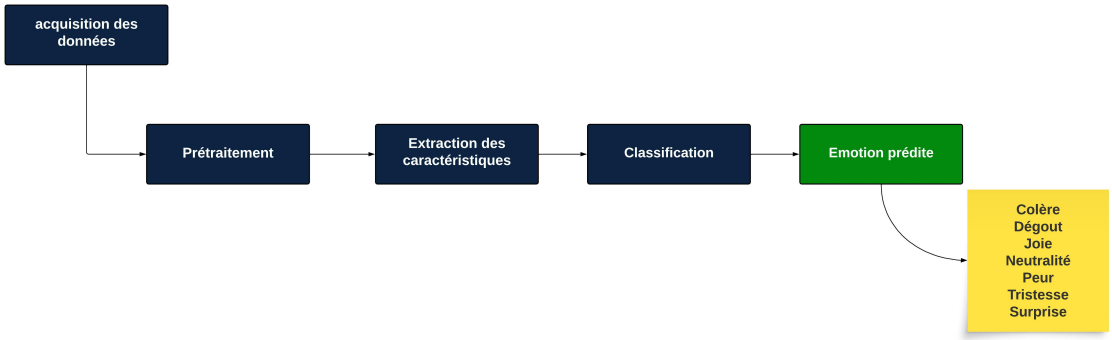


Fig. 1.2 : Les étapes de la RAE

1. **Pré-traitement des données :** Consiste à éliminer le bruit , normalisé et préparer les données afin de faciliter leurs traitements par les algorithmes de classification.

2. **Extraction des caractéristiques :** c'est une étape importante qui consiste à extraire les informations pertinentes des données d'entrée afin de les représenter de manière appropriée pour une analyse ultérieure.

3. **Classification :** au niveau de cette étape, les caractéristiques extraites seront introduites comme un vecteur d'entrée à des algorithmes de classification, qu'ils soient classiques ou basés sur l'apprentissage automatique, afin d'attribuer une émotion à la donnée d'entrée.

Il existe plusieurs types de système RAE, en fonction de la source de données utilisée. Les types de RAE les plus courants sont les suivants :

- La reconnaissance des émotions à partir des expressions faciales,
- La reconnaissance des émotions à partir de la parole,
- La reconnaissance à partir des signaux physiologiques (EEG, ECG…)

Dans ce travail, on s'intéresse à la reconnaissance des émotions à partir des expressions faciales et de la parole.

1.4.1 Reconnaissance faciale des émotions

Les différentes études affirment que les composantes non verbales représentent deux tiers de la communication humaine, parmi lesquelles les expressions faciales qui traduisent des états émotionnels internes, des intentions ou des communications sociales d'une personne.

Depuis des décennies, le décodage de ces expressions émotionnelles a fait l'objet de recherches dans le domaine de la psychologie et c'est en 1978 que les psychologues **Ekman Friesen** ont réussi à définir un système de codage des actions faciales, en anglais appelé *Facial Action Coding System (FACS)*[5]. Ce dernier est basé sur l'encodage des mouvements de muscles faciaux spécifiques appelés unités d'action (UA), qui reflètent des changements momentanés distincts dans l'apparence du visage. Le système $FACS$ repose sur la description de 46 UA identifiées par un numéro dans la nomenclature $FACS$.

Pour reconnaître les émotions faciales, les UAs individuelles sont détectées et le système classifie la catégorie de visage selon leurs combinaison comme le montre la figure 1.3 :

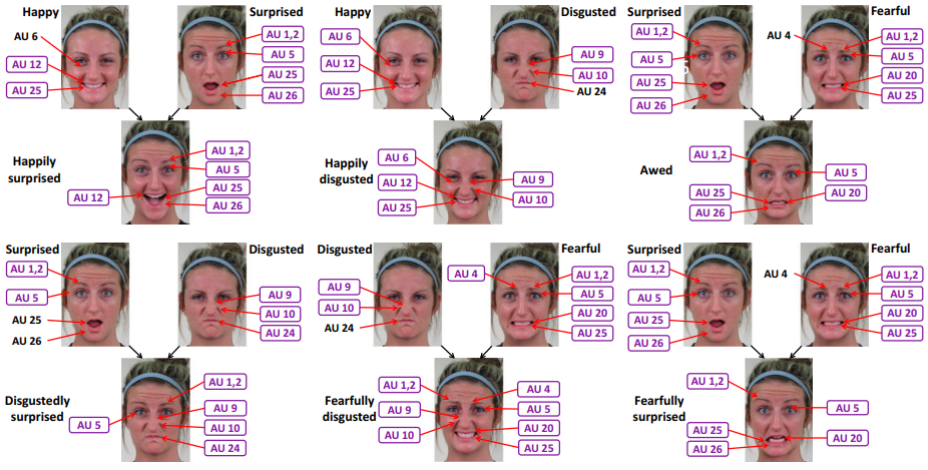


FIG. 1.3 : Illustration des unités d'action de six émotions composées [6]

Cependant, la reconnaissance des émotions faciales peut également se baser sur l'extraction de points de repère faciaux, qui représentent le positionnement spatial de points et de régions spécifiques du visage tels que le coin des yeux ou le bout du nez. Ces points de repère sont ensuite suivis et analysés pour mesurer le mouvement du visage au fil du temps.

Le tableau 1.2 ci-dessous donne les expressions faciales des émotions primaires :

TAB. 1.2 : Les émotions basiques et leurs expressions faciales associées

Émotions	Expression faciale
Joie	• Les coins des lèvres tirés vers le haut. • Les rides aux coins des yeux
Tristesse	• Sourcil extérieur vers le bas • Les coins des lèvres tirés vers le bas
Colère	• Sourcils froncés, • Lèvres serrées, • Paupières supérieures et inférieures relevées.
Dégoût	• Lèvre supérieure retroussée • Nez froncé
Peur	• Sourcils relevés et rapprochés • Bouche ouverte
Surprise	• Yeux et bouche ouverts • Des rides au niveau du front

Bien que la *RFE* soit une technique prometteuse pour la reconnaissance des émotions dans différents domaines, elle présente également des limites en termes de variabilité interindividuelle. Les expressions faciales peuvent varier considérablement d'une personne à une autre, en fonction de leur âge, de leur genre, de leur culture et de leur environnement d'où la difficulté de généralisation du modèle de détection.

1.4.2 Reconnaissance vocale des émotions

La parole ,connue comme étant une composante de communication verbale, est un flux audio continu qui est représenté comme un signal.

Ce type de reconnaissance repose sur l'analyse approfondie du processus de génération du signal vocal. Elle consiste à extraire des caractéristiques spécifiques de la voix du locuteur contenant des informations émotionnelles. Ces caractéristiques sont des mesures quantitatives qui capturent des aspects acoustiques et prosodiques de la parole tels que : l'énergie du signal, la fréquence fondamentale(pitch), la durée, le rythme du signal et les transitions spectrales…

Chacun de ces paramètres fournit des informations uniques sur la manière dont les émotions sont exprimées. La fréquence fondamentale, par exemple, permet de détecter les variations de hauteur de la voix tandis que l'énergie de la voix peut refléter l'intensité de l'émotion.
La RVE vise donc à comprendre et à interpréter les informations émotionnelles contenues dans la voix humaine. En analysant les caractéristiques acoustiques et en les associant à des émotions spécifiques et cela en se basant sur les différentes techniques d'apprentissage automatique. Le tableau ci-dessous 3.3 illustre certaines caractéristiques vocales associées aux émotions primaires :

TAB. 1.3 : Les émotions basiques et leurs caractéristiques vocales associées

Émotions	Caractéristique de la voix
Joie	• Énergie vocale élevée. • Volume élevé • Rythme rapide • Intensité élevée • Variation mélodique marqué • Durée des pauses plus courte
Tristesse	• Faible énergie • Voix basse • Rythme lent • Intensité réduite • Durée des pauses plus longue
Colère	• Énergie vocale élevée. • Volume élevé • Rythme rapide et irrégulier • Intensité élevée, • Variation mélodique abrupte,
Dégoût	• Énergie vocal variable • Volume variable • Rythme variable • Variation mélodique instable, • Durée des pauses variable
Peur	• Énergie vocale variable • Volume élevé • Variation mélodique instable, • Rythme rapide et saccadé • Durée des pauses imprévisible
Surprise	• Énergie vocale variable • Volume élevé • Variation mélodique brusque, • Rythme rapide et irrégulier, • Durée de pause courte

Il est important de noter que ces caractéristiques vocales peuvent varier en fonction de la culture, de l'individu et du contexte. Cette variabilité constitue l'un des principaux défis de la reconnaissance vocale des émotions.

1.4.3 Domaine d'application

La reconnaissance automatique des émotions est un domaine de recherche en plein essor qui couvre un large champs d'applications dont on cite notamment :

(a) **L'interaction homme-machine :** améliorer les interactions avec les systèmes informatiques en permettant aux machines de détecter les émotions des utilisateurs et d'adapter leur comportement en conséquence [7].

(b) **Santé mentale :** détecter les signes précoces de troubles mentaux tels que la dépression, l'anxiété, le stress et l'autisme, permettant ainsi une intervention précoce et un suivi adapté des patients.

(c) **Marketing :** analyser les réactions émotionnelles des clients aux produits et services, ce qui peut aider les entreprises à améliorer leurs stratégies de marketing.

(d) **Détection de fraude :** détecter les émotions suspectes ou trompeuses dans des situations telles que les entretiens d'embauche, les interrogatoires ou les transactions financières, contribuant ainsi à renforcer la sécurité et la fiabilité des processus.

(e) **Éducation :** évaluer l'engagement des étudiants, leur niveau de concentration et leur réaction émotionnelle lors de tâches d'apprentissage. Cela peut aider les enseignants à adapter leur approche pédagogique pour une meilleure compréhension et rétention des informations [8].

1.5 Travaux connexes

Au cours des dernières décennies, de nombreux travaux ont été réalisés pour explorer les différentes dimensions des émotions, allant des émotions basiques jusqu' aux émotions plus complexes et subtiles. Les chercheurs ont étudié la relation entre les expressions faciales, les mouvements du corps, la prosodie de la voix, et les émotions ressenties, afin de développer des modèles et des algorithmes permettant de reconnaître et de comprendre ces signaux émotionnels.

Dans l'état de l'art actuel, de nombreuses approches ont été explorées, allant des techniques classiques basées sur l'extraction de caractéristiques manuelles aux méthodes plus récentes exploitant les réseaux de neurones profonds. Les ensembles de données annotées ont également joué un rôle crucial dans le développement de modèles performants, permettant ainsi d'entraîner et de tester les algorithmes de reconnaissance des émotions.

1.5.1 Reconnaissance vocale des émotions

Les premiers efforts pour reconnaître les émotions à partir du signal de la parole étaient généralement basées sur des méthodes d'apprentissage automatique et de traitement du signal. En suivant le même chemin que la reconnaissance automatique de la parole, il y a

eu de nombreuses implémentations basées sur les HMM [9], les GMM et les SVM.

Parmi ces études, on cite les travaux de Lalitha et al. [10] qui ont réussit à atteindre une exactitude de 82.1% en classifiant les signaux de la base de donnée *EmoDB* avec le classifieur SVM qui reçoit en entrée un vecteur caractéristique contenant : Entropie des coefficients de la transformée en ondelettes discrète exploitée par l'énergie de Teager, Coefficients cepstraux prédictifs linéaires , le pitch et d'autres. Pour la même base de données les auteurs de [11] ont songé à établir un système à trois étages du SVM pour un seulement les coefficients MFCCs et ont atteint 68% d'exactitude.

Dans l'article [12], une différente combinaison de caractéristiques constituée de : coefficients MFCCs, l'énergie, pitch, flux spectral, l'affaiblissement spectral et la stationnarité spectrale a été extraite puis classifié avec le SVM. Le système a été entraîné et testé sur la base de données $Emo - DB$ et a aboutit à une exactitude moyenne de $86,6\%$. Cette précision a encore augmenté pour atteindre 100% en considérant une classification binaire des émotions : positives ou négative, ce qui indique que plus le nombre de classes augmente, plus la exactitude est vouée à diminuer.

Au cours des années 2000, avec les progrès réalisés dans le domaine de l'apprentissage profond, les chercheurs ont orienté leurs efforts vers l'exploitation des avantages offerts par les réseaux de neurones. L'article [13] propose d'utiliser des réseaux neuronaux profonds (RNP) afin d'extraire des caractéristiques pertinentes telles que les coefficients MFCC, et des informations sur la hauteur (le pitch) à partir de données brutes. Une classification primaire des segment du signal de parole est effectuée qui va être par la suite utilisée pour la classification de l'énoncé. Les résultats expérimentaux démontrent que l'approche proposée conduit à une amélioration de 20%.

Le modèle proposé en [14] est constitué de trois tours de CNN : un CNN 1D qui reçoit en entrée le signal temporel brute, un CNN 2D pour la modélisation du spectrogramme de Mel 2D temps-fréquence et un troisième CNN 3D pour la modélisation dynamique temporelle-spatiale, afin d'apprendre des caractéristiques multimodales profondes au niveau des segments à partir du signal original. Une stratégie de fusion au niveau du score a été adoptée comme méthode de fusion multi-CNN afin d'intégrer les différents résultats pour la classification finale des émotions. Des expériences menées sur deux ensembles de données $AFEW5.0$ et $BAUM - 1$, démontrent des performances prometteuses.

L'étude comparative [15] réalisée en utilisant une combinaison des bases de données $CREMA - D$, $RAVDESS$, $SAVEE$, $IEMOCAP$ et $TESS$ conclut que l'utilisation du spectrogramme de Mel comme méthode d'extraction de caractéristiques améliore de manière significative la précision des mesures. De plus, l'utilisation de multiples ensembles de données permet de réduire la propension du modèle à apprendre les caractéristiques spécifiques des enregistrements, étant donné que les ensembles de données présentent une diversité sonore grâce à l'utilisation de différents équipements d'enregistrement.

1.5.2 Reconnaissance faciale des émotions

Dans les années 1990, les premières avancées dans le domaine de la reconnaissance des émotions faciales ont été réalisées en utilisant des méthodes classiques qui demeurent pertinentes à ce jour. Et malgré la diversité des approches conventionnelles étudiées, les étapes suivies restent généralement les mêmes, d'abord la détection de la région du visage qui joue un rôle crucial dans le prétraitement, suivi de l'extraction de caractéristiques géométriques, d'apparence, ou d'une combinaison des deux, et enfin la classification à l'aide de techniques telles que le SVM, le HMM ou le Random Forest. [16], [17].

En dépit du succès notable des méthodes traditionnelles de reconnaissance faciale, au cours de la dernière décennie, les chercheurs se sont orientés vers l'approche de l'apprentissage profond [18] en raison de sa grande capacité à extraire des caractéristiques hautement discriminantes. De plus, elle est plus flexible et puissante, permettant d'obtenir des résultats plus précis et robustes, comme on peut constater dans l'article de référence [19] où plusieurs modèles ont été développés dans le but de classifier six émotions de base ainsi que l'état neutre. Lorsqu'ils ont été évalués sur l'ensemble de données FER-2013, le classifieur SVM a obtenu une exactitude de $45,95\%$, tandis que le CNN a donné de meilleurs résultats avec une exactitude de $20,72\%$. En revanche, sur le jeu de données CK+, le CNN a atteint une exactitude de $98,4\%$. Une implémentation en temps réel a été réalisée en utilisant l'outil de détection de visage de la bibliothèque *OpenCV* à partir d'une vidéo. Le système a réussi à classifier de manière fiable certaines des sept émotions étudiées.

Mollahosseini et al. [20] ont proposé un CNN profond pour la reconnaissance des émotions faciales à partir de plusieurs bases de données disponibles. Après avoir extrait les repères faciaux des données, les images ont été réduites à une taille de 48×48 pixels. L'architecture utilisée se compose de deux couches de convolution-pooling, auxquelles sont ajoutés deux modules de style Inception.

Dans l'article [21] la reconnaissance des expressions faciales en temps réel est basée sur l'opérateur adaptatif de Canny avec l'algorithme *AAM* (Active Appearance Model) ce qui a conduit à une réduction de la complexité de calcul et une amélioration de la précision de la localisation des points caractéristiques. Une autre technique pour l'extraction des caractéristiques a été mise en oeuvre dans [22] en utilisant l'histogramme de descente de gradient. Afin d'éliminer l'égalisation de l'histogramme et le bruit des images pour améliorer le contraste, un filtre médian a été appliqué aux données brutes.

D'autres chercheurs se sont concentrés sur l'impact des techniques de prétraitement dans leurs recherches. Dans l'étude mentionnée [23], l'augmentation des données, la correction de la rotation, le cadrage, l'échantillonnage et la normalisation de l'intensité ont été évalués sur trois bases de données disponibles, à savoir $CK+$, $JAFFE$et $BU-3DFE$. Les résultats ont montré que la combinaison de toutes ces étapes de prétraitement est plus efficace que leur application individuelle. Une autre étude a également examiné différentes combinaisons de techniques [24], telles que la détection de visage, le recadrage, la normalisation globale, la normalisation locale et l'égalisation d'histogramme. L'application de la seule détection de visage a permis d'obtenir une exactitude de $86,08\%$, tandis que la combinaison de toutes ces techniques a conduit à une exactitude de $97,06\%$.

Afin de mieux comprendre les caractéristiques extraite par les couches de convolutions, les chercheurs [25] ont examiné la relation entre les caractéristiques utilisées par ces réseaux, les $FACS$ et les unités d'action (UA). Les résultats ont été évalués sur les ensembles de données $CK+$, $NovaEmotions$ et $FER-2013$.

1.5.3 Reconnaissance des émotions à travers des informations audiovisuelles

Certaines émotions sont plus facilement détectées à travers la parole, tandis que d'autres sont mieux exprimées à travers les expressions faciales. Afin d'obtenir une compréhension plus complète de l'état émotionnel d'un individu, les experts dans ce domaine ont envisagé d'exploiter plusieurs sources d'informations, notamment l'information audiovisuelle. Les étapes de traitement restent généralement les mêmes, à l'exception d'une étape supplémentaire qui consiste à fusionner les informations. Cette fusion peut être réalisée au niveau des caractéristiques extraites ou après la classification, au niveau des scores obtenus.

Le modèle proposé en [26] est constitué de deux blocs CNN, le premier responsable de l'extraction des informations visuelles, tandis que le second est dédié à l'extraction des caractéristiques à partir de la représentation temps-fréquence du signal audio. Les deux sorties des deux blocs sont fusionnées puis passées à un empilement de deux couches entièrement connectées qui effectuent la classification, conduisant ainsi à une amélioration de 6% d'exactitude.

Dans l'étude [27], une autre forme de fusion au niveau de la décision a été réalisée en utilisant deux classifieurs GMM (Gaussian Mixture Model) de manière à sélectionner la décision finale avec la probabilité la plus élevée. Les auteurs ont exploité les caractéristiques géométriques pour les images et les caractéristiques prosodiques pour les audios.
D'autres travaux, tels que ceux présentés dans [28], ont effectué la fusion en utilisant un algorithme de répartition optimale des poids, où les poids de chaque mode sont mis à jour afin de minimiser l'erreur de rotation. Cette approche permet une fusion plus précise des informations provenant de différentes sources.
Par ailleurs, la fusion peut également être réalisée à l'aide de techniques d'apprentissage automatique, comme démontré dans l'étude [29], où deux classifieurs ANN (Artificial Neural Network) et K-NN (K-Nearest Neighbors) ont été utilisés pour la fusion, dans le but d'améliorer les performances du système.

Dans l'article [30], un système hybride de fusion a été développé, composé de quatre étapes distinctes. D'une part, il comprend le calcul de l'émotion à partir de la parole et des expressions faciales de manière séparée, et d'autre part, il effectue la détection des émotions en fusionnant les caractéristiques extraites. Enfin, la dernière étape consiste à calculer l'émotion finale à partir des étapes précédentes.

Dans l'article [31], les auteurs ont utilisé une méthode de sélection par étapes pour ne conserver que les caractéristiques les plus pertinentes, ce qui a permis d'obtenir une

meilleure précision de reconnaissance. Ils ont également proposé un système multi-classifieurs basé sur le classifieur de l'Analyse Discriminante Linéaire de Fisher. Le principe consiste à créer un classifieur binaire pour chaque émotion, et lorsque le modèle résulte en la présence de deux émotions différentes en même temps, la décision revient au classifieur global.

L'objectif fondamental de l'apprentissage approfondi est de prendre une donnée brute en entrée et d'obtenir le résultat souhaité à la fin du processus. Cette réalisation a été accomplie dans l'étude [32] en développant une architecture de réseau de neurones convolutifs (CNN) pour extraire les caractéristiques à partir des échantillons audio, et en utilisant le modèle Rasnet-50 pour extraire les caractéristiques des images. Les deux vecteurs de sortie des modèles précédents sont ensuite concaténés pour être alimentés dans un réseau de neurones récurrents à mémoire à court terme (LSTM). Le tableau 1.4 résume les travaux cités précédemment :

Tab. 1.4 : Tableau récapitulatif des travaux connexes.

Référence	Modalité	Bases de données	Classifieur	Exactitude
[33]	RVE	Emo-DB, SAVEE RAVDESS	CNN	92%
[10]	RVE	EmoDB	SVM	82.14 %
		RAVDESS		71.61%
[34]	RVE	EMO-DB	CNN	86.1%
		IEMOCAP		64.3%
[11]	RVE	EMO DB	SVM	68%.
[19]	RFE	FER 2013	SVM	45,95%
[24]	RFE	CK+, JAFFE, MUG	CNN	97,06%
[27]	Fusion	leurs propres données	GMM	80%
[30]	Fusion	leurs propres données	SVM	97.5%

1.6 Conclusion

Dans ce premier chapitre, nous avons abordé les notions fondamentales liées aux émotions, ainsi que les différentes approches de leur classification. Nous avons également défini le concept de reconnaissance automatique des émotions, qui peut être réalisée à partir de diverses sources telles que la parole et les expressions faciales, qui seront les objets d'étude de notre recherche. Nous avons établi les liens entre les émotions exprimées et les informations audiovisuelles, en examinant comment les émotions peuvent être véhiculées à travers le signal vocal et les expressions du visage. De plus, nous avons souligné certains domaines d'application de la reconnaissance automatique des émotions, en mettant en évidence notamment le domaine de l'interaction homme-machine. Enfin, nous avons recensé les travaux les plus pertinents dans notre domaine d'étude.

En conclusion, nous avons introduit les bases nécessaires pour se familiariser avec le thème de notre projet. Dans le prochain chapitre, nous allons nous concentrer sur les techniques avancées utilisées dans ce domaine, en mettant l'accent sur les réseaux de neurones, les méthodes de pré-traitement des données et l'extraction des caractéristiques.

Chapitre 2

Techniques de prétraitement, d'extraction des caractéristiques et de classification

2.1 Introduction

Les systèmes de reconnaissance automatique des émotions se constituent de trois blocs essentiels comme il a déjà été indiqué dans le chapitre précédent qui sont : le prétraitement, l'extraction des caractéristiques et enfin la classification. Dans ce chapitre, nous aborderons les techniques essentiels de ces derniers. Nous discuterons des différentes techniques de prétraitement spécifiquement adaptées aux données d'expression faciale et de signaux vocaux. Ensuite, nous nous pencherons sur l'extraction des caractéristiques, une étape importante pour représenter les informations pertinentes présentes dans les données. Enfin, nous nous concentrerons sur la classification des émotions à l'aide d'algorithmes d'apprentissage automatique et approfondi.

2.2 Prétraitement

Le prétraitement joue un rôle clé dans le processus global de reconnaissance des émotions. Cette étape vise à améliorer la qualité des données d'entrée et à mettre en évidence les régions d'intérêt afin de les préparer pour les étapes ultérieures d'extraction de caractéristiques et de classification. Les techniques de prétraitement diffèrent selon le type de données traitées. Dans cette section, nous allons détailler ces méthodes pour les signaux de la parole et pour les images faciales.

2.2.1 Techniques de prétraitement des signaux de la parole

Les étapes du prétraitement du signal de la parole illustrées dans la figure 2.1 permettent la préparation de ce dernier pour la phase de l'extraction précise des informations qu'il contient. Ces étapes sont : l'application du filtre de préaccentuation, la segmentation et le fenêtrage.

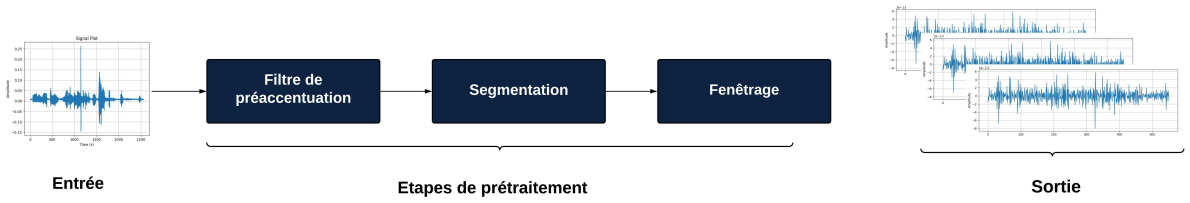


FIG. 2.1 : Étapes du prétraitment de signal de la parole.

2.2.1.1 Filtre de préaccentuation

C'est une technique utilisée dans le domaine du traitement du signal audio pour améliorer la netteté et la clarté des hautes fréquences d'un signal. Il s'agit d'un filtre passe-haut qui atténue les basses fréquences et amplifie les hautes fréquences. Il est défini comme suite :

$$y(n) = x(n) - \alpha \dots x(n-1) \tag{2.1}$$

$y(n)$: Signal de sortie du filtre.
x(n) : Échantillon d'entrée du signal audio àl'instant n.
$x(n-1)$: Échantillon précédent du signal audio.
α : Coefficient de préaccentuation ayant une valeur comprise entre $0,9$ et 1[35].

L'idée derrière le filtre de préaccentuation est de compenser la perte d'énergie dans les hautes fréquences qui peut se produire lors de l'enregistrement ou de la transmission d'un signal audio due à leur sensibilité aux distorsions. La figure 2.2 montre l'effet du filtre de préaccentuation sur un signal audio.

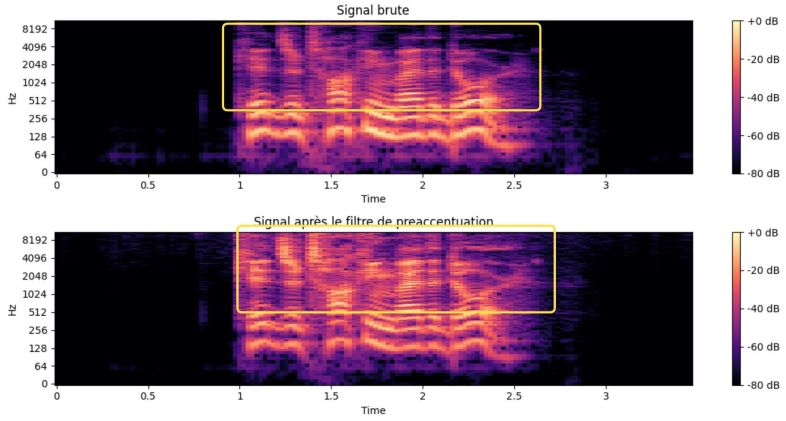


FIG. 2.2 : Effet du filtre de préaccentuation : après l'application du filtre de préaccentuation, l'énergie des hautes fréquences a augmenté comme démontré dans la deuxième figure par le changement de couleur.

2.2.1.2 Segmentation

Le signal de parole est de nature non stationnaire ce qui rend sa manipulation complexe. Cependant, il se comporte comme un signal stationnaire et invariant pendant une courte durée. C'est précisément à ce stade que la technique segmentation intervienne.
La segmentation est une étape fondamentale du traitement du signal audio qui consiste à diviser le signal continu en segments courts et facile à manipuler dits trames ou *frames*.Elle est définie par les paramètres suivants :

- **Taille de trame :** elle est généralement choisie pour assurer la stationnarité du signal à l'intérieur de chaque trame. Les valeurs couramment utilisées se situent entre 20 et 30 ms, permettant ainsi de capturer des portions du signal où les propriétés acoustiques sont relativement constantes.
- **Décalage entre trames (overlap) :** le décalage entre les trames détermine le chevauchement entre les segments adjacents. Souvent, un chevauchement de 50% est utilisé, ce qui signifie que chaque trame se chevauche avec la moitié de la trame précédente et la moitié de la trame suivante. Cela garantit une meilleure transition entre les segments tout en préservant les variations temporelles.

En résumé, la segmentation sert à faciliter l'analyse et le traitement ultérieur du signal en permettant une approche plus ciblée et spécifique à chaque segment. De plus, en segmentant le signal, on peut extraire des caractéristiques locales spécifiques à chaque partie du signal, ce qui permet d'obtenir des informations plus précises et détaillées sur les différentes composantes du signal.

2.2.1.3 Fenêtrage

Le fenêtrage est une étape spécifique qui intervient après la segmentation afin de réduire les discontinuités aux extrémités de la trame qui peuvent provoquer des distorsions dans les caractéristiques spectrales et temporelles du signal, ce qui risque d'affecter la précision des analyses ultérieures.

Il consiste à appliquer une fonction de fenêtre à chaque trame individuellement, qui est généralement une fonction mathématique, telle que la fenêtre de *Hamming* ou la fenêtre de *Hann*, qui réduit l'amplitude des échantillons au bord de la trame. Le choix de la fonction de fenêtre dépend du contexte et des objectifs spécifiques de l'analyse.

Au cours de notre projet, nous avons décidé d'appliquer la fenêtre de *Hamming* présentée dans la figure 2.3, ayant été largement utilisée et validée empiriquement dans le domaine, offrant ainsi une approche fiable et un bon compromis entre résolution temporelle et atténuation des effets de bord.

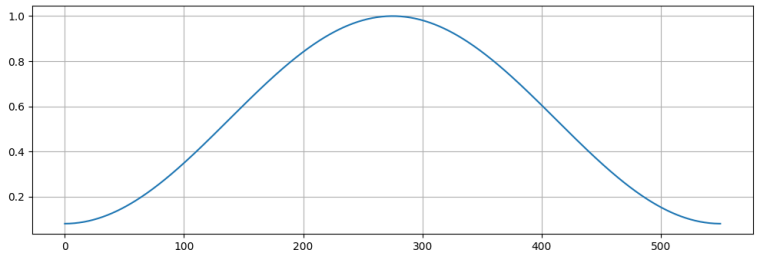
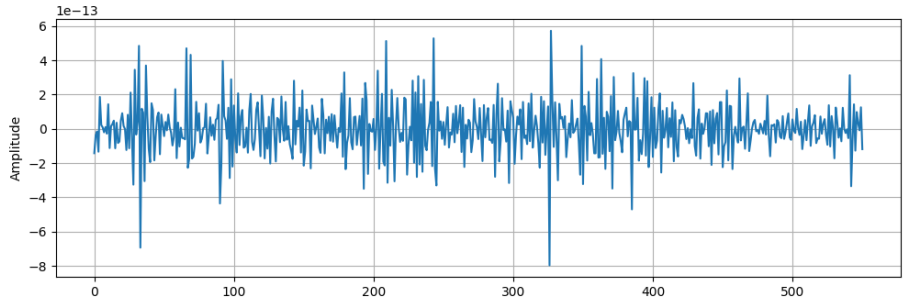


Fig. 2.3 : Fenêtre de Hamming.

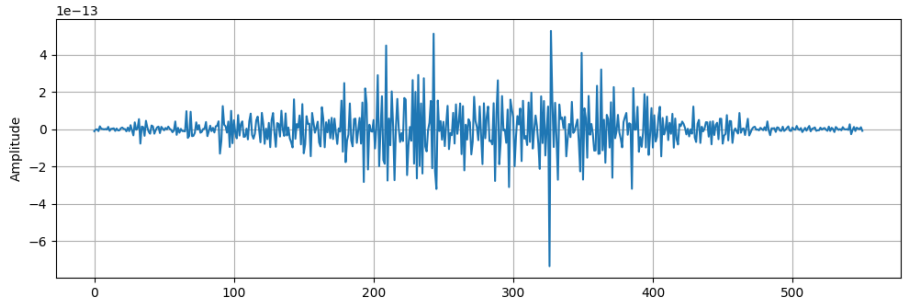
La fenêtre de *Hamming* est définie mathématiquement par la formule spécifique 2.2, qui permet de calculer les coefficients de la fenêtre pour chaque échantillon de la trame. Ces coefficients varient entre 0 et 1, et ils sont utilisés pour pondérer les échantillons du signal lors du fenêtrage. Il est à noter que **N** représente la taille de la fenêtre.

$$w(n) = 0.54 - 0.46 \cdot \cos\left(\frac{2\pi n}{N-1}\right) \tag{2.2}$$

La figure 2.4 illustre l'effet du fenêtrage en utilisant la fenêtre de hamming.



(a) Signal segmenté.



(b) Signal fenêtré.

FIG. 2.4 : Effet du fenêtrage.

2.2.1.4 Normalisation

La normalisation est une technique de prétraitement qui vise à éliminer les écarts d'échelle entre les différentes caractéristiques ou variables d'un ensemble de données, ce qui permet de les rendre comparables et de faciliter les analyses et les modélisations ultérieures. Il existe différents types de normalisation notamment :

- **Normalisation Min-max :** cette méthode utilise les valeurs minimale et maximale des données pour effectuer la mise à l'échelle dans une plage entre 0 et 1. Comme le montre l'équation :

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{2.3}$$

- **Normalisation Z-score :** également appelée standardisation, cette technique transforme les valeurs des données pour qu'elles aient une moyenne nulle et un écart type de 1. Sa formule mathématique est la suivante :

$$x_{norm} = \frac{x - \mu}{\sigma} \tag{2.4}$$

μ : moyenne
σ : l'écart-type

2.2.2 Techniques de prétraitement des images

Avant d'effectuer la classification, les images passent par les étapes de prétraitment indiquées sur la figure 2.5 qui sont la détection du visage, le recadrage, la transformation de l'image en niveaux de gris, la normalisation et enfin la detection des contours.

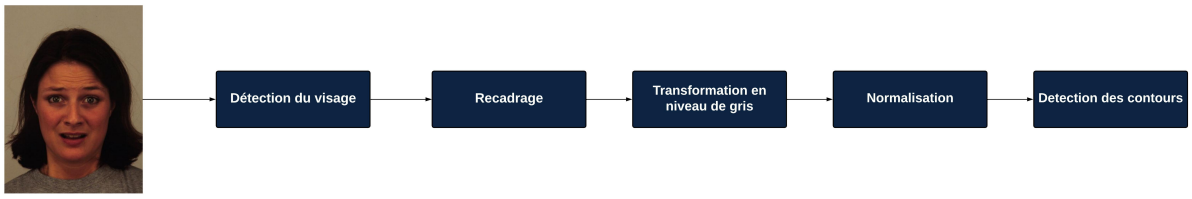


Fig. 2.5 : Étapes de prétraitement des images.

2.2.2.1 Détection du visage

L'algorithme de cascades de Haar, développé par Viola et Jones[36], est très utilisé pour la détection de visage dans de plusieurs systèmes de reconnaissance faciale des émotions. Il est basé sur des classifieurs en cascade qui sont entraînés à reconnaître les caractéristiques distinctives des visages. Les étapes de cette technique comprennent :

- La sélection des caractéristiques à l'aide de la méthode de Haar, qui utilise des filtres en forme de rectangle, présentés dans la figure 2.6 afin de capturer les variations d'intensité dans une image ;
- L'apprentissage des classificateurs en utilisant un ensemble d'exemples positifs et négatifs pour apprendre à distinguer les visages des arrière-plans non intéressants ;
- La création de la cascade en combinant plusieurs classificateurs en série permettant ainsi la réduction du nombre de fenêtres de recherche à analyser et par conséquent l'accélération de la détection et enfin la détection des visages.

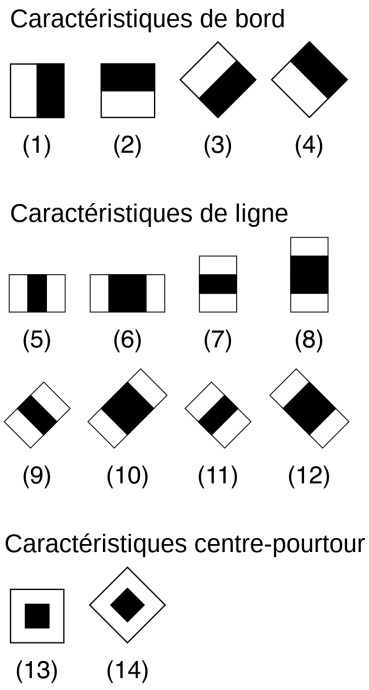


FIG. 2.6 : Exemples des caractéristiques de Haar

Le choix de cette méthode est motivé par sa vitesse de détection remarquable, permettant ainsi une analyse des visages avec une latence minimale tout en maintenant une précision de détection relativement élevée, comparable à celle d'algorithmes plus lents, ainsi que son taux de faux positifs très faible, ce qui signifie que les erreurs de détection sont minimisées ce qu'il garantit une meilleure précision dans l'identification des visages et des émotions. La figure 2.7 représente un exemple d'application de cet algorithme.

FIG. 2.7 : Visage détecté en appliquant l'algorithme de cascades de Haar.

2.2.2.2 Normalisation

Après avoir détecter les régions d'intérêt, il est important de redimensionner et normaliser les images lors de l'entraînement du modèle CNN. De plus, dans notre application, la classification ne repose pas sur les informations de couleur, ce qui signifie que les trois canaux RGB ne sont pas nécessaires. Ainsi, il est préférable de convertir les images en

niveaux de gris, ce qui permet de réduire à la fois le temps de calcul requis et le nombre de paramètres du modèle. La figure 2.8b montre l'image résultante à la fin de la chaîne de prétraitement.

(a) Image du visage détecté.

(b) Image en niveaux de gris.

FIG. 2.8 : Image résultante du recadrage et transformation en niveau de gris.

2.2.2.3 Détection de contours

La détection de contours est un processus basé sur des méthodes mathématiques, qui vise à identifier les zones d'une image où il y a des changements brusques d'intensité lumineuse. Ces changements de propriétés visuelles, tels que les variations de couleur ou de luminosité, peuvent indiquer des bordures et des contours significatifs dans l'image. Il existe un grand nombre de méthodes de détection des contours et notre choix s'est porté sur le filtre de **Canny** en raison de sa capacité à détecter efficacement la forme de la bouche, des yeux et des traits.

Le détecteur de contours de Canny, développé par *John F. Canny* et présenté dans son article de 1986 [37], repose sur quatre étapes clés. Tout d'abord, un filtre gaussien est utilisé pour réduire le bruit de l'image. Ensuite, le gradient de l'image est calculé pour déterminer la direction et l'amplitude des variations d'intensité. Les pixels présentant un fort gradient et une orientation correspondant à un contour sont conservés, tandis que les autres pixels sont supprimés. Afin d'éviter les faux positifs, une technique appelée suppression des non-maximaux est employée pour affiner les bords détectés. Par la suite, un seuillage à double seuil est appliqué pour classer les pixels en contours forts, contours faibles et non-contours. Enfin, une méthode de suivi des contours par seuillage hystérésis est utilisée pour relier les contours faibles aux contours forts, créant ainsi des contours continus et précis

(a) Image originale.

(b) Image de contours.

Fig. 2.9 : Détection de contours en appliquant le filtre de Canny.

2.3 Extraction des caractéristiques

L'objectif principal de l'extraction des caractéristiques est d'extraire des informations pertinentes et efficaces qui facilitent la tâche de la classification ultérieure. Dans cette section, nous allons explorer les concepts théoriques des différentes techniques d'extraction des caractéristiques utilisées dans notre projet pour les deux modalités : signaux de la parole et les images faciales.

2.3.1 Caractéristiques des signaux de la parole

Les caractéristiques des signaux de la parole utilisées dans la reconnaissance vocale des émotions peuvent être regroupées en deux catégories principales :

- **Caractéristiques prosodiques :** ce sont des caractéristiques qui se rapportent à l'aspect mélodique, rythmique et intonatif du signal vocal. Elles comprennent des mesures telles que la fréquence fondamentale (F0), l'énergie, la durée des segments, les variations d u signal de hauteur de voix (Pitch)... .
- **Caractéristiques spectrales :** cette catégorie se concentre sur les propriétés fréquentielles du signal, telles que la distribution des fréquences et les pics spectraux. Elles peuvent inclure des mesures telles que le spectre de puissance, le spectre de fréquence, les coefficients de Mel-Cepstral… .

Il convient de souligner que le domaine de la reconnaissance vocale des émotions propose un large éventail de caractéristiques pouvant être utilisées. Cependant, il n'existe pas de formule universelle pour choisir les caractéristiques les plus appropriées.

Dans ce qui suit, on présentera l'ensemble de caractéristiques qu'on a évalué dans le cadre de notre projet. Notre sélection de caractéristiques s'est basée sur une revue approfondie de la littérature existante, des connaissances théoriques et des expérimentations préliminaires.

2.3.1.1 Mel-Frequency Cepstral Coefficients (MFCCs)

Mel-Frequency Cepstral Coefficients (MFCCs) ou les coefficients cepstraux de fréquence en échelle mel en français, sont des caractéristiques largement utilisés dans le domaine du traitement de la parole.

Ces coefficients sont calculés sur la base de la capacité auditive humaine en approximant la réponse non linéaire du système auditif humain. Ce dernier ne perçoit pas les fréquences sur une échelle linéaire, mais plutôt sur une échelle logarithmique. Et donc le mel c'est l'échelle perceptive qui modélise cette perception non linéaire des fréquences.
Le calcul des MFCCs s'effectue suivant les étapes présentées dans la figure 2.10.

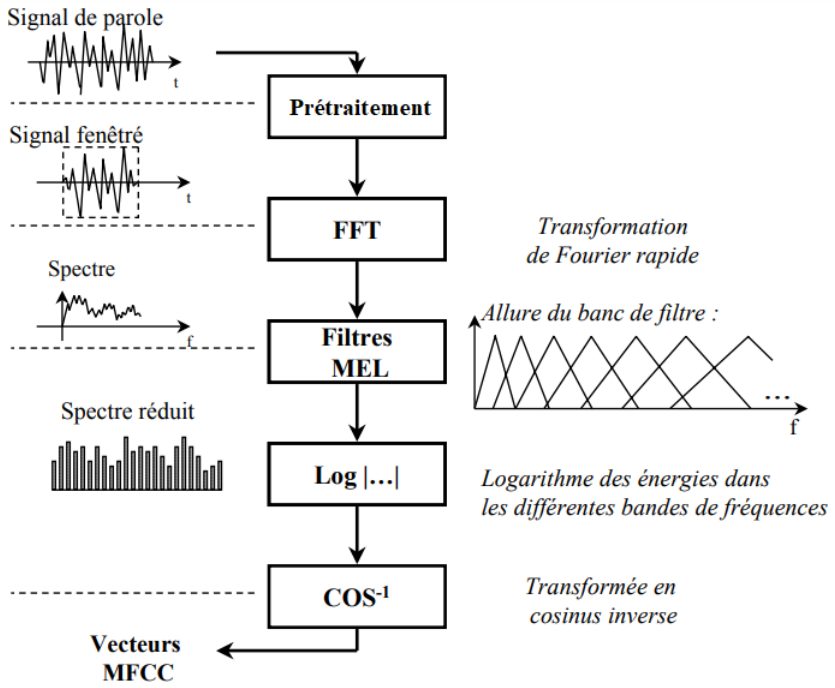


Fig. 2.10 : Étapes de calcul des MFCC

- **Conversion en domaine fréquentiel :** elle s'effectue en calculant la transformée de Fourier (FFT) pour chaque trame issue du prétraitement du signal vocal afin d'obtenir le spectre de puissance qui représente la distribution des fréquences dans la trame.

- **Banc de filtres :** ce sont des filtres triangulaires répartis sur l'échelle de Mel, appliqués au spectre de puissance de chaque trame pour représenter la sensibilité de l'oreille humaine aux différentes fréquences. L'équation mathématique pour convertir la fréquence normale **f** en échelle de Mel **m** est la suivante :

$$m = 2595log(1 + \frac{f}{700}) \tag{2.5}$$

- **Conversion logarithmique :** une fonction logarithmique est appliquée aux énergies spectrales obtenues par les filtres précédents pour transformer le modèle de

fréquence non linéaire de la perception humaine en échelle linéaire, facilitant ainsi l'inférence directe.

- **Transformation de cosinus discret (DCT) :** la DCT est appliquée aux valeurs logarithmiques afin d'obtenir les coefficients cepstraux de fréquence de Mel. La DCT réduit la dimensionnalité des données et saisit les caractéristiques les plus pertinentes du spectre de puissance.

2.3.1.2 Spectrogramme

Le spectrogrammes est une représentation visuelle de la distribution des fréquences d'un signal au fil du temps permettant de capturer des informations importantes et plus spécifiques sur les pics spectraux, la distribution de l'énergie dans différentes bandes de fréquences.
En utilisant le spectrogramme, il devient possible d'observer les transitions et les modulations de fréquence spécifiques qui sont associées à certaines émotions. Par exemple, une augmentation de l'énergie dans les hautes fréquences peut être associée à une expression de colère, tandis qu'une augmentation dans les fréquences graves peut indiquer une tristesse ou une mélancolie. La figure 2.11 présente un exemple d'un spectrogramme d'un signal vocal.

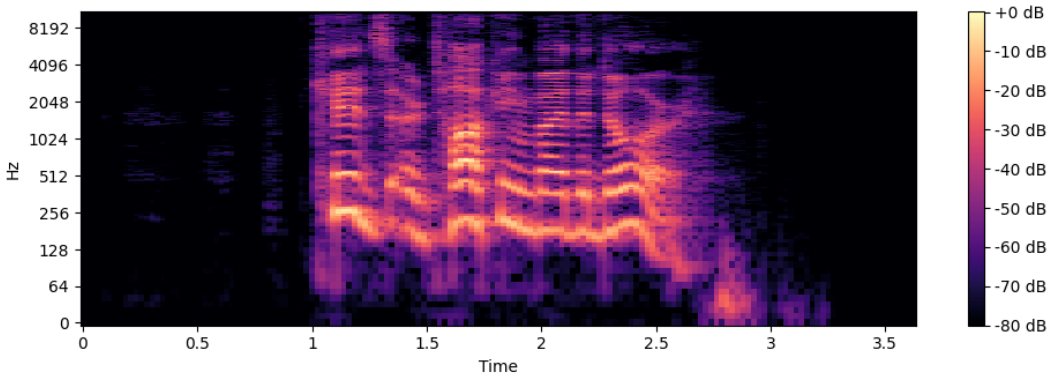


Fig. 2.11 : Spectrogramme

2.3.1.3 Spectrogramme de mel

Le spectrogramme de mel est une variante des spectrogrammes traditionnels basés sur l'échelle de Mel. La forme non linéaire de ce spectrogramme aide à mieux comprendre les émotions car les humains perçoivent le son sur une échelle logarithmique. Ainsi, le spectrogramme en log-mel correspond à la représentation du temps en fonction de la fréquence en log-mel, qui a été obtenue lors de l'étape 3 du calcul des MFCCs. La figure 2.12 permet une visualisation d'un spectrogramme de mel pour un signal donné.

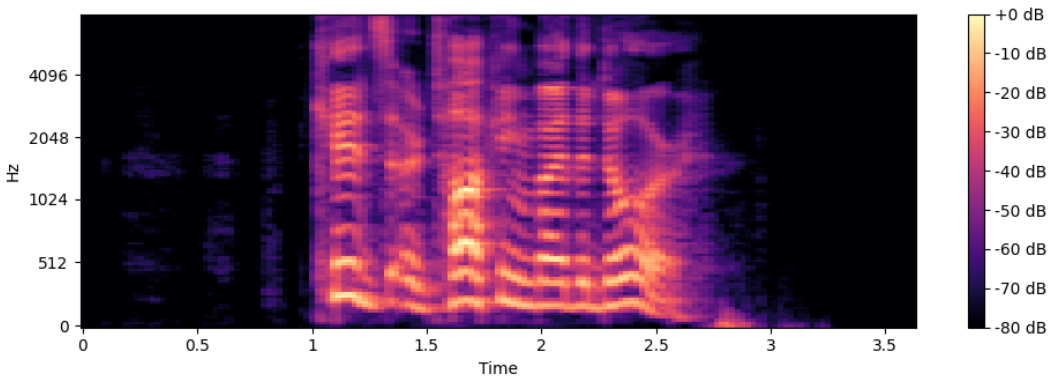


FIG. 2.12 : Spectrogramme de mel

2.3.1.4 Zero crossing rate (ZCR)

Il s'agit d'une mesure utilisée en traitement du signal audio qui capture les variations de signe ou de polarité de la forme d'onde de la parole et peut fournir des informations pertinentes permettant de différencier entre les émotions.
Par exemple, la colère peut être associée à des transitions rapides et des variations de signe importantes, tandis que la tristesse peut être caractérisée par des variations plus lentes et moins de transitions de signe.

2.3.1.5 Pitch

Le pitch, ou signal de hauteur de voix, correspond à la perception de variations ascendantes et descendantes dans le ton de la voix[38]. Il est influencé par la tension des cordes vocales et la pression de l'air qui les entoure, car il est généré par leurs vibrations. Le pitch est caractérisé par deux paramètres importants : la fréquence de vibration des cordes vocales, appelée fréquence fondamentale f_0, et la vitesse de l'air qui les traverse au moment de leur ouverture.

Un pitch élevé est généralement associé à un discours plus musical, joyeux et excité, tandis qu'un pitch plus bas peut être associé à des émotions telles que la tristesse, la colère ou la déception.

2.3.2 Caractéristiques des images

En vision par ordinateur, l'extraction de caractéristiques visuelles est une étape essentielle où des transformations mathématiques sont appliquées aux pixels d'une image numérique. Ces caractéristiques visuelles sont utilisées pour représenter de manière plus précise certaines propriétés visuelles de l'image afin de les exploiter dans des traitements ultérieurs tels que la reconnaissance faciales des émotions. Dans le cadre de notre projet, nous avons utilisé une technique appelée l'histogramme de gradient orienté afin d'extraire des informations pertinentes qui seront exploitées lors de la classification.

2.3.2.1 Histogramme de gradient orienté HoG

L'histogramme de gradient orienté ou Histogram of oriented gradient (HoG) est un descripteur de caractéristiques introduit par *Navneet Dalal* et *Bill Triggs* [39] en juin 2005 dans le cadre de leurs recherches sur la détection des piétons. L'idée essentielle derrière l'histogramme de gradient orienté est de décrire l'apparence locale et la forme d'un objet dans une image par la distribution de l'intensité du gradients ou de la direction des contours.
Les étapes de calcul du descripteur HoG impliquent la division de l'image en petites régions adjacentes appelées cellules. Pour chaque cellule, l'histogramme des directions du gradient ou des orientations des contours sont calculées en utilisant les pixels à l'intérieur de cette cellule. En combinant tous ces histogrammes, on obtient le descripteur HoG. Pour améliorer les résultats, les histogrammes locaux sont normalisés en contraste. Cela est réalisé en calculant une mesure de l'intensité sur des zones plus larges appelées blocs, et en utilisant cette valeur pour normaliser toutes les cellules du bloc. Cette normalisation permet d'obtenir une meilleure résistance aux changements d'illumination et aux ombres. La figure 2.13 présente un schéma simplifié des étapes du descripteur HoG.

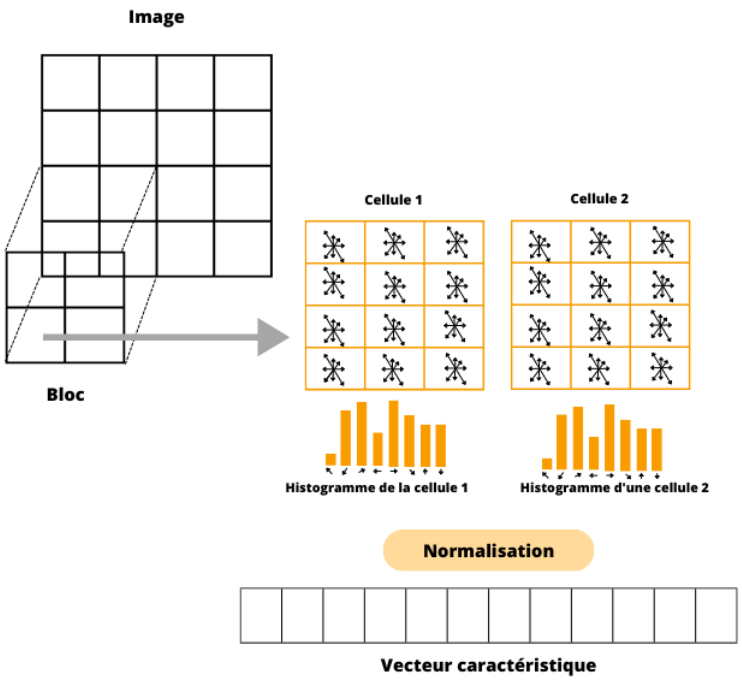


Fig. 2.13 : Schéma des étapes du descripteur HoG

2.4 Classification

Cette section est dédiée aux algorithmes utilisés dans la classification des émotions faciales et vocales. Nous avons établi deux approches. La première approche est basé sur l'algorithme d'apprentissage automatique SVM et pour la seconde approche nous avons sélectionné les réseaux de neurones convolutifs (CNN).

2.4.1 Algorithmes d'apprentissage automatique

2.4.1.1 Machines à vecteurs de support

Les machines à vecteurs de support ou Support Vector Machines (SVM) sont des modèles d'apprentissage automatique supervisés largement utilisés dans le domaine de la classification, y compris la reconnaissance automatique des émotions.

L'objectif d'un modèle SVM est de trouver une frontière qui sépare les classes dans l'espace des caractéristiques, de manière à maximiser la marge comme indiqué dans la figure 2.14, qui est la distance entre la frontière et les points de données les plus proches de chaque classe. Tous les points situés d'un côté de la frontière sont étiquetés comme 1, et tous les points situés de l'autre côté sont étiquetés comme -1. Les points les plus proches du plan de séparation de données sont appelés **vecteurs de support**.

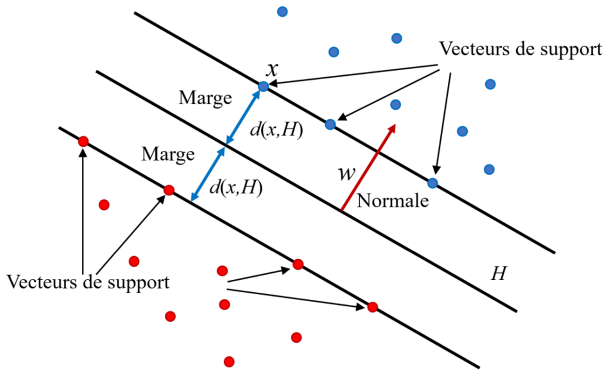


FIG. 2.14 : Schéma de machines à vecteurs de support

La machine à vecteur de support est caractérisé par les paramètres suivants :

- **Noyau :** Le noyau est une fonction qui permet de transformer les données d'un espace de dimension initiale vers un espace de dimension supérieure afin de rendre les données linéairement séparables dans l'espace de dimension supérieure, même si elles ne le sont pas dans l'espace de dimension initiale. Le choix du noyau dépend de la nature du problème et des caractéristiques des données. Les noyaux les plus utilisés sont : le noyau linéaire, le noyau RBF (Radial Basis Function) et le noyau polynomial.

- **Gamma :** Le paramètre *gamma* est spécifique aux noyaux non linéaires. Il détermine la portée d'influence des exemples d'entraînement sur la définition de la frontière de décision. Une valeur de *gamma* élevée signifie que les exemples d'entraînement les plus proches ont un poids élevé dans la classification, tandis qu'une valeur de *gamma* faible signifie que les exemples d'entraînement ont un poids plus uniforme.

- **Paramètre C :** Ce paramètre contrôle la pénalité associée à l'erreur de classification. Il détermine l'équilibre entre la maximisation de la marge et la minimisation des erreurs de classification sur les exemples d'entraînement. Une valeur plus élevée de C signifie que les erreurs de classification sont pénalisées de manière plus significative, ce qui conduit à un modèle qui cherche à classifier correctement autant d'exemples d'entraînement que possible.

2.4.2 Réseaux de neurones convolutifs

Les réseaux de neurones convolutifs, en anglais *Convolutional neural network* (CNN) est un algorithme d'apprentissage approfondi largement utilisé dans le domaine de l'intelligence artificielle, notamment dans les applications de vision par ordinateur telles que la classification d'images, la détection d'objets et la reconnaissance des expressions faciales. Les premières initiatives au CNN remontent à 1980, lorsque Fukushima [40] a introduit la carte auto-organisée, appelée "self-organizing map", pour extraire des caractéristiques à l'aide de l'apprentissage non supervisé. Ensuite, en 1998, Yann LeCun [41] a développé le premier réseau de neurones convolutif pour la reconnaissance des chiffres de l'écriture manuscrite. L'architecture proposée était elativement simple, composée de trois couches de convolution, deux couches de Pooling et deux couches entièrement connectées.

Le CNN a été développé dans le but d'automatiser l'extraction des caractéristiques des données, éliminant ainsi la nécessité de les extraire manuellement. Il effectue cette extraction en attribuant des poids et des biais mémorisables qui décrivent différents aspects de la donnée d'entrée, permettant ainsi de les distinguer les uns des autres.

Le CNN est composé de deux blocs principaux illustrés dans la figure 2.15. Chaque donnée d'entrée traverse d'abord le premier bloc, qui comprend des couches de convolution et des couches de Pooling. Cette combinaison permet d'extraire efficacement les caractéristiques de la donnée. Ensuite, la sortie de ce bloc est aplatie avant d'être transmise au deuxième bloc construit à partir de couches entièrement connectées, qui utilisent les caractéristiques extraites précédemment pour effectuer des classifications ou des prédictions.

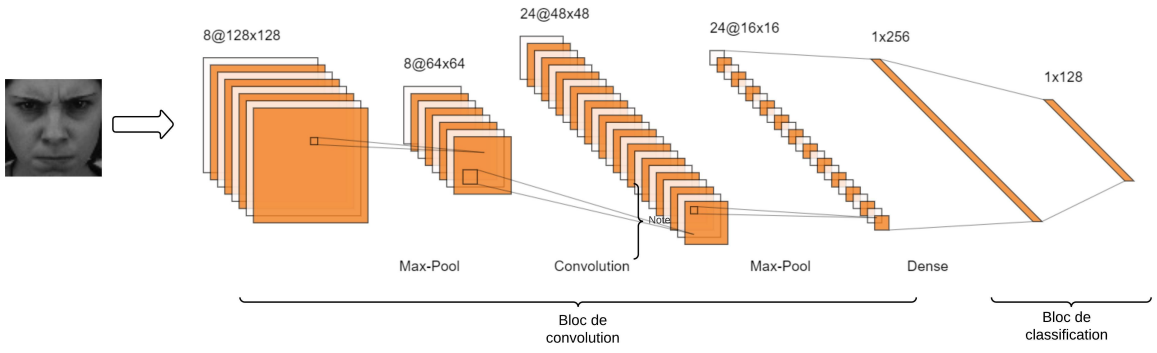


Fig. 2.15 : Architecture standard d'un CNN.

2.4.2.1 Couches de convolution (Convolution layer)

La couche de convolution est l'une des composantes clés d'un réseau de neurones convolutifs (CNN). Elle est responsable de l'extraction des caractéristiques importantes des données d'entrée. Cette couche effectue l'opération de la convolution qui consiste à effectuer le produit matricielle entre la matrice de l'image d'entrée et la matrice du filtre connu aussi comme le noyau dont la dimension est inférieure à celle de la matrice d'entrée. Cette opération est réalisée en superposant le filtre sur une région de l'image d'entrée et en effectuant une multiplication terme à terme entre les éléments de ce filtre et les éléments correspondants de la région d'entrée. Ensuite, les produits sont sommés pour obtenir une seule valeur, qui représente l'activation de la caractéristique détectée par le filtre dans cette région. Ces étapes sont représentées sur la figure 2.16.

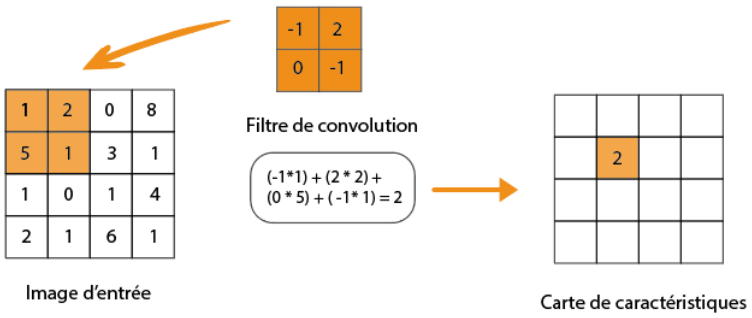


Fig. 2.16 : Exemple d'opération de convolution

2.4.2.2 Couches de pooling (Pooling layer)

Après une ou plusieurs couches de convolution, une couche de pooling est généralement ajoutée pour réduire la dimension spatiale des caractéristiques extraites tout en préservant les informations importantes.

La couche de pooling fonctionne en divisant la carte de caractéristiques en régions non disjointes, appelées fenêtres de pooling, et en appliquant une opération statistique sur chaque fenêtre pour produire une valeur agrégée. Les deux types de pooling les plus couramment utilisés sont le *max pooling* et le *average pooling*.

- **Max pooling :** pour chaque fenêtre de pooling, le max pooling sélectionne la valeur maximale parmi les activations présentes dans la fenêtre.
- **Average pooling :** à la place du maximum, l'average pooling calcule la moyenne des activations présentes dans chaque fenêtre de pooling.

La figure 2.17 représente un exemple d'opérations de max-pooling and average pooling.

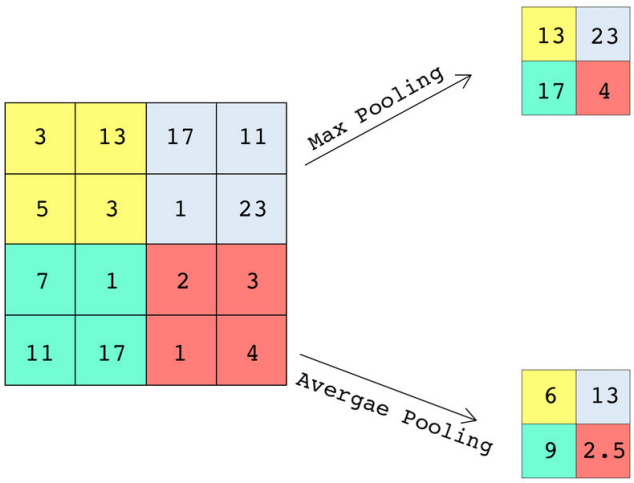


FIG. 2.17 : Pooling

2.4.2.3 Couches d'aplatissement (Flatten layer)

Cette couche se situe vers la fin du bloc convolutif qui sert à convertir les cartes de caractéristiques multidimensionnelles obtenues par les couches de convolution en un vecteur unidimensionnel, comme le montre la figure 2.18. Cela permet de connecter les caractéristiques spatiales extraites aux couches entièrement connectées du réseau. En résumé, la couche d'aplatissement facilite la transition entre les couches de convolution et les couches entièrement connectées en transformant les caractéristiques spatiales en un vecteur unidimensionnel pour le traitement ultérieur.

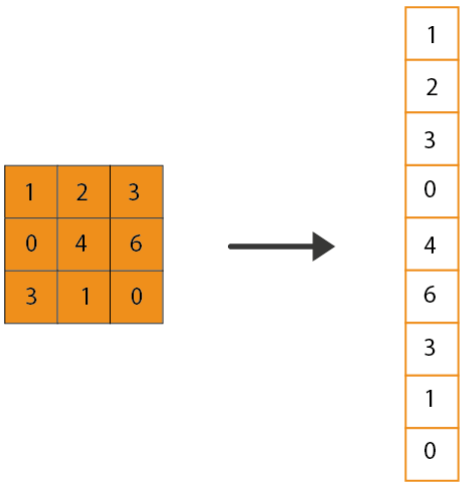


FIG. 2.18 : Opération d'aplatissement

2.4.2.4 Couches entièrement connectées (Fully connected layers)

Les couches entièrement connectées sont les éléments principaux du bloc de classification d'un CNN et ont pour objectif de prendre le vecteur fourni par la couche d'aplatissement en entrée et d'apprendre les meilleurs paramètres pour classifier l'image dans une classe spécifique. Plusieurs couches entièrement connectées sont généralement empilées pour renforcer l'apprentissage des caractéristiques et améliorer la précision des prédictions comme indiqué dans la figure 2.19. La dernière couche entièrement connectée fournit la sortie du CNN, représentée par un vecteur de probabilités correspondant aux différentes classes possibles. Les neurones de cette couche utilisent la fonction d'activation **softmax** pour fournir les probabilités estimées pour chaque classe, indiquant ainsi la probabilité avec laquelle l'image d'entrée appartient à chaque classe.

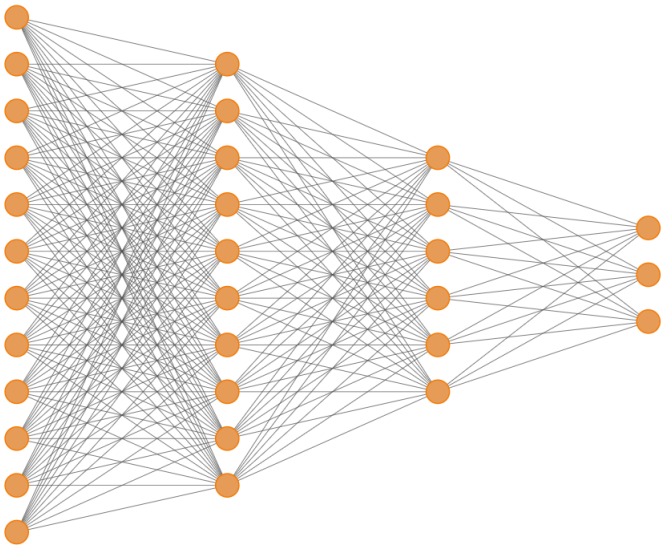


FIG. 2.19 : Couches entièrement connectées.

2.4.2.5 Paramètres d'un CNN

Filtres

Les filtres sont des matrices de poids utilisées dans les couches de convolution des CNN. Ils sont responsables de la détection de motifs spécifiques dans une image, tels que des contours, des textures ou des formes. Le nombre et la taille des filtres des couches de convolution dans un CNN sont des paramètres flexibles et adaptatifs, spécifiques au problème à résoudre, et leur détermination relève souvent d'un processus expérimental et itératif pour parvenir à une configuration optimale.

Stride

Le stride ou pas de déplacement, est un paramètre utilisé dans les couches de convolution. Il spécifie le pas de déplacement du filtre lors de l'opération de convolution sur l'image d'entrée en déterminant le nombre de pixels décalées horizontalement et verticalement à chaque déplacement.

Zero padding

Le zero padding est une opération qui consiste à ajouter des zéros autour de la matrice à l'entrée d'une couche de convolution. Cette opération permet de préserver la taille de la matrice résultante en sortie et d'éviter une réduction excessive de la dimension lors de la convolution.

Batch normalization

Le batch normalization ou normalisation par lot est une technique de régularisation appliquée lors de l'entraînement du CNN afin de stabiliser l'apprentissage du modèle en réduisant les variations de distribution des activations à chaque couche. Elle consiste à normaliser les vecteurs d'activation des couches cachées en utilisant la moyenne et l'écart type du lot courant.

Dropout

Le Dropout est une technique de régularisation largement utilisée dans les CNN. Il consiste à désactiver aléatoirement un certain pourcentage de neurones lors de l'entraînement du réseau. Cela permet d'éviter la sur-adaptation en réduisant la corrélation entre les neurones et en favorisant une meilleure généralisation du modèle.

Fonction d'activation

Les fonctions d'activation sont des éléments essentiels des CNN. Elles introduisent une non-linéarité dans le modèle, ce qui lui permet de saisir des relations complexes entre les données. Les fonctions d'activation prennent en entrée la somme pondérée des activations précédentes (ou des valeurs de pixels dans le cas de la première couche) et produisent une sortie non linéaire. Cette sortie est ensuite transmise à la couche suivante du réseau. Différentes fonctions d'activation sont utilisées dans les CNN, citons :

- **Rectified Linear Unit (ReLu) :**
  Elle remplace les valeurs négatives par zéro et laisse les valeurs positives inchangées. Elle est simple, rapide favorise la convergence de l'apprentissage.

$$F(z) = \max\{0, z\} \tag{2.6}$$

- **Softmax :**
  elle est généralement utilisée dans les couches de sortie des CNN pour effectuer une classification multi-classes. Elle transforme les valeurs en une distribution de probabilités, où la somme des valeurs est égale à 1. Elle permet de déterminer la classe la plus probable parmi plusieurs options.

$$F(z) = \frac{\exp(z_i)}{\sum_j \exp(z_i)} \tag{2.7}$$

2.5 Conclusion

Dans ce chapitre, nous avons exposé les différentes étapes et approches adoptées dans notre projet afin de garantir une compréhension claire de la méthodologie adoptée dans le prochain chapitre. Tout d'abord, nous avons décrit en détail les étapes de prétraitement des deux modalités. Ensuite, nous avons évoqué de manière approfondie les caractéristiques extraites des données. Enfin, nous avons fourni un aperçu général des algorithmes de classification utilisés, notamment le SVM (Support Vector Machine) et les réseaux de neurones convolutifs.

Dans le chapitre suivant, nous allons mettre en pratique les concepts introduits pour le développement d'un système de reconnaissance des émotions et discuter les résultats obtenus.

Chapitre 3

Méthodologie et résultats expérimentaux

3.1 Introduction

Ce chapitre est dédié à l'explication de la méthodologie adoptée et l'analyse des résultats expérimentaux obtenus. Tout d'abord, nous commençons par l'introduction des différentes bases de données utilisées, les métriques d'évaluations ainsi que l'ensemble de logiciels employés dans le cadre de ce projet. Ensuite, nous décrivons le protocole expérimental mis en place pour développer un système de reconnaissance de six émotions(joie, tristesse, colère, dégoût, peur, neutralité) qui est basé sur les données audiovisuelles. Les résultats de chaque modalité sont présentés et discutés, suivis d'une exploration de la fusion de scores et d'une étude comparative des performances atteintes.

3.2 Ensembles de données

Les bases de données disponibles pour la reconnaissance automatique des émotions sont regroupées en trois catégories en fonction du type d'émotions exprimées :

- **Base de données des émotions artificielles** : ses sujets sont des artistes professionnels et formés. Le processus de la collecte est très facile et garantit une grande variété d'émotions mais elles restent artificielles.
- **Base de données des émotions suscitées** : ses données sont collectées en créant une situation émotionnelle artificielle. Les données sont très proches de la nature, cependant les émotions peuvent ne pas être disponibles
- **Base de données des émotions naturelles** : elle est constituée de données réelles, complètement naturelles et fortement appropriées pour la reconnaissance des émotions dans des situations réelles.

En raison du faible nombre des échantillons des ensembles de données disponibles pour la reconnaissance des émotions humaines, nous avons travaillé avec différentes bases de données ce qu'il a permis à notre modèle d'enrichir son apprentissage afin qu'il soit performant pour des prédictions en temps réel. Les bases de données utilisées pour la reconnaissance vocale des émotions sont SAVEE, TESS, EMO DB et RAVDESS. Les images de la classification des émotions faciales appartiennent aux deux ensembles CK+ et KDEF et RAVDESS.

3.2.1 SAVEE

La base de données Surrey Audio-Visual Expressed Emotion (SAVEE) [42] a été enregistrée comme condition préalable au développement d'un système de reconnaissance automatique des émotions. Elle se compose d'enregistrements de 4 acteurs masculins britanniques exprimant sept émotions différentes : la colère, le dégoût, la peur, le bonheur, la tristesse, la surprise et la neutralité. Les données ont été enregistrées dans un laboratoire de médias visuels avec un équipement audiovisuel de haute qualité, puis traitées et étiquetées.

3.2.2 TESS

La base de données Toronto emotional speech set (TESS)[43] est composée de 2800 fichiers audio représentant sept émotions : la colère, le dégoût, la peur, la joie, la surprise agréable, la tristesse et la neutralité, enregistrés par de deux femmes âgées de 26 et 64 ans.

3.2.3 EMO DB

La base de données EMO-DB est la base de données émotionnelle allemande librement accessible. Elle a été créée par l'Institut des sciences de la communication de l'Université technique de Berlin, en Allemagne. Cinq hommes et cinq femmes ont participé à l'enregistrement des données. La base de données contient un total de 535 énoncés comprenant sept expressions qui sont la colère, l'ennui, l'anxiété, le bonheur, la tristesse, le dégoût et la neutralité.

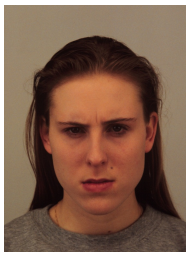
3.2.4 CK+

La base de données The Extended Cohn-Kanade (CK+) [44] contient 593 séquences vidéo dont 327 ont été étiquetées dans l'une des sept catégories d'expressions suivantes : colère, mépris, dégoût, peur, joie, tristesse et surprise. L'âge de ses sujets varie entre 18 et 50 ans parmi eux, 69% sont des femmes, 81% sont des Euro-Américains et 13% sont des Afro-Américains. La figure 3.1 illustre quelques exemples des six émotions de cette base de données.

Fig. 3.1 : Examples de six expressions de CK+ : A) Colère, B) Dégoût, C) Peur, D) Joie, E) Tristesse,) Surprise.

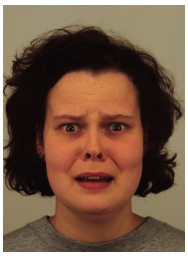
3.2.5 KDEF

Karolinska Directed Emotional Faces (KDEF) [45] est une base de données produite en 1998 et mise gratuitement à la disposition de la communauté des chercheurs. Elle comprend 4900 images représentant les six émotions basiques humaines : colère, dégoût, peur, joie, tristesse, surprise, ainsi que l'état neutre illustrées dans la figure 3.2. Cette base de données regroupe 70 acteurs amateurs, soit 35 femmes et 35 hommes, âgés de 20 à 30 ans. Chaque expression a été capturée sous cinq angles différents, cependant, pour notre application, nous avons utilisé uniquement les visages en position frontale.

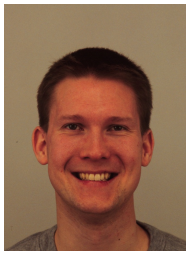


(a) Colère.

(b) Dégoût

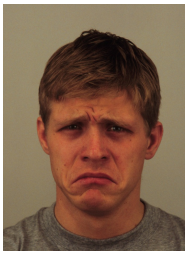


(c) Peur.



(d) Joie.

(e) Neutralité.

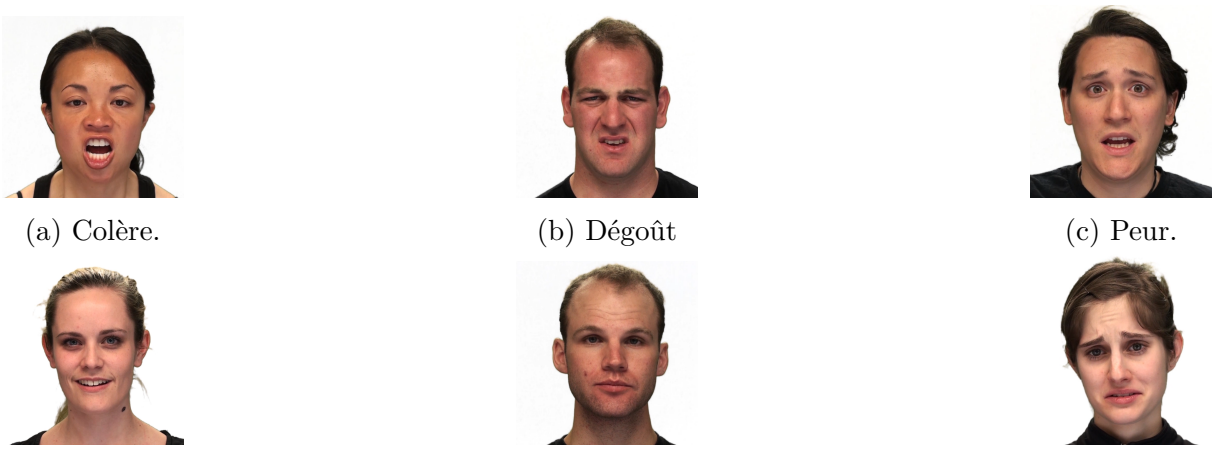


(f) Tristesse.

Fig. 3.2 : Échantillons des images de la base de données KDEF.

3.2.6 RAVDESS

The Ryerson Audio-Visual Database of Emotion and Song (RAVDES)[46] est une base de données multimodale validée de discours et de chants émotionnels. La base de données est équilibrée en matière de genre et se compose de 24 acteurs professionnels qui vocalisent des énoncés lexicalement appariés avec un accent nord-américain neutre. La parole comprend des expressions neutres, joyeuses, tristes, colériques, craintives, de surprise et de dégoût. Elle est composée de 7356 fichiers dont seulement 1440 sont des fichiers audio de parole. Elle est disponible sur trois formats : Audio seulement (.wav), Audiovidéo (.mp4), et Vidéo seulement sans son. La figure 3.3 présente des exemples des images prises de RAVDESS.



(a) Colère. (b) Dégoût (c) Peur.

(d) Joie. (e) Neutralité. (f) Tristesse.

FIG. 3.3 : Échantillons des images de l'ensemble de données RAVDESS.

3.2.7 Distributions des échantillons

Les figure 3.4 et 3.5 fournissent une visualisation de nombre des échantillons contenus dans chaque classe (émotion) pour les bases de données des signaux vocaux et des expressions faciales respectivement.

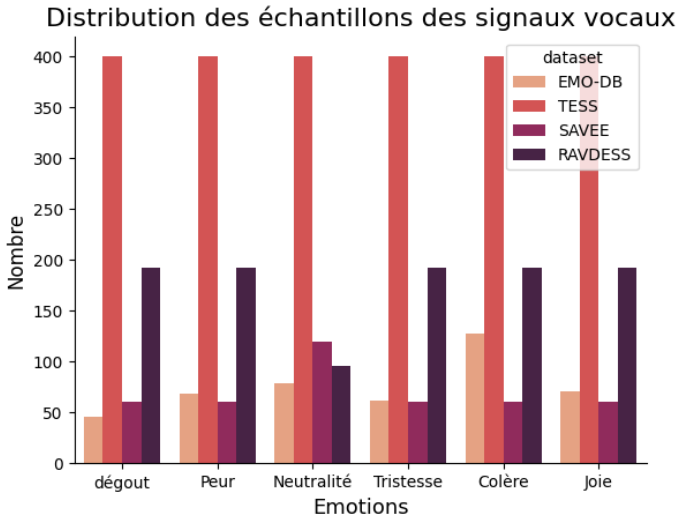


FIG. 3.4 : Distribution des échantillons de chaque base de données des signaux vocaux pour les six émotions : Colère, dégoût, joie, neutralité,peur et tristesse

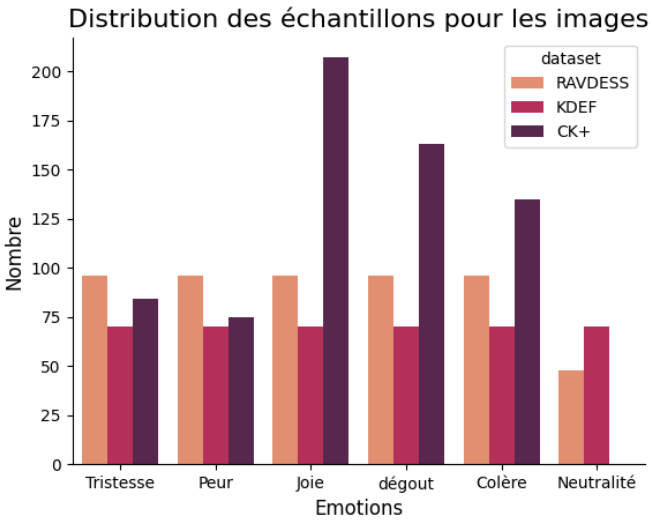


FIG. 3.5 : Distribution des échantillons de chaque base de données des images faciales pour les six émotions : Colère, dégoût, joie, neutralité,peur et tristesse

3.3 Logiciels et librairies

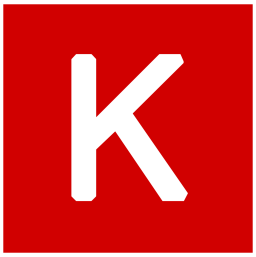
3.3.1 Python

Python est un langage de programmation interprété, orienté objet et de haut niveau. Sa syntaxe claire et sa grande flexibilité en font un choix populaire dans de nombreux domaines tels que le développement web, l'analyse de données, l'apprentissage profond et l'intelligence artificielle. Il prend en charge la modularité grâce aux modules et aux packages, ce qui facilite la réutilisation du code. Python est disponible gratuitement sur toutes les principales plates-formes, ce qui en fait un choix accessible et largement adopté.

3.3.2 Tensorflow

TensorFlow est une plateforme open source d'apprentissage automatique développé par Google. Elle permet de créer, entraîner et déployer des modèles d'apprentissage automatique, en particulier des réseaux de neurones, pour diverses applications. TensorFlow offre une grande flexibilité et prend en charge le calcul sur des structures de données multidimensionnelles appelées tenseurs. Il est largement utilisé pour ses fonctionnalités avancées, notamment la distribution des calculs sur plusieurs processeurs ou GPU, et est devenu une bibliothèque populaire dans le domaine de l'apprentissage automatique.

3.3.3 Keras



Keras est une bibliothèque open source d'apprentissage automatique. Intégrée à TensorFlow, elle offre une interface haut niveau et abstraite pour créer, entraîner et déployer des réseaux de neurones profonds. Elle fournit des fonctionnalités conviviales pour construire des modèles d'apprentissage automatique, tout en offrant une flexibilité pour leur la personnalisation et leur ajustement.

3.3.4 Scikit-Learn

Scikit-learn est une bibliothèque Python open source dédiée à l'apprentissage automatique. Elle offre une gamme d'outils pour effectuer des tâches telles que la classification, la régression, le regroupement et la réduction de la dimensionnalité, entre autres. Scikit-learn est conçue pour être compatible avec d'autres bibliothèques Python populaires, notamment NumPy et Pandas.

3.3.5 Librosa

Librosa est une bibliothèque Python dédiée au traitement de la musique et du signal audio . Elle propose des fonctionnalités avancées pour le chargement, l'analyse et la manipulation de données audio. Grâce à Librosa, il est possible d'extraire efficacement des caractéristiques audio, telles que les spectrogrammes ou les coefficients MFCC, essentiels pour l'analyse et l'apprentissage automatique. Cette bibliothèque est largement utilisée par les professionnels de l'audio et les chercheurs travaillant dans le domaine du traitement du signal audio.

3.3.6 OpenCV

OpenCV (Open Source Computer Vision Library) est une bibliothèque open source largement utilisée pour le traitement d'images et la vision par ordinateur. Elle propose des fonctionnalités avancées pour la manipulation, l'analyse et la compréhension des images et des vidéos. OpenCV est écrite en C++ mais offre également des

interfaces pour les langages de programmation tels que Python et Java. Cette bibliothèque est utilisée dans de nombreux domaines, tels que la reconnaissance faciale, la détection d'objets, la réalité augmentée, la vision industrielle, et bien d'autres. OpenCV est réputée pour sa performance élevée, sa large gamme d'algorithmes et sa grande compatibilité avec diverses plateformes.

3.3.7 Google Colaboratory

Google colaboratory, souvent raccourcie en "Colab" , est une plateforme en ligne gratuite proposée par Google. elle est basée sur des notebooks Jupyter qui des interfaces interactives où les utilisateurs peuvent écrire, exécuter et partager du code, ainsi que visualiser les résultats.

Colab permet aux utilisateurs d'accéder à ces notebooks Jupyter directement depuis leur navigateur web, sans nécessiter d'installation locale. Il offre également des fonctionnalités avancées telles que l'accès à des ressources de calcul, notamment les GPU et les TPU, pour accélérer le traitement des tâches intensives en calcul, comme l'apprentissage automatique.

3.4 Métriques d'évaluation

3.4.1 Exactitude (*Accuracy*)

La métrique la plus simple pour l'évaluation des modèles est la précision. Il s'agit du rapport entre le nombre de prédictions correctes et le nombre total de prédictions effectuées pour un ensemble de données.

$$\textit{Exactitude (\%)} = \frac{\textit{Nombre de prédictions correctes}}{\textit{Nombre total de prédictions effectuées}} \times 100 \tag{3.1}$$

3.4.2 Matrice de confusion

La matrice de confusion est le paramètre le plus utilisé lors de l'évaluation des performances d'un modèle d'apprentissage automatique. C'est une matrice carrée dont les dimensions dépendent de nombre de classes. Chaque ligne représente une classe réelle et chaque colonne représente une classe estimée. La case (L,C) contient le nombre d'éléments de la classe réelle L qui ont été estimés appartenir à la classe C. Elle permet une bonne visualisation de la distribution des étiquettes prédites. La figure 3.6 contient un exemple d'une matrice de confusion de classification binaire.

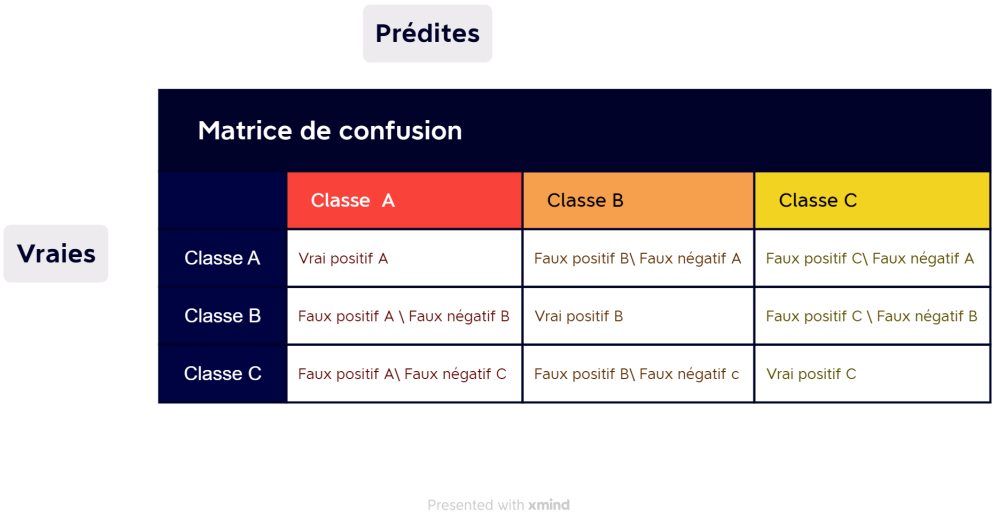


FIG. 3.6 : Modèle de la matrice de confusion

- **Vrai positif (VP) :**Nombre d'échantillons qui sont réellement positifs et qui sont prédits positifs.
- **Vrai négatif (VN) :** Nombre d'échantillons qui sont réellement négatifs et qui sont prédits négatifs.
- **Faux positif (FP) :** Nombre d'échantillons qui sont en réalité négatifs mais prédits positifs. Ces erreurs sont également appelées erreurs de type 1.
- **Faux négatif (FN) :** Nombre d'échantillons qui sont en fait positifs mais prédits négatifs. Ces erreurs sont également appelées erreurs de type 2.

3.4.3 Précision

La précision mesure la proportion d'instances correctement classées parmi les instances prédites comme positives pour une classe spécifique. En d'autres termes, elle indique à quel point le modèle est précis lorsqu'il prédit une classe particulière. Elle est calculée par la formule suivante :

$$Prcision_i = \frac{Nombre\ d'instances\ correctes\ pour\ une\ classe_i}{Nombre\ total\ d'instances\ attribues\ \ la\ classe_i} \times 100 \tag{3.2}$$

Pour le rappel dans le cadre d'une classification de n classes :

$$Prcision = \frac{\sum_{i=1}^{n} \text{précision}_i}{n} \times 100 \tag{3.3}$$

3.4.4 Rappel

Le rappel, également appelé sensibilité ou taux de vrais positifs, mesure la proportion d'instances correctement classées parmi toutes les instances réelles positives pour une classe spécifique. En d'autres termes, il indique à quel point le modèle est capable de rappeler ou de capturer les instances réelles positives. Il est calculé par la formule suivante :

$$Rappel_i = \frac{Nombre\ d'instances\ correctes\ pour\ une\ classe_i}{Nombre\ total\ d'instances\ appartenant\ \ la\ classe_i} \times 100 \tag{3.4}$$

Pour le rappel dans le cadre d'une classification de n classes :

$$Rappel = \frac{\sum_{i=1}^{n} \text{rappel}_i}{n} \times 100 \tag{3.5}$$

3.5 Protocole expérimental global

Notre travail repose sur la réalisation et l'implémentation d'un système automatique capable de reconnaître l'émotion exprimée par un individu à travers son expression faciale et sa voix en temps réel.
Afin d'arriver à notre objectif, nous avons suivi un processus présenté en 3.7. Le système est divisé en deux parties majeures : la reconnaissance faciale des émotions et la reconnaissance vocale des émotions. Nous avons proposé pour chaque modalité un algorithme d'apprentissage automatique qui est le SVM et un autre d'apprentissage approfondi qui est le CNN et nous allons choisir celui qui fournit les meilleurs résultats tout en respectant la contrainte du coût et du temps de calcul. À la fin, nous allons effectuer une fusion des scores (prédictions) des deux modalités en utilisant des opérations arithmétiques simples pour arriver à une décision finale. Le système sera implémenté sur une carte Raspberry Pi et testé en temps réel.

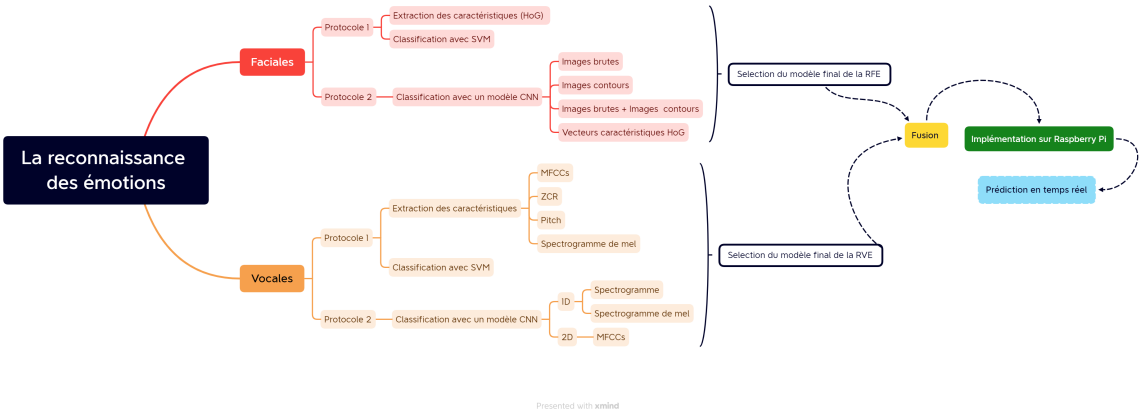


Fig. 3.7 : Schéma récapitulatif de la démarche de notre projet.

3.5.1 Entraînement du CNN

Les entraînements sont réalisés sous l'environnement Python, utilisant le framework **Tensorflow**. Les conditions d'entraînement établies sont les suivantes :

- **Subdivision de l'ensemble de données :** L'ensemble de données a été subdivisé sur 3 sous-ensemble : entraînement, validation et test, avec les pourcentage de 80%, 10% et 10% respectivement.
- **Fonction de coût :** La fonction de coût utilisée est la "*categorical cross-entropy*", ou entropie croisée catégorique. Ce choix est approprié car le problème est un problème multi-classe, où les classes sont encodées à l'aide d'un encodage one-hot.
- **Taux d'apprentissage :** On a utilisé un *callback* de mise à jour du taux d'apprentissage avec une réduction d'un facteur de 0.1 si le coût ne diminue pas au bout de 3 époques consécutives.
- **Fonction d'optimisation :** Adam.

3.6 Performances atteintes

3.6.1 Résultats de la reconnaissance vocale des émotions

Pour cette première modalité, nous avons proposé deux protocoles. En premier lieu, nous effectuons une classification avec un algorithme d'apprentissage automatique (SVM) en évaluant une combinaison de caractéristiques d'un signal audio.
En deuxième lieu, nous développons deux modèles CNN : le premier est basé sur une architecture CNN à une seule dimension (CNN 1D) qui reçoit en entrée le vecteur des caractéristiques extraites, tandis que le deuxième est un CNN à deux dimensions (CNN 2D) qui reçoit en entrée les spectrogrammes des signaux audio.

3.6.1.1 Protocle 1 : Classification en utilisnat le SVM

Ce protocole est basé sur la classification des émotions avec le SVM qui est un algorithme d'apprentissage automatique simple et largement utilisé. Son implémentation sur l'environnement python a été réalisée à l'aide de la librairie **scikit-learn** mentionnée précédemment.
Dans nos premiers tests, nous sommes contentés de l'utilisation d'une seule base de données qui est RAVDESS, en éliminant les signaux audio des émotions (surprise et calme). Les données sont ensuite subdivisées sur deux ensembles comme le montre le tableau suivant 3.1 :

Tab. 3.1 : Répartition des différents ensembles : entraînement et test

Ensembles	Pourcentage (%)	Nombre d'audio
Entraînement	80	844
Test	20	212

Effet des techniques de prétraitement

Avant de présenter les résultats d'évaluation des différentes caractéristiques, nous tenons à mettre en évidence l'effet des techniques de prétraitement utilisées sur les performances de la RVE. Pour cela nous avons fixé les MFCCs comme caractéristiques, avec le noyau Rbf du SVM. Les résultats obtenus sont illustrés dans le tableau ci-dessous 3.2 :

TAB. 3.2 : Effet des prétraitements

Technique de prétraitement	Exactitude (%)	Précision (%)	Rappel (%)
Aucun	30	19	30
Filtre de préaccentuation	28	30	28
Normalisation	64	63	64
Normalisation + préaccentuation	67	68	67

Les résultats du tableau démontrent clairement l'effet positif des techniques de prétraitement sur les performances du système. L'application de la normalisation seule a conduit à des améliorations significatives avec une augmentation de l'exactitude de 30% à 64%, tandis que la combinaison des deux techniques a donné les meilleurs résultats en termes d'exactitude, de précision et de rappel. Cela s'explique par le fait que la normalisation réduit les variations de valeurs entre les caractéristiques, alors que le filtre de préaccentuation améliore la résolution et l'accentuation des caractéristiques vocales pertinentes.

Par conséquent, tous les autres tests effectués au cours de cette section ont été réalisés après l'application de ces prétraitements.

Évaluation des caractéristiques

Comme nous l'avons précédemment mentionné dans le deuxième chapitre, la sélection des caractéristiques ne peut être déterminée de manière exacte, mais plutôt par des tests pratiques. Ainsi, notre premier objectif était de choisir les caractéristiques les plus pertinentes.

Dans ce projet, nous avons d'évalué quatre types de paramètres, à savoir les MFCCs, le ZCR, les coefficients du spectrogramme de mel (Mel) et le pitch. D'abord, nous avons effectué des évaluations individuelles de chacune de ces caractéristiques, puis nous avons exploré différentes combinaisons possibles tout en variant le noyau du SVM entre RBF et linéaire comme le montre de tableau 3.3. Pour extraire ces caractéristiques, nous avons utilisé la bibliothèque Python **Librosa**.

Tab. 3.3 : Résultats d'évaluation des caractéristiques

Caractéristiques	Noyaux	Dimension du vecteur	Exactitude (%)	Précision (%)	Rappel (%)
MFCCs	rbf	13	71	72	71
	linéaire	13	50	49	50
	Rbf	65	73	74	73
	linéaire	65	60	63	60
Pitch	Rbf	228	42	49	42
	linéaire	228	28	29	28
mel	Rbf	128	39	42	39
	linéaire	128	37	37	37
ZCR	Rbf	228	42	44	42
	linéaire	228	29	30	29
mel, MFCCs	Rbf	193	61	62	61
	linéaire	193	66	67	66
Pitch, MFCCs	Rbf	293	59	59	59
	linéaire	293	51	52	51
ZCR, MFCCs	Rbf	293	65	66	65
	linéaire	293	58	59	58
Pitch, mel, MFCCs	Rbf	412	61	62	61
	linéaire	412	56	56	56
Pitch, mel, ZCR	Rbf	584	59	61	59
	linéaire	584	57	60	57
Pitch, ZCR, mel, MFCCs	Rbf	649	56	58	56
	linéaire	649	59	61	59

Réduction de dimensionnalité : Dans le but d'optimiser les ressources et le temps de calculs, nous avons appliqué une méthode statistique de réduction de dimensionnalité appelée **analyse en composantes principales (ACP)** avant de fournir les vecteurs d'entrée à notre classifieur. L'ACP permet donc de transformer un ensemble de variables d'origine en un nouvel ensemble réduit de variables non corrélées, appelées composantes principales.

Tab. 3.4 : Résultats après l'application de l'ACP

Caractéristiques	Noyaux	Dimension du vecteur	Exactitude (%)	Précision (%)	Rappel (%)
Sans ACP					
Pitch, Mel, MFCCs	Rbf	412	61	62	61
Pitch, ZCR, Mel, MFCCs	Rbf	649	56	58	56
Avec ACP					
Pitch, Mel, MFCCs	Rbf	310	61	62	61
Pitch, ZCR, Mel, MFCCs	Rbf	400	56	58	56

- Donc, l'ACP identifie les directions principales le long desquelles les données varient le plus, permettant ainsi de compresser les informations tout en préservant les caractéristiques les plus significatives comme le montre les résultats illustrés dans le tableau 3.4.

- En analysant les résultats du tableau d'évaluation des caractéristiques 3.3, il est évident que les MFCCs avec une dimension de 65, qui incluent des statistiques telles que la moyenne, la variance, l'écart-type, le minimum et le maximum des coefficients tout au long du signal, offrent les meilleures performances en termes d'exactitude, de précision et de rappel lorsqu'ils sont combinés avec le noyau RBF du SVM. Par conséquent, pour le reste de ce travail, nous continuons nos tests en utilisant les MFCCs comme caractéristiques principales.

Augmentation des données

Après avoir déterminé les techniques de prétraitement à utiliser, les caractéristiques significatives et les paramètres de la SVM, nous avons décidé d'augmenter le nombre d'échantillons en utilisant d'autres bases de données en plus de $RAVDESS$. Nous avons choisi d'utiliser à la fois la base de données $SAVEE$ et $TESS$. Cette approche nous permet de collecter des échantillons vocaux à la fois masculins et féminins, ce qui évite de rendre notre modèle spécifique à un seul genre et permet d'améliorer sa généralisation et sa robustesse de notre modèle.

Le tableau 3.5 illustre la nouvelle répartition des échantillons en deux ensembles : entraînement et test. Tandis que le tableau 3.6 représente une comparaison des performances de la SVM après et avant l'augmentation des données.

Tab. 3.5 : Nouvelle répartition des différents ensembles : entraînement et test

Ensembles	Pourcentage (%)	Nombre d'audio
Entraînement	80	3100
Test	20	776

Tab. 3.6 : Tableau comparatif des performances des modèle avec 1BDD et 3BDD

Nombre de base de données	Exactitude (%)	Précision (%)	Rappel (%)
1	73	74	73
3	88	88	88

La figure 3.8 représente la matrice de confusion de la classification des six émotions.

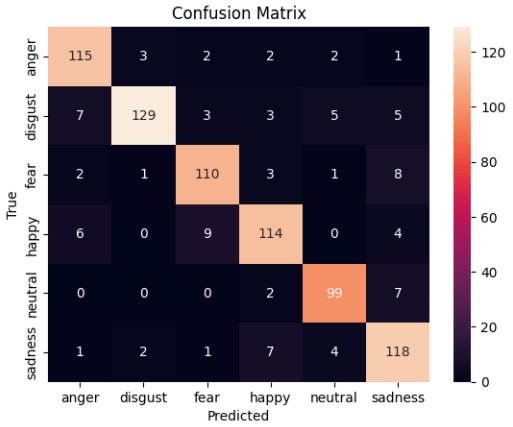


FIG. 3.8 : Matrice de confusion

3.6.1.2 Protocole 2 : Classification en utilisant le CNN

Dans ce deuxième protocole, nous avons opté pour l'utilisation de réseaux de neurones convolutifs (CNN) afin de mieux exploiter les informations extraites des signaux audio et d'améliorer les performances de notre système.

Pour cela nous avons adopté deux approches différentes : La première consiste à réaliser notre propre architecture $CNN1D$ qui traite les vecteurs contenant les coefficients MFCCs d'un signal audio. Tandis que la deuxième repose sur la création d'un $CNN2D$ qui reçoit à son entrée la représentation en 2D du signal, appelée le spectrogramme.

CNN 1D :

L'architecture de ce modèle est constitué de trois couches de convolution, deux couches de Pooling, suivies par une couche d'aplatissement et enfin 2 couches entièrement connectées responsables de la classification.

Après chaque couche de convolution 1D, une couche de normalisation par lot (batch normalization) est appliquée pour normaliser la sortie de chaque couche, afin d'éviter des distributions variables des caractéristiques entre les données d'entraînement et de test. Nous avons adapté le dropout afin d'éviter le problème du surapprentissage.

L'architecture détaillée est illustré dans le tableau suivant 3.7 :

Tab. 3.7 : Architecture du modèle proposé (CNN 1D)

Type	Forme	Paramètres
Convolution 1D	(None, 65, 256)	1536
Normalisation par lot	(None, 65, 256)	1024
Convolution 1D	(None, 65, 128)	163968
Pooling	(None, 33, 128)	0
Normalisation par lot	(None, 33, 128)	512
Convolution 1D	(None, 33, 64)	41024
Normalisation par lot	(None, 33, 64)	256
Pooling	(None, 17, 64)	0
Dropout	(None, 17, 64)	0
Couche d'Applatissement	(None, 1088)	0
Couche entièrement connectée	(None, 128)	139392
Dropout	(None, 128)	0
Couche de classification	(None, 6)	774
Nombre totales de paramètres		**348486**

Nous avons évalué les performances de cette architecture pour différentes configurations des bases de données :

- **Première configuration :** combinaison des trois bases des données ; $TESS$, $SAVEE$ et $RAVDESS$.
- **Deuxième configuration :** combinaison des trois bases des données précédentes mais cette fois-ci en éliminant toutes les répétitions existant dans $RAVDESS$.
- **Troisième configuration :** combinaison des trois bases des données : $TESS$, $SAVEE$ et $EMODB$.
- **Quatrième configuration :** combinaison des quatre bases de données : $TESS$, $SAVEE$, $EMODB$ et $RAVDESS$.

Le tableau 3.8 illustre les performances de l'architecture proposée en fonction des configurations des bases de données :

Tab. 3.8 : Tableau récapitulatif des résultats de combinaison de différentes bases de données

	Exactitude (%)			Coût		
Configuration	Entraînement	Validation	Test	Entraînement	Validation	Test
1	91	85	90	0.43	0.56	0.44
2	90	85	87	0.27	0.41	0.37
3	97	93	91	0.07	0.23	0.13
4	89	83	84	0.26	0.44	0.49

Nous avons testé les quatre modèles précédents en effectuant des prédictions sur une base de données, Crema D, qui n'a jamais été utilisée lors de leur entraînement. Cela nous permet de mieux évaluer la capacité de généralisation des modèles et leurs performances sur de nouvelles données. Les résultats de cette évaluation sont présentés dans le tableau suivant 3.9 :

Tab. 3.9 : Architecture du modèle proposé (CNN 2D)

Modèle	1	2	3	4
Exactitude (%)	18	27	23	21

- D'après les deux tableaux précédents, le meilleur modèle en terme d'exactitude et Coût est celui entraîné sur TESS , SAVEE et EMO DB atteignant une exactitude de 93%, tandis qu'il représente de faible capacité de généralisation.
- En comparant les deux premières configuration, l'exactitude sur l'ensemble de test a diminué de **3%**, après la suppression des échantillons répétés dans la base de donnée RAVDESS, donc on a éliminer la possibilité que les données de l'ensemble de test soient les même que celle de l'entraînement, c'est ce qui explique cette diminution en terme d'exactitude. Cependant, le modèle entraîné sur la 2éme configuration, fournis les meilleure capacités de généralisation.
- La quatrième configuration montre que le fait de fournir plus de données ne garantit pas de meilleurs résultats.
- Au final, en se basant sur les résultats de ce tableau, nous avons choisi la configuration fournissant les meilleurs résultats en terme de généralisation.Nous avons donc choisi le deuxième modèle.

CNN 2D

Cette méthode consiste à analyser à l'aide d'un CNN 2D, les représentations visuelles des signaux audio, à savoir les spectrogrammes et les spectrogrammes de Mel afin d'exploiter les motifs fréquentiels et temporels présents.

Nous avons proposé une architecture simple, composée de quatre couches de convolution 2D, deux couches de Pooling suivies par une couche d'aplatissement puis deux couches entièrement connectées dédiées à la classification. Nous avons adopté également, le dropout et les couches de normalisation de lots. Cette architecture reçoit à son entrée des images de taille 224*224 en RGB normalisées (en divisant les pixels sur 255).

L'architecture de ce modèle est détaillée dans le tableau ci-dessous 3.10 :

Tab. 3.10 : Architecture du modèle proposé (CNN 2D)

Type de couche	Forme	Paramètres
Convolution 2D	(None, 111, 111, 64)	1792
Convolution 2D	(None, 109, 109, 64)	36928
Pooling	(None, 36, 36, 64)	0
Convolution 2D	(None, 34, 34, 32)	18464
Normalisation par lot	(None, 34, 34, 32)	128
Convolution 2D	(None, 32, 32, 16)	4624
Pooling	(None, 16, 16, 16)	0
Dropout	(None, 16, 16, 16)	0
Couche d'Applatissement	(None, 4096)	0
Couche entièrement connectée	(None, 128)	524416
Dropout	(None, 128)	0
Couche de classification	(None, 6)	774
Nombre totales de paramètres		**587126**

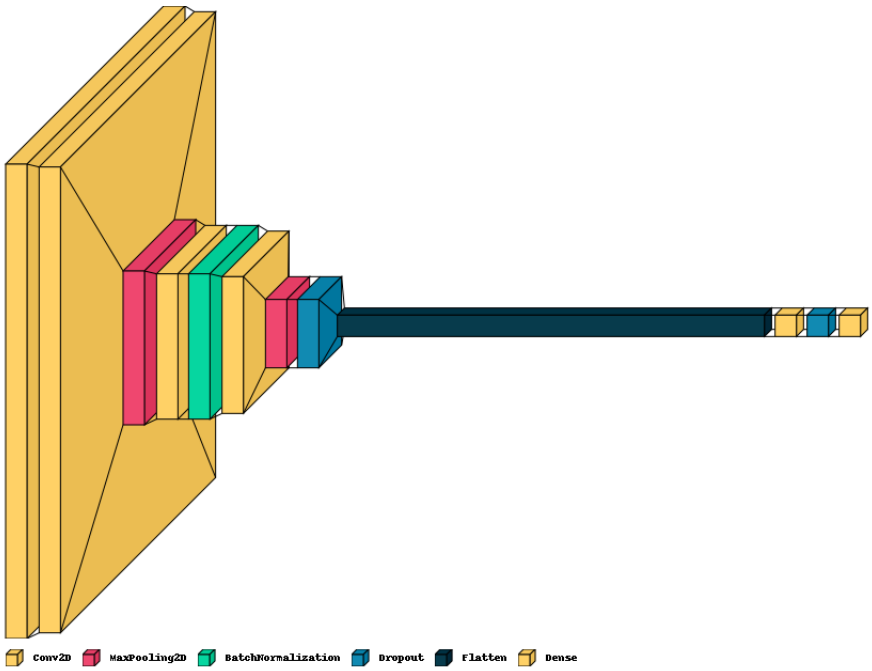


Fig. 3.9 : Architecture du modèle proposé (CNN 2D)

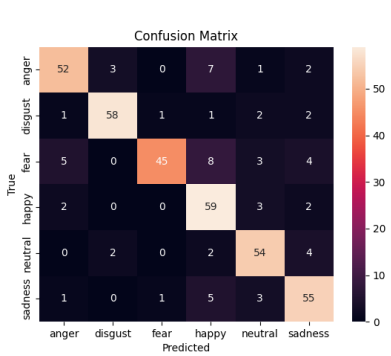
Nous avons mené une étude comparative des performances d'un CNN 2D en exploitant les spectrogrammes et les spectrogrammes de Mel, en adoptant la deuxième configuration de base de données. Les résultats sont illustrés dans le tableau suivant 3.11 :

Tab. 3.11 : Tableau comparatif

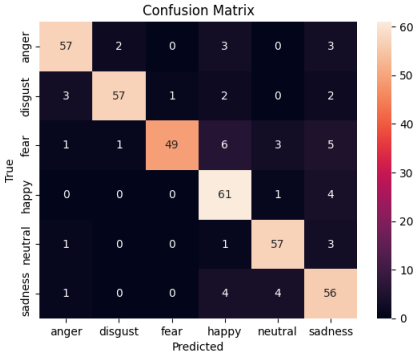
	Exactitude (%)			Coût		
Entrées	Entraînement	Validation	Test	Entraînement	Validation	Test
Spectrogramme	87	85	83	0.31	0.44	0.43
Spectrogramme de mel	95	89	87	0.13	0.31	0.36

En analysant les résultats du tableau, nous pouvons clairement constater que l'utilisation des spectrogrammes de mel a conduit à une amélioration significative des performances par rapport aux spectrogrammes standard. En termes d'exactitude, nous observons une augmentation de 4% sur l'ensemble de test, passant de 83% à 87%. De plus, le modèle entraîné sur les spectrogrammes de mel a présenté des valeurs de coût plus faibles lors de l'entraînement (0,13), de la validation (0,31) et des tests (0,36) par rapport au modèle utilisant les spectrogrammes. Ces résultats mettent en évidence l'efficacité et la robustesse supérieures des spectrogrammes de mel dans la reconnaissance des émotions par rapport aux spectrogrammes traditionnels.

L'amélioration des performances observée avec les spectrogrammes de mel s'explique par leur capacité à représenter les caractéristiques fréquentielles du signal vocal de manière plus adaptée à la perception auditive humaine. En effet, comme il a été mentionné dans le chapitre précédent, les spectrogrammes de mel utilisent une échelle de fréquences non linéaire (échelle de Mel) qui tient compte de la sensibilité différente de l'oreille humaine aux différentes fréquences. Cette représentation permet de mettre l'accent sur les bandes de fréquences pertinentes pour l'analyse de la parole, ce qui peut aider le modèle à extraire des caractéristiques plus discriminantes pour la reconnaissance des émotions. La figure **??** représente les matrices de confusion de modèles entraînés sur les spectrogrammes et les spectrogrammes de Mel respectivement.



(a) Spectrogramme



(b) Spectrogramme de mel

3.6.1.3 Discussion

Dans le cadre du développement d'un système performant de reconnaissance vocale des émotions, tout en tenant compte des contraintes liées à l'implémentation sur une carte électronique, nous avons réalisé une série de tests en explorant plusieurs aspects : les caractéristiques utilisées, les bases de données utilisées et les méthodes de classification. Nous avons examiné trois méthodes en particulier : SVM, CNN 1D et CNN 2D. Ci-dessous un tableau 3.12récapitulatif des meilleurs résultats obtenus pour chaque approche :

Tab. 3.12 : Tableau récapitulatif des performances des modèles de la RVE.

Méthode	Base de données utilisée	Performance sur l'ensemble de test	Performance sur l'ensemble de donnée de Emo DB
SVM	Tess , Savee , Ravdess	88%	18 %
CNN 1D	Tess , Savee , Ravdess	87%	40 %
CNN 2D	Tess , Savee , Ravdess	87%	41%

L'analyse des résultats du tableau 3.12 révèle que le modèle SVM a obtenu une performance plus élevée que les deux modèles CNN sur l'ensemble de test. Cependant, il a présenté une faible exactitude, de seulement 18% sur l'ensemble de données de Emo_DB, ce qui suggère une capacité de généralisation limitée. En revanche, les modèles basés sur CNN, à savoir CNN 1D et CNN 2D, ont montré de meilleures performances, surpassant le SVM de 22% en termes d'exactitude. Ces modèles sont donc capables de capturer des caractéristiques plus discriminantes dans les signaux vocaux et de mieux généraliser sur des données variées.

Après avoir éliminé le modèle SVM de la sélection, il nous reste d'en sortir avec une décision concernant le modèle de CNN final à implémenté par la suite.

Étant donné que les deux modèles ont presque les mêmes performances, on a effectué une étude comparative en terme de nombre de paramètres et le temps d'exécution, montré sur le tableau 3.13 ci dessous :

Tab. 3.13 : Tableau comparatif

Modèle	Nombre totale de paramètres	Temps d'exécution (ms)
CNN 1D	348,486	15
CNN 2D	587,126	14000

D'après les résultats du tableau 3.13, le CNN 1D possède moins de paramètres et un temps d'inférence réduit comparant à ceux du CNN 2D. En tenant compte de ces facteurs, nous avons fini par choisir le modèle CNN 1D comme étant le plus optimal. Son nombre de paramètres réduit et son temps d'inférence plus court en font un choix adapté pour des prédictions efficaces et rapides, tout en respectant les contraintes de ressources et de temps réel de la carte électronique.

Nous concluons cette discussion, en comparant nos résultats obtenus à partir des différents modèles, avec ceux des travaux présentés dans l'état de l'art.

Nous remarquons qu'avec le modèle SVM nous a dépassé les auteurs de l'article [12], qui ont utilisé une combinaison de caractéristiques (MFCC, pitch, énergie ..) extraites de la base de données Emo DB, avec une exactitude de 1.4%. D'autre part, avec notre architecture proposée de CNN 2D, nous sommes parvenus à la même conclusion que celle des auteur de l'article [15] concernant l'efficacité des spectrogrammes de Mel, tout en les surpassant avec une exactitude de plus de 30% que ça soit en comparant avec leurs résultats obtenue à partir du modèle qu'ils ont proposé ou bien en utilisant le modèle pré-entrainé Resnet18.

Enfin, notre modèle CNN 1D a montré de meilleurs résultats (87%) que ceux obtenus par les auteurs de [47] qui ont utilisé la base de données RAVDESS, et cela en les surpassant avec une exactitude de 5% d'exactitude.

3.6.2 Résultats de la reconnaissance faciale des émotions

3.6.2.1 Protocole 1: Classification en utilisant le SVM

Le premier protocole expérimental pour la classification des émotions faciales repose sur une approche traditionnelle qui consiste à extraire manuellement les caractéristiques puis mettre en oeuvre une méthode de classification parmi les algorithmes d'apprentissage automatique. Nous avons choisi la technique de l'Histogramme de gradient orienté (HoG) pour l'étape de l'extraction des caractéristiques et l'algorithme SVM pour l'étape de la classification des émotions.

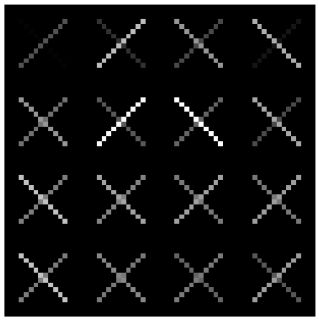
Le calcul des vecteurs caractéristiques est effectué grâce au module *feature* de la bibliothèque *scikit-image*. Le choix des paramètres tels que le nombre d'orientations, la taille des cellules et le nombre de cellules par bloc est crucial pour garantir une extraction d'information efficace. Nous avons effectué des tests en modifiant ces paramètres et en visualisant les images résultantes. Afin d'étudier l'effet du choix de ces paramètres, nous avons sélectionné trois combinaisons spécifiques, qui sont répertoriées dans le tableau.3.14.

Tab. 3.14 : Tableau récapitulatif des paramètres de HoG sélectionnés.

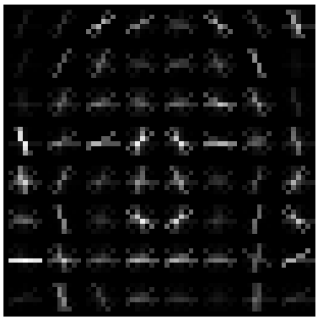
Combinaison	Nombre de bins	Taille de la cellule	Nombre de cellules dans un bloc	Taille du vecteur caractéristique
1	2	(16,16)	3	72
2	9	(8,8)	1	576
3	9	(3,3)	2	14400

La figure 3.11 représente les résultats de l'extraction des caractéristiques HoG pour les trois configurations de paramètres sélectionnées.

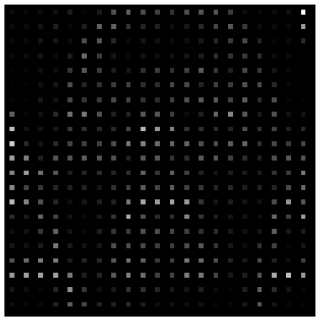
(a) Image originale.



(b) première configuration.



(c) deuxième configuration.



(d) troisième configuration.

FIG. 3.11 : Visualisation des caractéristiques HoG.

Les données sont divisées en deux sous-ensembles : l'ensemble d'entraînement regroupe 80% des données globales et les 20% restantes sont consacrées pour la phase de test comme décrit en 3.15. Les vecteurs caractéristiques des images d'entrées sont normalisées en utilisant la normalisation Z-score avant d'être introduits au classifieur SVM.

TAB. 3.15 : Tableau de la distribution des données d'entraînement et de test

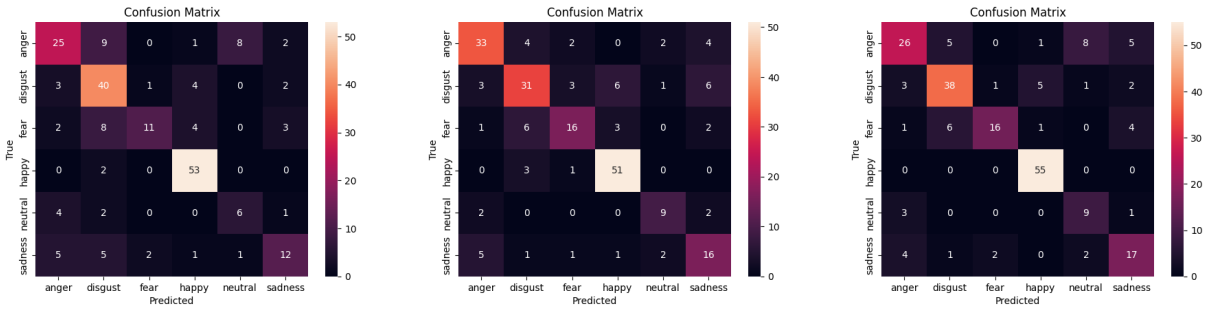
Ensemble	Nombre d'échantillons	Pourcentage (%)
Entrainement	867	80
Test	217	20
Total	1084	100

Le tableau 3.16 présente les résultats de la classification des caractéristiques obtenues à partir de la première configuration des paramètres HoG en utilisant trois noyaux : Polynôme, linéaire et Rbf. Dans cette configuration, le noyau Rbf obtient les meilleurs résultats, avec une exactitude de **74,19%**.

TAB. 3.16 : Tableau des résultats de la classification avec trois noyau du SVM de la première configuration des paramètres HoG.

Noyau	Exactitude d'entraînement (%)	Exactitude de test (%)	Précision d'entraînement (%)	Précision de test (%)	Rappel d'entraînement (%)	Rappel de test (%)
polynôme de degré 2	76,70	67,74	77,66	68,30	76,70	67,74
linéaire	82,81	71,88	82,99	71,83	82,81	71,88
Rbf	84,89	74,19	84,97	75,15	84,89	74,19

Les figures 3.12a, 3.12b et 3.12c présentent les matrices de confusion des noyaux polynomiale, linéaire et rbf respectivement. En analysant ces matrices, on observe que le noyau polynomial offre les meilleures performances pour prédire les données de la classe dégoût. D'autre part, le noyau linéaire se distingue par sa capacité à classifier efficacement les images de l'émotion colère. Le noyau Rbf, quant à lui, présente des performances moyennement supérieures aux deux autres noyaux dans toutes les classes.



(a) Noyau polynôme. (b) Noyau linéaire. (c) Noyau rbf.

FIG. 3.12 : Matrices de confusion des trois modèles SVM pour trois noyau différents.

Le tableau 3.17 présente les performances de la deuxième configuration des paramètres du descripteur HoG.

TAB. 3.17 : Tableau des résultats de la classification avec trois noyau de la deuxième configuration des paramètres de HoG

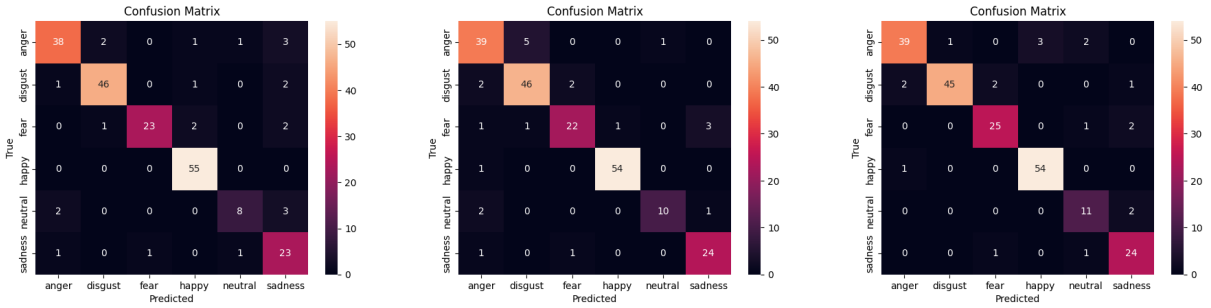
Noyau	Exactitude d'entraînement (%)	Exactitude de test (%)	Précision d'entraînement (%)	Précision de test (%)	Rappel d'entraînement (%)	Rappel de test (%)
polynôme de degré 2	99,53	88,94	99,54	89,52	99,53	88,94
linéaire	100	89,86	100	89,92	100	89,86
Rbf	99,53	91,24	99,54	91,63	99,53	91,24

On constate une amélioration significative des performances sur l'ensemble d'entraînement, avec une exactitude atteignant 100% pour le noyau linéaire, contre $89,86\%$pour l'ensemble de test. Le noyau Rbf obtient les meilleurs résultats, avec seulement 8% d'écart

entre les sous-ensembles d'entraînement et de test.

L'amélioration observée est attribuable à l'augmentation du nombre de bins d'orientation. En effet, un nombre plus élevé de bins d'orientation permet de capturer davantage d'informations sur l'orientation du gradient, facilitant ainsi la distinction entre les différentes classes pour le SVM. Cette augmentation est particulièrement bénéfique pour les émotions où il est facile de se confondre, telles que la colère, le dégoût et l'état neutre.

Les figures 3.13a, 3.13b et 3.13c illustrent les matrices de confusion des noyaux : polynomial, linéaire et Rbf respectivement. Elles offrent une visualisation détaillée des prédictions des données de test pour chaque classe, permettant ainsi de constater facilement l'amélioration de l'exactitude.



(a) Noyau polynôme. (b) Noyau linéaire. (c) Noyau rbf.

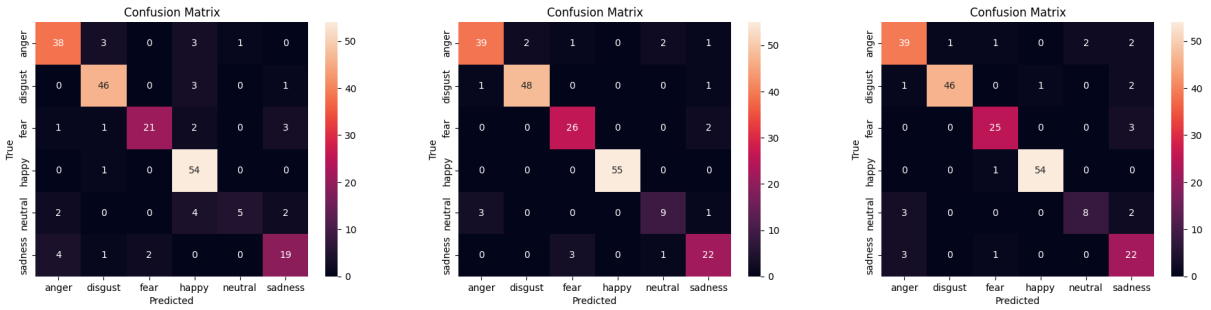
FIG. 3.13 : Matrices de confusion des trois modèles SVM pour trois noyau différents.

Les résultats de la troisième et dernière configuration sélectionnée pour les paramètres du descripteur HoG, sont présentés dans le tableau 3.18. On a fixé le nombre de bins d'orientation et on a changé les deux autres paramètres. Le modèle SVM de noyau linéaire obtient les meilleurs résultats avec 91, 70%. Nous remarquonsa travers les résultats de l'entraînement, tous les noyaux ont atteint une exactitude de 100%.

TAB. 3.18 : Tableau des résultats de la classification des caractéristiques obtenues avec la troisième configuration en utilisant trois noyaux différents du SVM.

Noyau	Exactitude d'entraînement	Exactitude de test	Précision d'entraînement	Précision de test	Rappel d'entraînement	Rappel de test
polynôme de degré 2	100 %	84,33 %	100 %	84,51 %	100 %	84,33 %
linéaire	100 %	91,70 %	100 %	91,71 %	100 %	91,70 %
Rbf	100 %	89,40 %	100%	89,83 %	100 %	89,40 %

La figure 3.14 est une illustration des matrices de confusion. La précision de classification est élevée et se rapproche de celle de la deuxième configuration.



(a) Noyau polynôme. (b) Noyau linéaire. (c) Noyau rbf.

FIG. 3.14 : Matrices de confusion des trois modèles SVM pour trois noyau différents.

3.6.2.2 Protocole 2: Classification en utilisant le CNN

La reconnaissance faciale des émotions humaines est un défi qui nécessite un modèle capable de saisir des informations complexes à partir des expressions de visage, et bien que l'algorithme SVM montre résultats satisfaisants, il est toujours intéressant d'exploiter les réseaux de neurones convolutifs qui sont populaires dans le domaine de la vision par ordinateur.

Dans ce deuxième protocole, nous avons proposé notre propre architecture de réseau de neurones convolutifs (CNN). Nous avons effectué plusieurs tests en introduisant différents types de données à l'entrée du modèle. Nous avons testé la classification à partir des images brutes, des images de contours, la combinaison des deux, ou encore les vecteurs caractéristiques obtenus à partir du descripteur HoG.

Classification des vecteurs caractéristiques du descripteur HoG à l'aide un modèle CNN 1D

Nous avons effectué la classification des caractéristiques HoG extraites précédemment à l'aide d'un modèle à base de réseau de neurones convolutif dont l'architecture est détaillé dans le tableau 3.19 afin d'effectuer une étude comparative avec le protocole 1. Nous avons fixé les configurations du tableau 3.14. Le tableau 3.20 représente les résultats obtenus.

Tab. 3.19 : Tableau des paramètres de l'architecture CNN 1D

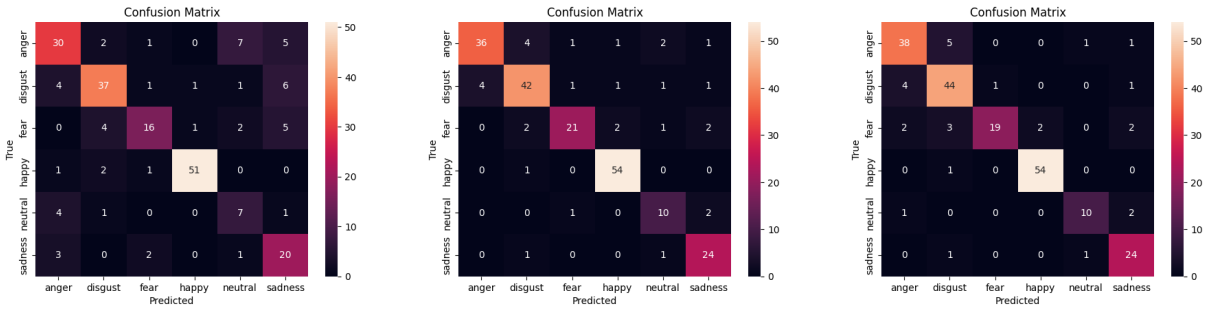
Type de couche	Taille de la sortie	Paramètres
Convolution	(None, 574, 36)	144
Convolution	(None, 285, 64)	11584
Pooling	(None, 142, 64)	0
Convolution	(None, 140, 128)	24704
Pooling	(None, 70, 128)	0
Dropout	(None, 70, 12)8	0
Couche d'applatissement	(None, 8960)	0
Couche entièrement connectée	(None, 620)	5555820
Dropout	(None, 620)	0
Couche de classification	(None, 6)	3726
Nombre totales de paramètres		**5,595,978**

Tab. 3.20 : Tableau des résultats de la classification des caractéristiques HoG avec le CNN

Configuration de paramètres HoG	Ensemble	Coût	Exactitude (%)	Précision(%)	Rappel (%)
1	Entrainemet	0,350	87,31	90,38	83,08
	Validation	0,75	77,26	80,77	72,41
	Test	0,76	74,19	80,31	71,41
2	Entrainement	0,03	99,36	99,61	99,36
	Validation	0,335	87,36	89,16	85,06
	Test	0457	86,18	87,74	85,71
3	Entrainement	0,016	99,62	99,62	99,62
	Validation	0,336	87,36	93,75	86,21
	Test	0,458	87,10	88,26	86,21

En premier lieu, nous remarquons une amélioration au niveau de la première configuration par rapport au protocole 1 de $6,72\%$. Les performances des deux dernières configurations sont assez proches l'un de l'autre et figurent toujours meilleurs que ceux de la première configuration, cela est dû au nombre d'orientation différent.

La figure 3.15 présente la matrice de confusion de la classification des vecteurs caractéristiques HoG avec le modèle CNN. Il existe une légère différence entre les matrices 3.15b et 3.15c comparant à la matrice 3.15a.



(a) Première configuration (b) Deuxième configuration. (c) Troisième configuration.

FIG. 3.15 : Matrices de confusion de la classification des caractéristiques HoG avec un modèle CNN.

En conclusion, le choix de nombre d'orientation de gradient influence de façon importante les performances du modèle.

Classification des images à l'aide d'un modèle CNN 2D

Nous avons construit notre architecture à zéro. Nous avons commencé d'abord par deux couches de convolution et puis nous en ajoutons d'autres au fur et à mesure. Nous avons testé plusieurs combinaisons possibles de paramètres : nombre de couches de convolution, nombre de couches de Pooling, nombre et taille de filtres par couche, dropout, strides et le zero padding.

La répartition des données utilisées pour l'entraînement, la validation et le test est présentée dans le tableau 3.21.

TAB. 3.21 : Répartition des classes deux deux bases de données KDEF et CK+.

Classes	Entrainement	Validation	Test	Total
Colère	164	21	20	205
Dégoût	186	24	23	233
Joie	222	28	27	227
Neutralité	56	7	7	70
Peur	116	14	15	145
Tristesse	123	15	16	154
Total	867	109	108	1084

Notre sélection est basée sur deux axes principaux : les performances atteintes et la taille du modèle car au final notre but est d'implémenter la solution sur une carte électronique Raspberry Pi.

Le tableau 3.22 récapitule les performances de quelques modèles évaluées ainsi que les détails de leurs architectures et leurs tailles.

Tab. 3.22 : Tableau des résultats des différentes architectures CNN

Architecture	Couches de convolution	Couches de Pooling	Droout	Couches entièrement connectées	Exactitude (%)		Coût		Nombre de paramètres (Millions)
					Entrainement	Test	Entrainement	Test	
1	2	2	1	1	95	85	0,13	0,49	11,13
2	3	2	3	1	94	89	0,15	0,43	4,32
3	4	3	3	2	87	81	0,35	0,53	1,96
4	5	4	6	2	82,58	76,56	0,51	0,69	1,39
5	5	4	5	2	92,56	85,24	0,23	0,49	1,56

En dépit de leurs exactitudes élevées, la première et deuxième architectures sont éliminées en raison au nombre élevé de leurs paramètres. Le quatrième modèle est le le moins performant avec une exactitude $76,56\%$. Au final, nous avons choisi **la dernière architecture** qui atteint une exactitude de **85,24%** pour les données de tests et un faible coût de **0,49**.

La figure 3.16 offre une vue globale de l'architecture sélectionnée et le tableau 3.23 illustre le nombre de couches de convolution, de Pooling, de Dropout et les couches entièrement connectées implémentées dans l'architecture sélectionnée, en spécifiant pour chaque couche le nombre et la taille des filtres.

Tab. 3.23 : Architecture détaillée du modèle CNN

Type de couche	Forme de la sortie	Nombre de Paramètres
Convolution	(None, 62, 62, 36)	360
Convolution	(None, 60, 60, 64)	20800
Pooling	(None, 30, 30, 64)	0
Convolution	(None, 28, 28, 128)	73856
Pooling	(None, 14, 14, 128)	0
Convolution	(None, 12, 12, 228)	262884
Pooling	(None, 6, 6, 228)	0
Dropout	(None, 6, 6, 228)	0
Convolution	(None, 4, 4, 256)	525568
Pooling	(None, 2, 2, 256)	0
Dropout	(None, 2, 2, 256)	0
Couche d'Applatissement	(None, 1024)	0
Couche entièrement connectée	None, 512	524800
Dropout	(None, 512)	0
Couche entièrement connectée	(None, 256)	131328
Dropout	(None, 256	0)
Couche de classification	(None, 6)	1542
Nombre totales de paramètres		**1,541,138**

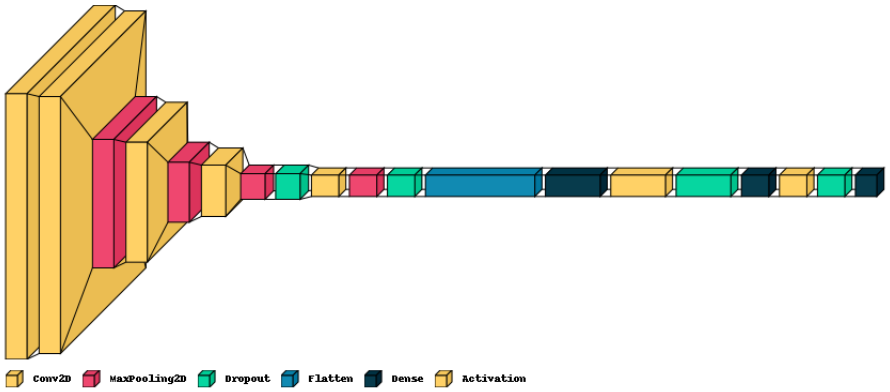


FIG. 3.16 : Vue globale de l'architecture CNN.

Durant le processus d'amélioration des performances du modèle choisi, nous avons procédé par **une augmentation des données** dans le but d'enrichir l'apprentissage du modèle en l'exposant à d'autres types d'images similaire à ceux du cas réel.Nous avons procédé par deux types d'augmentation des données :

1. Création des échantillons de façon artificielle à partir des données de l'ensemble d'entraînement disponible en utilisant des techniques géométriques notamment :
   - L'étirement des données (Data stretching) : elle consiste à étirer l'image le long d'un axe selon un certain angle appelé angle de cisaillement.
   - L'égalisation d'histogramme : elle consiste à ajuster le contraste en égalisant les valeurs de l'intensité des pixels dans une image donnée.
   - Le retournement de l'image
2. Ajout des échantillons d'une troisième base de données *RAVDESS*.

Un tableau comparatif 3.24 démontre l'effet progressif de l'augmentation des données sur les performances du modèle entraînée.

TAB. 3.24 : Tableau des résultats de l'augmentation des données.

Bases de données	Ensemble	Cout	Exactitude	Précision	Rappel
Avant augmentation des données					
KDEF, CK+	Entraînement	0.23	92,56%	92,38%	91,27%
	Validation	0.36	90,44%	90,33 %	90,44%
	Test	0.49	85.24%	85,03 %	85%
Après augmentation des données					
KDEF, CK+	Entrainement	0.0146	99,63%	99,63%	99,63%
	Validation	0.1011	97,62%	97,62%	97,62%
	Test	0.2321	**91,86 %**	91,86%	91,06%
KDEF, CK+, RAVDESS	Entrainement	0.1891	92,93%	94,29%	94,29%
	Validation	0.3887	89;55%	91,39%	86,82%
	Test	0.3011	88,74%	89,00%	87,03%

D'après le tableau 3.24, l'augmentation des données d'entraînement en utilisant des techniques géométriques a assuré une exactitude de 91%. L'ajout des échantillons d'une nouvelle base de données, *RAVDESS*, a fournit résultat de **88,74%** sur l'ensemble de test. Cette diminution d'exactitude est justifiée par le fait d'introduire des échantillons d'une distribution différente. Cependant, le modèle reste robuste et plus générale car l'ensemble d'apprentissage est plus varié.

Pour conclure, les résultats suggèrent que les techniques d'augmentation des données, ainsi que l'ajout d'un ensemble de données diversifié, peuvent être des stratégies efficaces pour améliorer les performances d'un modèle et le rendre plus robuste et capable de gérer différentes conditions et expressions.

Classification des images brutes avec leurs images de détection de contours correspondantes

Afin d'étudier l'effet de la détection des contours sur la reconnaissance faciale des émotions, nous avons mis en oeuvre trois scénarios avec trois types d'entrées : les images brutes, les images de contours puis les deux à la fois.

Pour cela, nous avons utilisé la quatrième architecture avec la distribution des données du tableau 3.21.
La classification des deux types d'images est réalisé grace à un modèle CNN à deux entrées composé de deux sous-réseaux : le premier reçoit en entrée l'image brute et le deuxième reçoit son image de contours détectés correspondante. Les détails du modèle développé sont décrit dans la figure 3.17.

Nous avons appliqué le filtre Canny et au final nous avons obtenu les résultats démontré dans le tableau 3.25. L'exactitude obtenue avec uniquement les images de contours est de 68% contre une exactitude de $76,56\%$ pour les images originales. La combinaison de l'image brute avec son image de contours conduit à une augmentation de 5% et donc une exactitude de $81,25\%$.

Tab. 3.25 : Tableau des résultats des images brutes, images de contours détectés et leurs combinaison.

	Coût			Exactitude		
Données d'entrée	Entraînement	Validation	Test	Entraînement	Validation	Test
Images brutes	0.5109	0.7632	0.6924	82,58%	73,02%	76,56%
Images détection de contour	0.5849	0.9179	0.7136	78,08%	69,84%	68,75%
Les deux jointes	0.2046	0.4220	0.5040	92,41%	86,15%	81,25%

Pour conclure, il est important de noter que la détection des contours ne garantit pas des performances optimales en soi. Bien que l'utilisation des deux types d'images puisse améliorer les résultats de prédiction, il convient de noter que cette approche conduit à un doublement des paramètres du modèle alors que nous avons obtenu de meilleurs résultats

en utilisant une architecture plus simple.

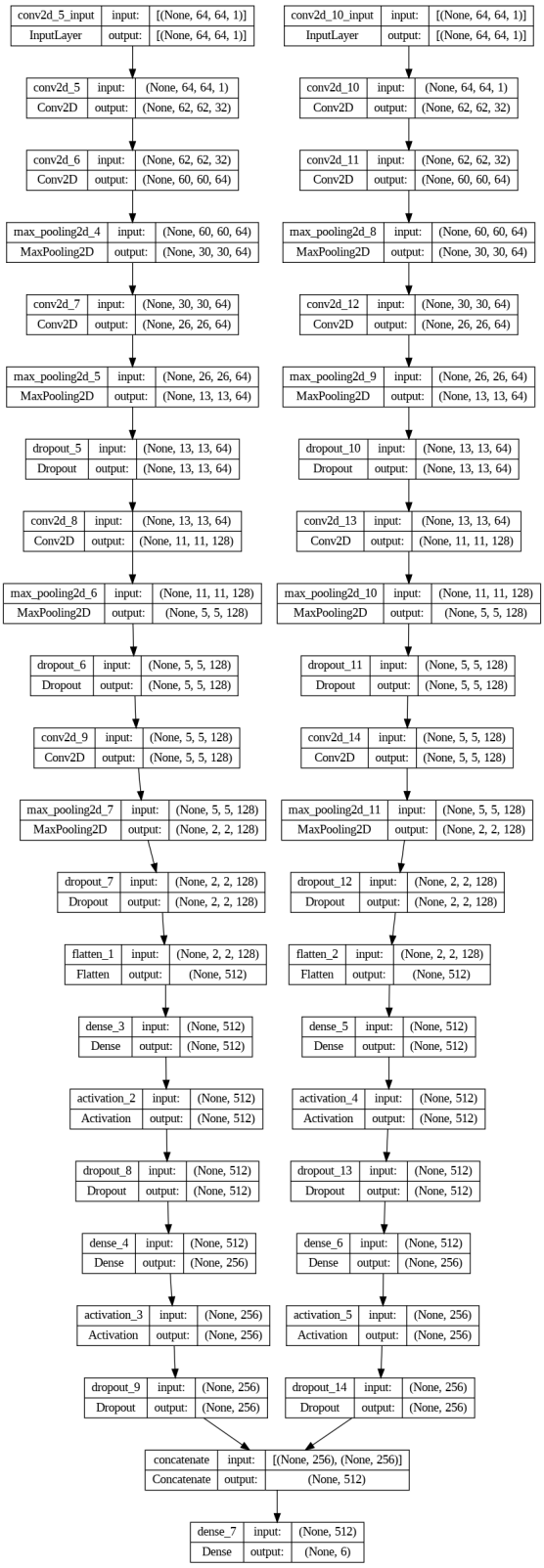


FIG. 3.17 : Modèle de base de la matrice de confusion

3.6.2.3 Discussion

Afin de vérifier nos conclusions concernant les modèles de la RFE, nous avons effectué une évaluation finale sur les images de RAVDESS dans le but de tester leur capacité de généralisation. Nous avons sélectionné les modèles ayant montré les meilleurs performances. Les résultats présentés dans le tableau 3.26 confirment notre propos.

Tab. 3.26 : Résultats de l'évaluation des modèles.

Bases de données	Type de données d'entrée	Algorithme	Performance sur l'ensemble de test	Performances sur l'ensemble RAVDESS
KDEF, CK+	HoG	CNN	87,10%	32%
KDEF, CK+	HoG	SVM	91,70%	39%
KDEF, CK+	Image brute	CNN	91,86%	46%

Le tableau 3.27 synthétise les meilleurs résultats obtenus pour différentes approches. Nous constatons que la classification des images brutes avec les réseaux CNN garantit les meilleures performances.

En effet, le CNN démontre sa capacité à capturer des informations pertinentes pour la classification lors de la phase d'entraînement. En comparaison, lors de l'extraction manuelle des caractéristiques, une perte de détails discriminants peut se produire comme dans notre cas avec HoG ou Canny. De plus, la classification est sensibles aux paramètres de ces algorithmes, ce qui limite leur généralisation à des images en dehors de la base de données d'entraînement.

Tab. 3.27 : Tableau récapitulatif des meilleurs résultats obtenus pour chaque cas traité.

Base de donnée	Augmentation des données	Type de donnée d'entrée	Algorithme	Performances
KDEF, CK+	X	HoG	SVM	91,70%
KDEF, CK+	X	HoG	CNN	87,10%
KDEF, CK+	X	Image brute et l'image des contours correspondante	CNN	81,25%
KDEF, CK+	X	Image brute	CNN	85,24%
KDEF, CK+	✓	Image brute	CNN	**91,86%**
KDEF, CK+, RAVDESS	✓	Image brute	CNN	**88,74%**

Pour conclure, nous comparons nos travaux vec quelques articles publiés.
Nous concstatons que nous sommes arrivé à des résultats similaure à ceux de l'article [48], qui ont atteint un taux d'exactitude maximum de $85,7\%$ pour la base de données JAFFE. De plus, nos résultats étaient proches de ceux de l'article [49] pour l'approche Hog+SVM appliquée à la base de données KDEF. Alors qu'ils ont utilisé uniquement KDEF, nous avons utilisé une combinaison de KDEF+CK. Par ailleurs, l'article [50] a obtenu des performances allant jusqu'à 96% pour la base de données CK+ en utilisant uniquement la méthode HoG et SVM. Finalement, nos résultats étaient presque identiques à ceux de l'article [18], qui a atteint un taux d'exactitude de 91,79% en utilisant leurs propre architecture CNN.
Nous tenons a rappeler que le modèle final de la reconnaissance faciale des émotions est celui entrainé sur trois bases de données, ayant atteint une exactitude de 88,74%.

3.7 Fusion des deux modèles de la reconnaissance des émotions

La fusion des signaux de la parole et des expressions faciales constitue une avancée majeure dans le domaine de la reconnaissance automatique des émotions. En combinant ces deux sources d'information, nous pouvons obtenir une compréhension plus précise et holistique des états émotionnels d'un individu.

Cette section est dédié à la fusion des deux modèles finaux de la RVE et la RFE choisis précédemment. Nous allons d'abord évoquer un aperçu général sur les méthodes de la fusion puis nous exposons la méthode suivie ainsi que les résultats obtenus et enfin nous terminerons avec une comparaison entre les deux modèles et leur fusion.

3.7.1 Aperçu général sur la fusion des modèles

La fusion de modèles fait référence au processus d'intégration de plusieurs modèles ou prédictions individuels en un seul modèle ou prédiction unifié. Elle vise à exploiter les forces des différents modèles afin d'améliorer les performances globales et d'obtenir de meilleurs résultats.
Sanderson and Paliwal [51] ont classé la fusion en deux grandes catégories : la fusion préclassification et la fusion post-classification comme indiqué dans la figure3.18.

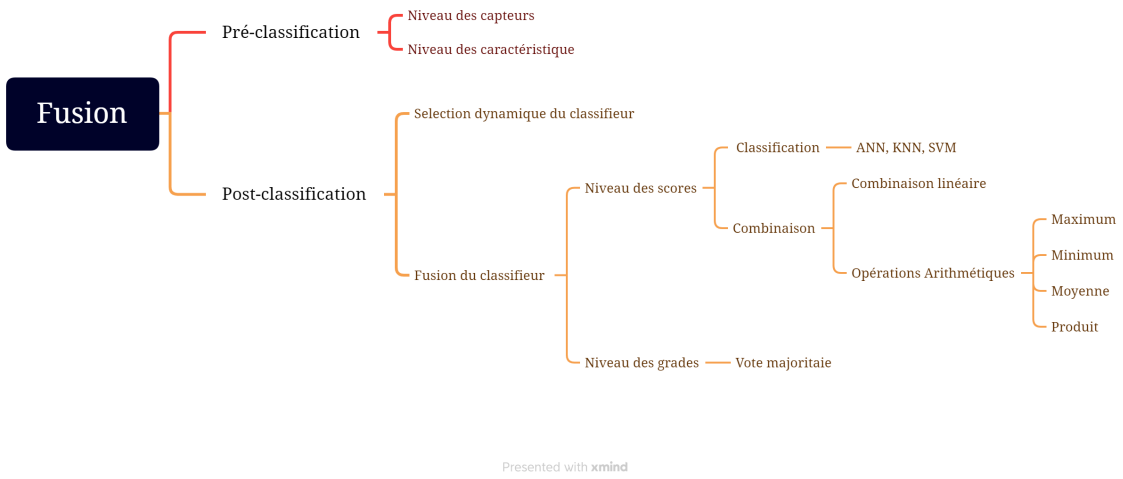


Fig. 3.18 : Différentes méthodes de la fusion.

3.7.1.1 Fusion préclassification

Ce type de fusion consiste à combiner les informations avant la classification, elle peut se réaliser à deux niveaux

- **Capteurs** : elle est possible dans le cas où les données sont de même nature mais

proviennent de différents capteurs par exemple les images de visage obtenues à partir de plusieurs caméras peuvent être combinées pour former une seule image de visage.

- **Caractéristiques** : en concaténant différentes caractéristiques extraites en un seul vecteur. Cependant, il faut veiller aux caractéristiques qui sont fortement corrélées et le problème de la malédiction de la dimension.

3.7.1.2 Fusion post-classification

Elle est réalisée après avoir effectué la classification. Donc il s'agit de fusionner les prédictions (les résultats) des classifieurs. Elle peut être accomplie de deux façons distinctes :

- **Système de sélection dynamique** : en mettant en place un des classifieurs où le résultat final est celui du classifieur le plus susceptible de prendre une décision correcte.
- **Fusion des classifieurs** : en prenant en compte les décisions de tous les classifieurs. Cette fusion s'établie au niveau des grades (rank) en effectuant par exemple un vote majoritaire ou au niveau des scores des prédictions.

  Les scores peuvent être combinés de deux manières : soit en les fusionnant avec un autre classifieur comme le SVM, soit en utilisant une combinaison linéaire simple des vecteurs de prédictions. Il existe également une approche simple et efficace qui consiste à appliquer des opérateurs arithmétiques tels que le produit, le maximum, le minimum et la moyenne aux scores.

  Dans le cadre de notre projet, nous avons choisi d'utiliser la méthode de la fusion des scores de prédictions en appliquant les différentes opérations arithmétiques. C'est une technique simple, peu coûteuse et ne nécessite aucun entraînement préalable, ce qui en fait une solution pratique et facile à implémenter. En d'autres termes, nous allons effectuer des opérations arithmétiques sur les vecteurs de probabilités d'appartenance à chaque classe (émotion) résultant de la classification.

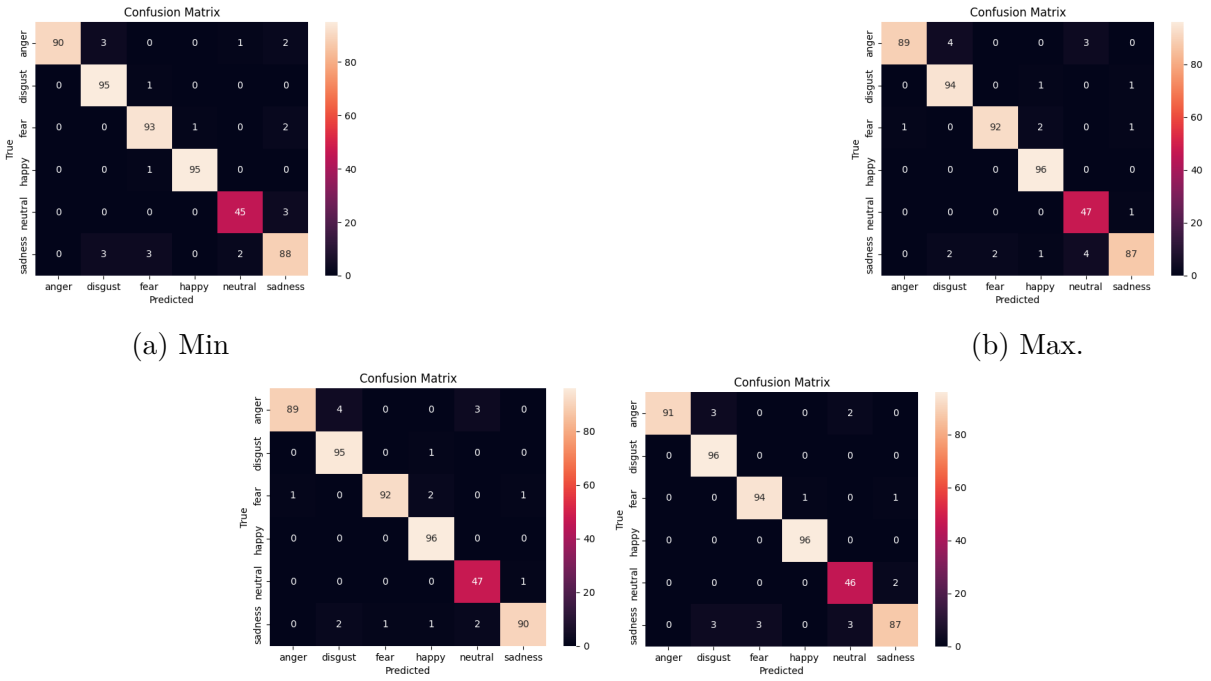
3.7.2 Résultats de la fusion des deux modèles de la RVE et la RFE

Comme indiqué précédemment, nous avons évalué la fusion des deux modèles de la RFE et la RVE avec les opérations arithmétiques suivantes : le maximum, le minimum, le produit et enfin la moyenne, sur les vecteurs de scores contenant les probabilités pour chaque émotion. Les résultats sont présentés dans le tableau 3.29 ci- dessous :

Tab. 3.28 : Comparaison des différentes opérations pour la fusion.

	Type de fusion			
	Min	Max	Produit	Moyenne
Exactitude (%)	96	96	97	96

La figure 3.19 représente les matrices de confusion correspondant à chacune des techniques précédentes.



(a) Min (b) Max.

(c) Moyenne. (d) Produit.

FIG. 3.19 : Matrices de confusion des quatre modèles de fusion.

En analysant le tableau 3.29 et les matrices de confusion 3.19 ci-dessus, nous remarquons que les quatre opérations de fusion fournissent pratiquement les même résultats. Nous avons donc décidé d'implémenter la fusion sur la Raspberry Pi en effectuant la moyennes des vecteurs de probabilités en raison de sa simplicité et le fait qu'elle prend en considération durant le calcul les scores des deux modèles.

3.7.3 Comparaison des résultats de la RVE et RFE avec leur fusion

Ci-dessous un tableau comparatif des performances des trois systèmes de reconnaissance des émotions basées sur les signaux vocaux, les expressions faciales et enfin la fusion des deux modalités :

TAB. 3.29 : Comparaison des résultats de la RFE, la RVE et leur fusion.

	Modalité		
	RVE	RFE	Fusion
Exactitude (%)	87	88.24	**96**

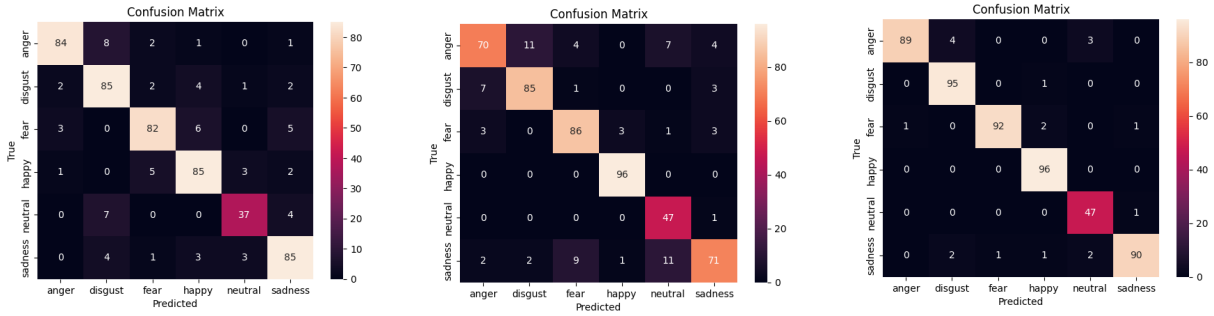
Nous constatons que la fusion a démontré une amélioration significative des performances dans notre système avec une augmentation d'exactitude de 9%.

3.7.3.1 Discussion

Le modèle de reconnaissance des émotions à partir de la voix (RVE) se concentre sur l'analyse des caractéristiques vocales et des modulations du signal. Il est capable de capturer les variations de ton, de rythme et d'autres paramètres acoustiques pour identifier les émotions exprimées dans la voix. Cependant, il peut rencontrer des difficultés lorsque les émotions sont faiblement exprimées, c'est là où les informations visuelles sont nécessaires pour une meilleure compréhension.

D'autre part, le modèle de reconnaissance faciale des émotions (RFE) s'appuie sur l'analyse des expressions faciales et des traits des émotions visibles. Il peut détecter les mouvements musculaires, les expressions du visage et les micro-expressions qui indiquent différentes émotions. Cependant, il peut également être sensible à des conditions d'éclairage variables, à des angles de vue différents ou à la présence d'obstacles visuels.

Par conséquent , grâce à la fusion des deux modalités, notamment, les informations vocales et visuelles, nous avons pu exploiter les avantages spécifiques de chaque modèle et compenser leurs éventuelles limitations. Comme le montrent les matrices de fusion illustrées sur la figure 3.20 :



(a) RVE (b) RFE. (c) Fusion.

FIG. 3.20 : Matrices de confusion de la RVE, RFE et de leur fusions.

En conclusion, nous soulignons que nous avons réussi à obtenir des résultats meilleurs que ceux des auteurs de l'article [52] qui ont arrivé à une exactitude de 80.08% et ceux des auteurs de l'article [26] qui ont arrivé à $69,42\%$.

3.8 Conclusion

En conclusion, ce chapitre méthodologie et résultats a permis de présenter les différentes composantes essentielles de notre étude. Nous avons commencé par décrire les bases de données utilisées. Ensuite, nous avons détaillé le protocole expérimental mis en place, en expliquant les différentes étapes et procédures suivies pour atteindre nos objectifs.

L'objectif principal de notre travail était d'explorer la reconnaissance des émotions à la fois à partir de l'expression faciale et de la voix. Nous avons présenté les résultats obtenus pour chacune de ces modalités et avons souligné la justification du choix de nos deux modèles finaux. Nous avons ensuite procédé à une fusion des scores des modèles pour parvenir à une décision finale, démontrant ainsi que cette dernière permet d'obtenir des informations plus riches sur l'état émotionnel d'une personne.En outre, nous avons constaté que la fusion des images faciales avec les données vocales conduit à de meilleures performances que l'utilisation de ces deux modalités de manière isolée. Cela met en évidence l'importance de prendre en compte plusieurs sources d'informations lors de la reconnaissance des émotions.

Dans le prochain chapitre, nous aborderons l'implémentation du modèle final sur la carte Raspberry Pi, ce qui nous permettra d'effectuer des prédictions en temps réel. Cette étape est essentielle pour une éventuelle application pratique de notre système, en offrant la possibilité de détecter les émotions en temps réel et d'adapter les interactions en conséquence.

Chapitre 4

Implémentation de la solution

4.1 Introduction

Dans ce chapitre, Nous commencerons par présenter les outils matériels et logiciels utilisés lors de cette implémentation, ainsi que le processus d'acquisition des données. Ensuite, nous détaillerons le prototype réalisé et les différentes étapes de tests et de démonstrations en temps réel effectuées. Enfin, nous mènerons une comparaison des performances de notre système avant et après l'implémentation.

4.2 Matériels et logiciels

4.2.1 Raspberry Pi

La Raspberry Pi se présente sous la forme d'une petite carte électronique qui n'est ni un microcontrôleur ni un microprocesseur, mais plutôt un micro-ordinateur, qui lorsqu'il est connecté à une souris, un clavier est un écran, fonctionne comme n'importe quel ordinateur présentant des performances relativement plus lentes.

Bien que la Raspberry Pi puisse être moins puissante qu'un ordinateur de bureau ou un ordinateur portable, elle offre néanmoins des performances de traitement, de mémoire et d'interfaces suffisantes pour des applications spécifiques telles que la reconnaissance automatique des émotions. Son faible encombrement et sa consommation d'énergie réduite en font une solution idéale pour des projets embarqués ou autonomes.

Dans ce projet, nous avons opté pour une carte Raspberry Pi 4 illustrée dans la figure 4.1, étant la dernière évolution des célèbres cartes Raspberry Pi, lancé en juin 2019. Présentant des performances révolutionnaires en termes de vitesse de processeur, de performances multimédias, de capacité mémoire et de connectivité par rapport à sa génération précédente, tout en maintenant une consommation d'énergie similaire.
Les spécifications techniques clés du produit peuvent être trouvées ci-dessous :

Tab. 4.1 : Spécification technique de la Raspberry Pi 4 [53]

Processeur	quad-core Cortex-A72 ,64-bit
Mémoire	8 GB LPDDR4 RAM
Connexion sans fil	Bluetooth 5.0 Wi-Fi 802.11b/g/n/ac wireless
Ports	2 x USB 3.0 2 x USB 2.0 1 x USB-C (alimentation seulement) 1 x GPIO 40 pin 2 x micro-HDMI Port caméra CSI
Alimentation	5V DC via un connecteur USB-C (minimum 3A) 5V DC via un en-tête GPIO (minimum 3A) Power over Ethernet (PoE) (nécessite un HAT pour PoE)

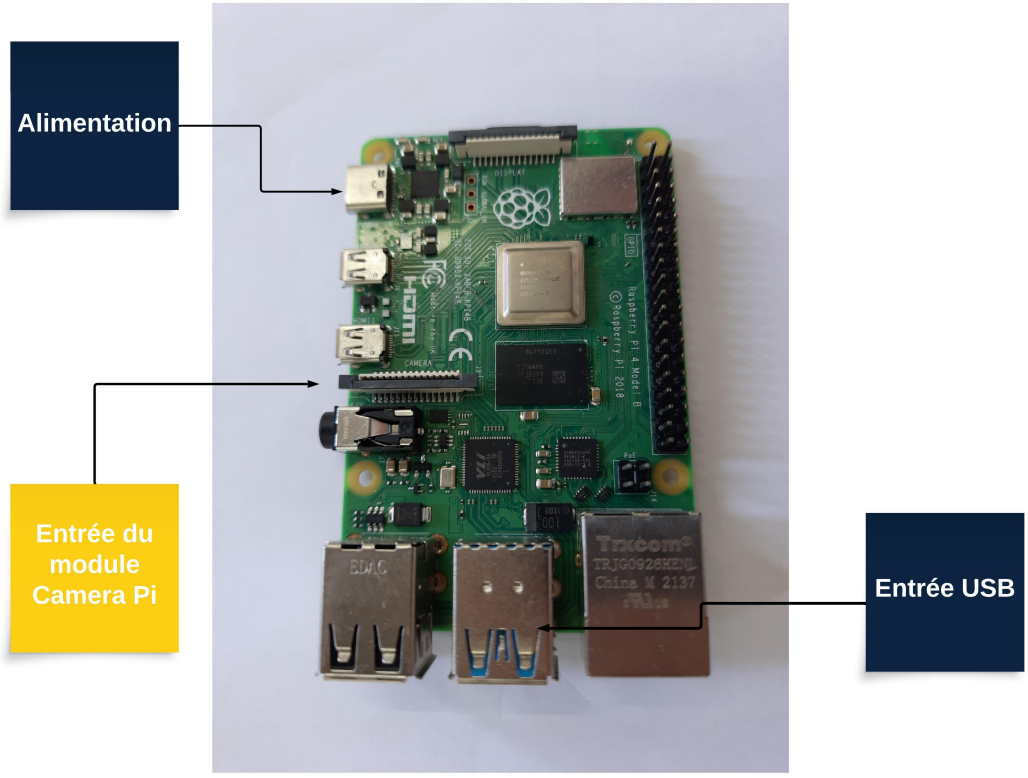
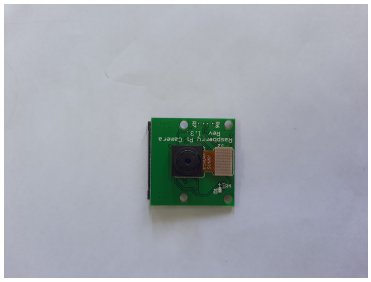


Fig. 4.1 : Raspberry Pi model B.

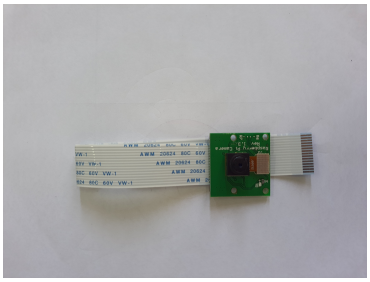
4.2.2 Module Camera Pi

Le module Camera Pi, illustré dans la figure 4.2a est un accessoire conçu pour la Raspberry Pi, permettant à cette dernière de capturer des images et des vidéos de haute qualité. Il se fixe facilement sur le port CSI (Camera Serial Interface Raspberry Pi via) de la Raspberry, via un câble-ruban, comme indiqué sur la figure 4.2b, offrant une intégration simple. Ce module est disponible en différentes versions, avec des résolutions et des fonctionnalités

variables.



(a) Module caméra.



(b) Caméra + le câble ruban.

Dans le cadre de notre projet nous avons effectué nos différents tests en utilisant le module Camera Rev 1.3 qui est caractérisé par une résolution maximale de 5 mégapixels, permettant de capturer des images détaillées. Il prend en charge l'enregistrement de vidéos haute définition jusqu'à 1080p à 30 images par seconde.

4.2.3 Microphone USB

Un microphone USB est un périphérique d'entrée audio qui se connecte à la Raspberry via un port USB. Il est conçu pour capturer les sons et les convertir en signaux électriques, qui sont ensuite transmis à l'ordinateur pour un traitement ultérieur, tel que la reconnaissance vocale des émotions. Pour notre circuit, nous avons utilisé Un microphone associé à un casque de référence Lenovo thinkplus Headphones G30.

4.2.4 Bitvise

Bitvise est un logiciel de client/serveur SSH [1] qui offre une solution sécurisée et conviviale pour accéder à des serveurs à distance. Il est largement utilisé pour se connecter à des systèmes Linux et autres dispositifs basés sur Unix depuis des machines Windows.

Dans notre projet, nous avons utilisé Bitvise SSH Client pour accéder à la ligne de commande de la Raspberry Pi depuis Windows et d'exécuter des commandes à distance, transférer des fichiers et effectuer d'autres opérations sur le système. Pour cela, il suffit d'installer Bitvise SSH Client sur le PC Windows, puis d'activer le service SSH et enfin, saisir l'adresse IP, le nom d'utilisateur et le mot de passe sur la page d'acceuil de Bitvise SSH Client comme indiqué dans la figure 4.3. La figure 4.4 montre comment on peut accéder et manipuler en même temps les fichiers sur la Raspberry Pi et le PC.

[1] Le protocole Secure Shell (SSH) est un protocole de réseau cryptographique permettant d'exploiter des services réseau en toute sécurité sur un réseau non sécurisé. Ses applications les plus notables sont la connexion à distance et l'exécution de la ligne de commande.

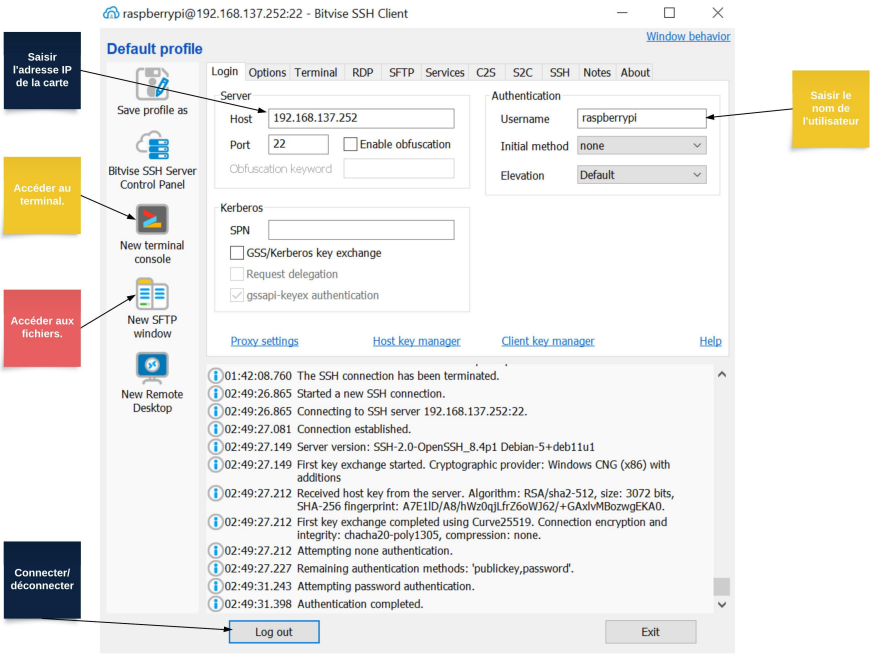


Fig. 4.3 : Page d'acceuil de Bitvise SSH Client

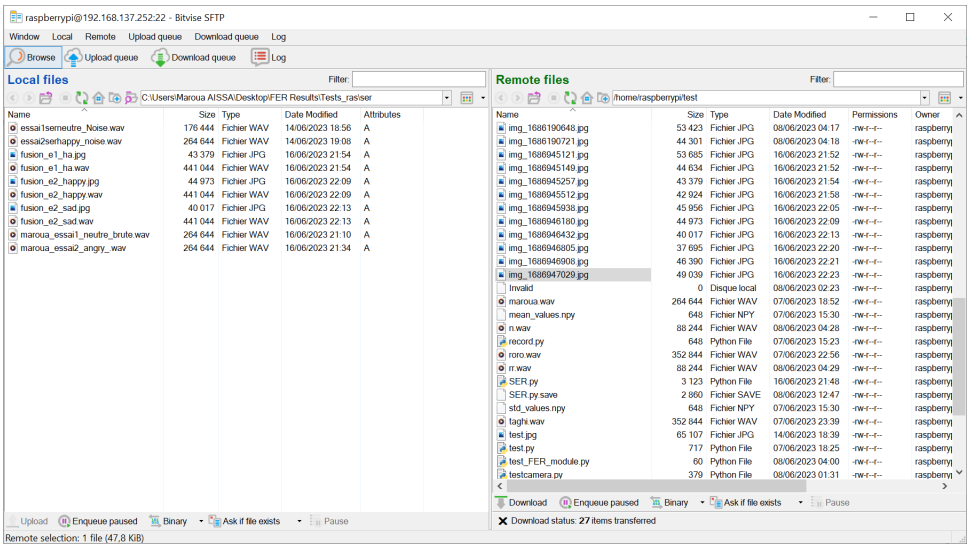


Fig. 4.4 : Accès aux fichiers.

4.2.5 TensorFlow Lite

TensorFlow Lite est une version légère et optimisée du framework TensorFlow, conçue spécifiquement pour les applications mobiles et les dispositifs à ressources limitées telles que que la Raspberry Pi. Il permet de déployer des modèles d'apprentissage automatique sur ces dispositifs embarqués, en tirant parti de leur capacité de calcul local, sans nécessiter une connexion internet constante ou une puissance de calcul importante.

Le processus d'utilisation de TensorFlow Lite comprend plusieurs étapes.

- **Entraînement** du modèle d'apprentissage automatique à l'aide de TensorFlow dans un environnement de développement classique.
- **Conversion** du modèle entraîné en un tampon plat compressé avec l'outil de conversion de TensorFlow Lite.
- **Optimisation** des ressources matérielles de la carte en procédant par une quantification en convertissant des nombres (poids du modèle entraîné) à virgule flottante, codés sur 32 bits en nombres entiers codés sur 8 bits plus efficaces.
- **Déploiement** du fichier compressé et optimisé sur la carte désirée.

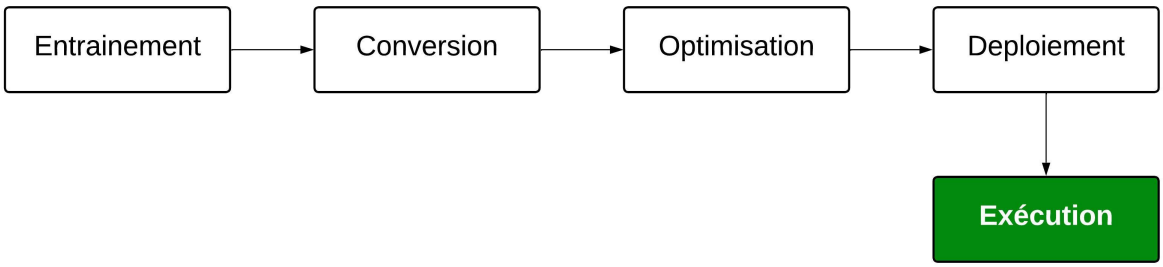


Fig. 4.5 : Étapes de l'implémetentaton du modèle TFlite.

TensorFlow Lite est utilisé dans le contexte de notre travail pour ses dénombrables avantages notamment :

- Sa légèreté.
- Sa capacité d'optimiser les performances avec une faible consommation d'énergie.
- La possibilité d'exécution des modèles d'apprentissage automatique sur des appareils embarqués, tels que la Raspberry Pi, en fournissant des résultats rapides et précis, tout en conservant une empreinte mémoire réduite.

4.3 Implémentation

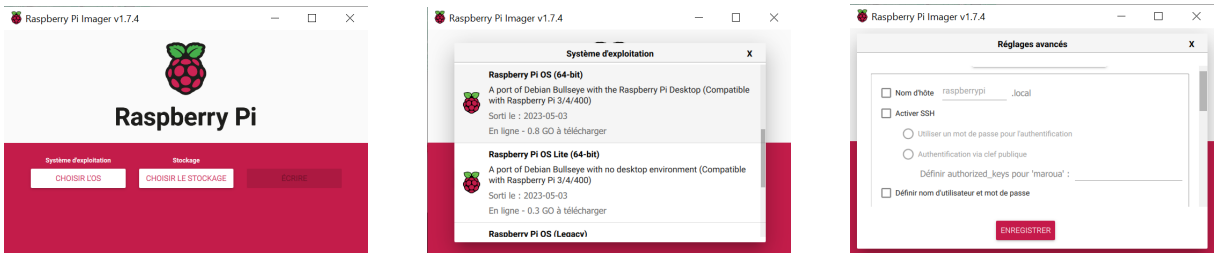
Dans cette section, nous allons présenter les outils que nous avons utilisés pour la configuration de la Raspberry Pi ainsi que les dépendances installées. Nous allons également évoquer la conversion du modèle en utilisant TFlite.

4.3.1 Configuration de la Raspberry Pi

La première étape de la préparation de la carte Raspberry Pi est l'installation du système d'exploitation(OS). Pour cela, nous avons besoin d'une carte SD de taille minimale de 8G de RAM. Nous utilisons le ***Raspberry Pi Imager*** pour écrire le fichier du OS sur la

carte SD.

Raspberry Pi Imager est un logiciel développé par la Fondation Raspberry Pi qui permet de simplifier le processus d'installation du système d'exploitation sur une carte Raspberry Pi. Grâce à cet outil convivial, il est possible de choisir parmi une variété de systèmes d'exploitation, tels que Raspbian, Ubuntu, ou d'autres distributions personnalisées, et les installer facilement sur la carte Raspberry Pi comme indiqué sur la figure 4.12c. Il permet également de la configurer en précisant le nom d'utilisateur, le mot de passe et les coordonnées du réseau WIFI au quel la carte peut se connecter comme illustré sur la figure 4.6c



(a) Page d'accueil . (b) Listes des OS proposés. (c) Paramètres à configurer.

FIG. 4.6 : Aperçu de la configuration de la carte SD avec Raspberry Pi Imager.

Raspberry Pi Imager fonctionne en téléchargeant le fichier du système d'exploitation sélectionné, puis en le gravant sur la carte SD ou microSD à l'aide d'un processus de copie bit à bit. Une fois le processus terminé, la carte est prête à être insérée dans le Raspberry Pi.

Pour notre application, nous avons installé une version légère du dernier OS sans interface du bureau Raspberry Pi OS LITE 64 bits sorti en mai 2023, compatible avec le processeur de la carte utilisée.

4.3.2 Installation des dépendances

Une fois le système d'exploitation est installé sur la carte, on procède comme suit :

- **Mise à jour du système d'exploitation** : afin d'assurer un bon fonctionnement il est impératif de mettre à jour le système d'exploitation de la Raspberry Pi, en exécutant ces deux commandes dans la fenêtre du terminal :
  - sudo apt-get update.
  - sudo apt-get upgrade.
- **Installation des librairies** : nous utiliserons la commande «pip» pour faciliter l'installation. Voici quelques-unes des bibliothèques essentielles que nous allons installer

- Librairie OpenCV : didiée la manipulation les images.
- Librairie Librosa : didiée la manipulation des signaux de la parole
- Libcamera : permet d'utiliser le module de caméra Raspberry Pi en intégrant efficacement le traitement d'image avec le processeur de la carte.
- Tensorflow Interpreter : c'est une interface d'interprétation pour l'exécution des modèles TensorFlow Lite.
- Sounddevice et soundfile : permet d'enregistrer et lire des fichiers audio à partir du microphone.

4.3.3 Conversion du modèle

Nous allons convertir les modelés de la RFE et la RVE en des modèles TFlite afin de les implémenter de la Raspberry Pi. Nous allons suivre les étapes décrites dans la section précédente. Le tableau 4.2 démontre la réduction de la taille des deux modèles après leurs conversion.

Tab. 4.2 : Comparaison de la taille des modèles Keras avec modèles Tflite.

Taille (MB)	Modèle de la RFE	Modèle de la RVE
Modèle Keras	17.72	4.055
Modèle Tflite	5.885	1.336

4.4 Tests exprimentaux sur la Raspberry Pi

Cette section est dédiée à la présentation des résultats des tests effectués sur la Raspberry Pi.

4.4.1 Élimination du bruit du microphone

Les bases de données utilisées pour entraîner notre modèle sont constituées d'enregistrements audio réalisés dans des conditions contrôlées, dans des laboratoires hautement équipés. En conséquence, ces enregistrements sont généralement de grande qualité et exempts de bruit. Cependant, lors de l'utilisation du microphone intégré à notre circuit, les enregistrements peuvent présenter un certain niveau de bruit indésirable.

Afin de remédier à cette situation et d'améliorer les prédictions du modèle en éliminant le bruit perturbateur, nous avons utilisé **Noise reduce** qui est un algorithme de réduction du bruit sur python qui diminue le bruit dans les signaux du domaine temporel comme la parole, la bioacoustique et les signaux physiologiques.

Il s'appuie sur une méthode appelée "spectral gating" qui est une forme de Noise Gate[2]. Il calcule le spectrogramme d'un signal (et éventuellement d'un signal de bruit) et estime

[2]Dispositif électronique ou un logiciel utilisé pour contrôler le volume d'un signal audio. Les portes de bruit atténuent les signaux qui se situent en dessous du seuil.

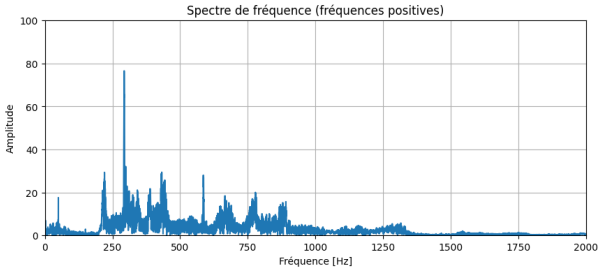
un seuil de bruit pour chaque bande de fréquence de ce signal (bruit). Ce seuil est utilisé pour calculer un masque, qui bloque le bruit en dessous du seuil variable en fonction de la fréquence.
Le tableau 4.3 montre l'effet de l'application des filtres sur les prédictions.

Tab. 4.3 : Prédictions en fonction des filtres appliquées.

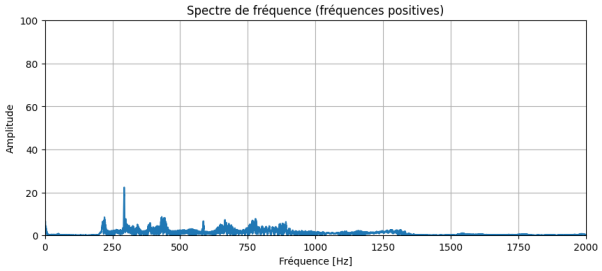
Essai	Émotion exprimée	Validation humaine	Signal bruité	Réduction du bruit	filtre de préaccentuation	Réduction du bruit puis filtre de préaccentuation
1	neutralité	neutralité/tristesse	tristesse	neutralité	tristesse	colère
2	colère	colère	neutralité	dégoût	colère	colère
3	neutralité	neutralité	tristesse	neutralité	dégoût	colère

L'utilisation des deux filtres engendre une augmentation de l'énergie du signal, ce qui induit une confusion pour le modèle et conduit à des prédictions erronées d'émotion, en particulier celle de la colère. Étant donné que le signal utile est faible comparant au bruit, il a été difficile de distinguer correctement les émotions de tristesse et de neutralité.

Pour une meilleure visualisation de l'effet de la réduction du bruit, nous avons effectué l'analyse de fourier sur le signal bruité et le signal filtré. Le résultat est démontré dans la figure 4.7.



(a) Signal bruité .



(b) Signal filtré.

Fig. 4.7 : Le signal vocal avant et après la réduction du bruit.

Comme indiqué dans la figure 4.8, les amplitudes de fréquences du bruit sont atténuées.

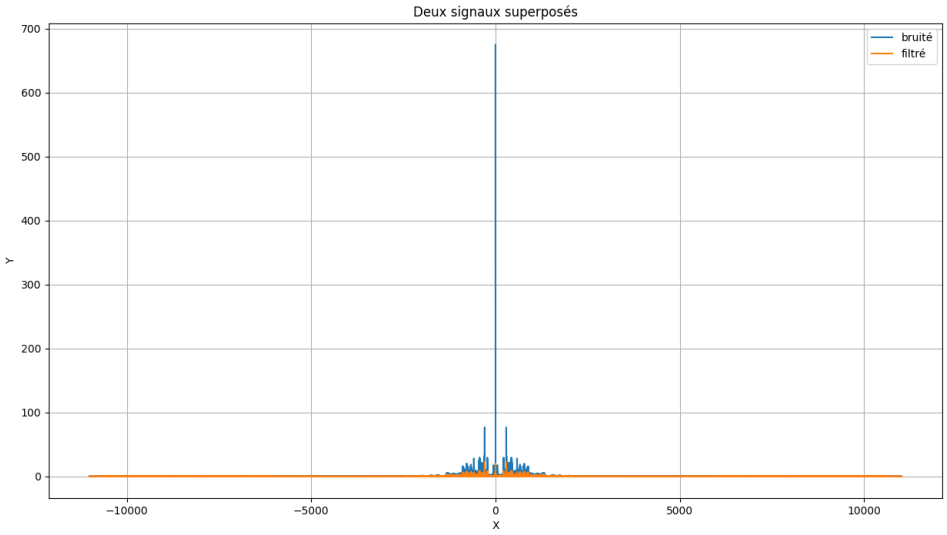


FIG. 4.8 : Les deux signaux superposés

4.4.2 Résultats des prédictions en temps réel

Nous avons effectué des tests en temps réel en utilisant le circuit représenté dans la figure 4.9. Le processus de test comprend plusieurs étapes. Tout d'abord, une image est capturée à l'aide du module caméra pour détecter l'émotion faciale. Ensuite, un fichier audio est enregistré à partir du microphone. Une fois les données acquises, elles sont soumises aux différentes étapes expliquées dans le deuxième chapitre de notre travail. Enfin, les vecteurs de prédictions pour les deux modalités, ainsi que leur fusion, sont affichés sur le terminal de la Raspberry comme le démontre la figure 4.10.

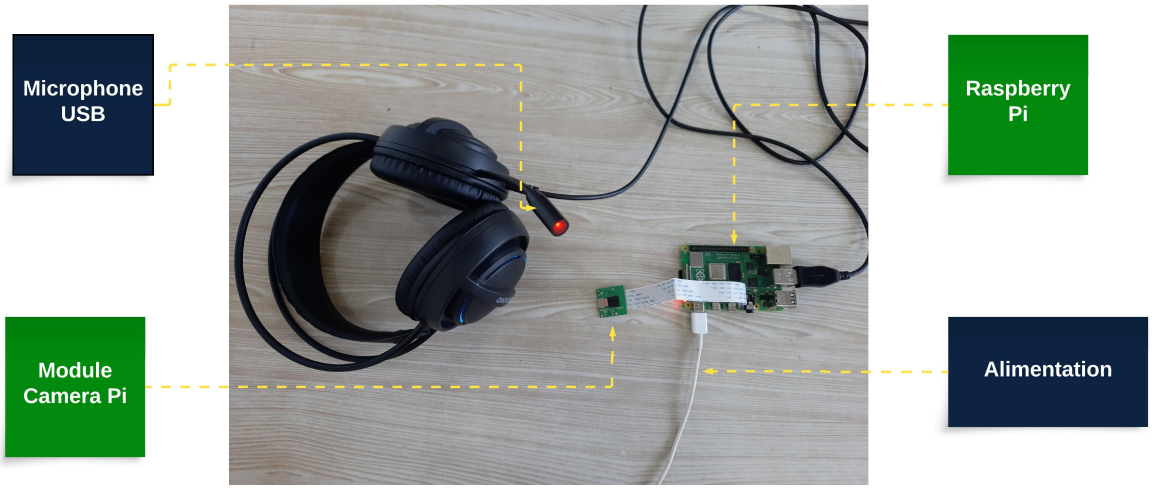


FIG. 4.9 : Montage du circuit.

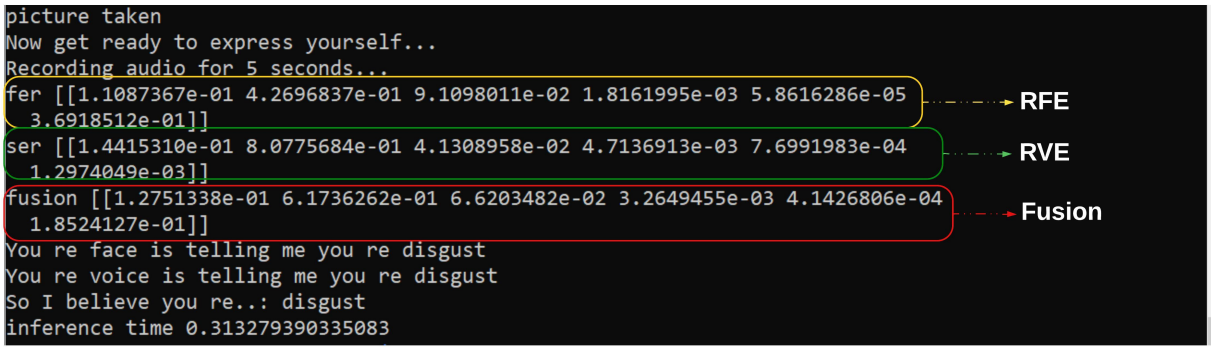


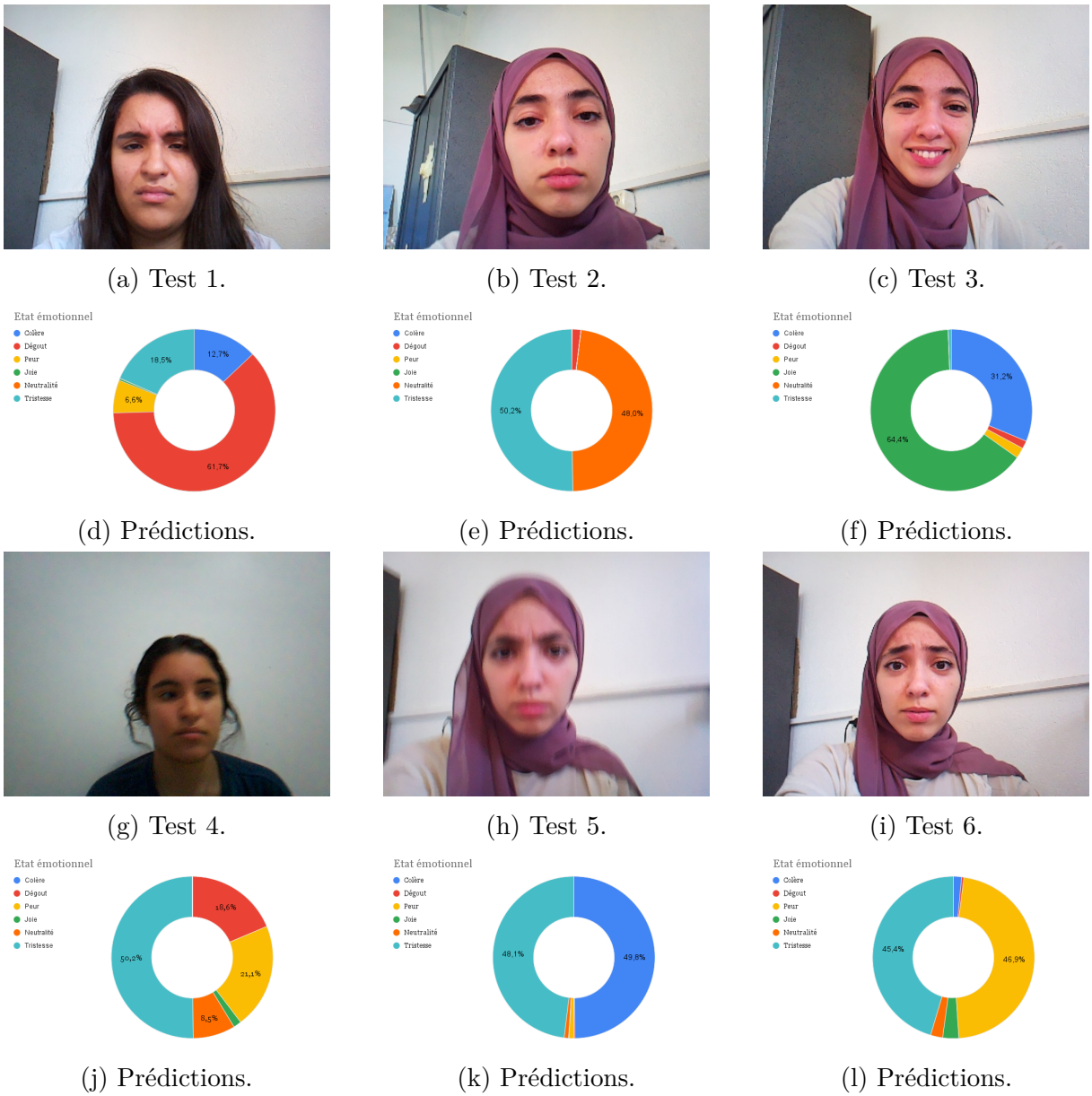
Fig. 4.10 : Résultats affichés sur la fenêtre du terminal.

Le tableau 4.4 récapitule les résultats obtenus des prédictions en temps réel. Il indique l'émotion dominante selon les deux modèles RVE et RFE puis la décision finale est prise lors de la fusion des deux vecteurs de scores de prédictions.
Nous observons que les deux modèles ne parviennent pas à détecter les émotions avec des probabilités identiques. Par exemple, une personne peut sembler neutre selon son expression faciale, mais triste selon sa voix en raison d'une faible énergie dans le signal vocal, comme illustré dans le test 2. Ceci met en évidence l'importance de combiner plusieurs sources d'information pour une meilleure détection des émotions.

Tab. 4.4 : Prédictions des deux modèles RFE et RVE puis la fusion.

Test	Résultats de la RFE	Résultats de la RVE	Résultats de la fusion des deux modalités
1	Dégoût	Dégoût	Dégoût
2	Neutralité	Tristesse	Tristesse
3	Joie	Colère	Joie
4	Tristesse	Tristesse	Tristesse
5	Tristesse	Colère	Colère
6	Tristesse	Peur	Peur
7	Peur	Neutralité	Neutralité
8	Peur	Neutralité	Peur
9	Joie	Dégoût	Joie

Les figures 4.11 et 4.12 illustrent les images des tests du tableau 4.4 ainsi que les vecteurs de prédictions de la fusion.



(a) Test 1. (b) Test 2. (c) Test 3.

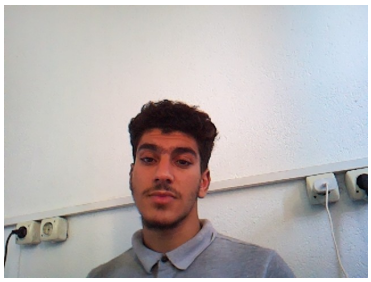
(d) Prédictions. (e) Prédictions. (f) Prédictions.

(g) Test 4. (h) Test 5. (i) Test 6.

(j) Prédictions. (k) Prédictions. (l) Prédictions.

FIG. 4.11 : Résultats des tests effectués en temps réel.

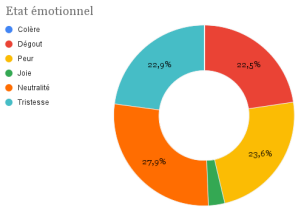
Dans le cas du test 2, nous constatons deux probabilités dominantes, celles de la tristesse et de la neutralité. En effet, ces deux émotions partagent des caractéristiques similaires en termes d'expressions faciales et de signaux vocaux. Les deux émotions se manifestent par une faible énergie dans les signaux de la parole.



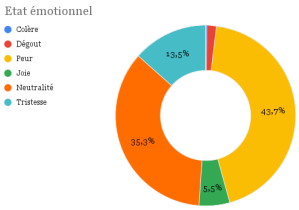
(a) Test 7.

(b) Test 8.

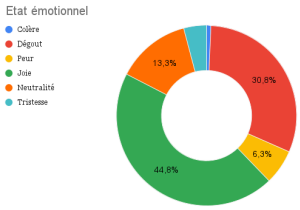
(c) Test 9.



(d) Prédictions.



(e) Prédictions.



(f) Prédictions.

Fig. 4.12 : Résultats des tests effectués en temps réels.

Concernant le test 8, notre sujet de test, possède une voix basse et avec une image prise d'un angle incliné , ce qui explique la confusion induite au niveau de l'émotion exprimée en détectant l'émotion peur. D'après les prédictions du test 9 la personne est joyeuse , or que l'émotion exprimée par était sensé être la colère cela montre la difficulté de généraliser un modèle surtout avec une architecture simple et un ensembles de données relativement petit.

4.4.3 Temps d'exécution

Afin d'évaluer notre application, nous avons comparé les temps d'exécution des trois étapes principales : l'extraction des caractéristiques des signaux de la parole, la détection du visage et le temps d'inférence pour les deux modalités,ainsi que leurs fusion, dans les deux environnements : Colab et Raspberry Pi. Les résultats sont présentés dans le tableau 4.5 ci-dessous.

Tab. 4.5 : Comparaison du temps d'exécution entre la Raspberry Pi et Google Colab.

Temps (ms)	Extraction des caractéristiques audio	Détection du visage	Inférence		
			RFE	RVE	Fusion
Raspberry Pi	2000∼7000	100∼145	∼236	3∼6	∼243
Google Colab	∼7000	∼127	∼66	170∼425	∼236

L'étape d'extraction des caractéristiques est la plus gourmande en terme de temps d'exécution, sa durée augmente avec l'augmentation de la durée de l'enregistrement vocal. L'inférence de la Reconnaissance Vocale des Emotions (RVE) est plus rapide sur Raspberry Pi que sur Google Colab, ce qui démontre l'efficacité de l'utilisation de TensorFlow Lite. Cependant, l'inférence de la Reconnaissance Faciale Emotionnelle (RFE) prend beaucoup

de temps, car elle est basée sur un modèle CNN 2D. En fin de compte, le temps d'inférence sur Raspberry Pi et sur Google Colab est similaire.

4.5 Conclusion

Ce chapitre constitue le résultat final de notre projet, qui consiste en l'implémentation de la solution établie. Nous avons abordé les aspects matériels et logiciels utilisés, ainsi que les étapes de configuration de la carte électronique et de conversion des modèles en une version plus légère. Enfin, nous avons présenté les résultats des prédictions en temps réel avec une comparaison du temps d'exécution entre la Raspberry Pi et Google Colab.

Ce chapitre nous a permis d'évaluer la capacité de généralisation de notre modèle et de franchir les premières étapes vers l'intégration des émotions humaines dans le domaine des systèmes embarqués.

Conclusion et perspectives

L'objectif de notre travail est de développer un système de reconnaissance automatique des émotions humaines a travers les expressions faciales et les signaux de la parole et enfin l'implémentation sur une carte Raspberry Pi.

En premier lieu, nous avons effectué une évaluation des caractéristiques des signaux audio afin de sélectionner ceux qui sont pertinents pour notre problématique. Ensuite nous avons établi la classifications des MFCCs, des spectrogrammes et les spectrogrammes de Mel à l'aide des réseaux de neurones convolutifs sur les quatre bases de données TESS, SAVEE, EMO DB et RAVDESS. En deuxième lieu, nous avons exploité la possibilité de classifier les caractéristiques HoG des images d'abord le SVM puis le CNN. De plus, nous avons étudié la possibilité de la classification des images des contours avec le CNN. A la fin nous avons choisi les deux modèles finaux à fusionner et implémenter et nous avons effectué des prédictions sur des données prises en temps réel.

À travers les expériences réalisées, nous concluons que :

- Les MFCCs sont les caractéristiques qui fournissent les meilleurs résultats pour la RVE.
- L'augmentation des données garantit des performances meilleures et une éventuelle généralisation .
- Les signaux de parole permettent une meilleure détection de certaines émotions, telles que la colère, tandis que les images faciales sont plus efficaces pour détecter d'autres émotions, comme la joie.
- La fusion permet une amélioration des résultats jusqu'à 96%.

Perspectives

Bien que nous ayons atteint notre objectif principal, ce travail n'est pas une finalité en soi et ne représente que le premier pas vers le développement d'un produit final, parmi les perspectives éventuelles en vue de l'évolution de ce projet nous citant :

- Collecte de données supplémentaires : afin d'améliorer les performances et rendre le système plus général, il est intéressant d'étendre l'ensemble d'apprentissage davantage de données pertinentes.

- Intégration dans des domaines spécifiques.
- Détection d'émotions complexes : En plus de la simple reconnaissance des expressions faciales et des signaux de la parole, il est possible d'entraîner le modèle à reconnaître d'autres types d'émotions.
- Intégration d'un système dynamique qui prend en considération la variation temporelle dans les séquences d'images.
- Interface utilisateur conviviale : Développer une interface utilisateur conviviale sur le Raspberry Pi. Cela permettrait aux utilisateurs d'interagir facilement avec le système, de visualiser les résultats de reconnaissance et, éventuellement, de contrôler d'autres fonctionnalités associées.

Bibliographie

1. PLUTCHIK, Robert. The nature of emotions : Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*. 2001, t. 89, n° 4, p. 344-350.
2. NAGHRAOUI, Djihane ; TAMAZIRT, Melyssa. *Benchmark pour la reconnaissance automatique des émotions*. 2020. Thèse de doct. Directeur : Mme. HANDOUZI Wahida/Co-directeur : Mr. RIMOUCHE Ali.
3. DUBEY, Monika ; SINGH, Lokesh. Automatic emotion recognition using facial expression : a review. *International Research Journal of Engineering and Technology (IRJET)*. 2016, t. 3, n° 2, p. 488-492.
4. MARÍN-MORALES, Javier ; HIGUERA-TRUJILLO, Juan Luis ; GRECO, Alberto ; GUIXERES, Jaime ; LLINARES, Carmen ; SCILINGO, Enzo Pasquale ; ALCAÑIZ, Mariano ; VALENZA, Gaetano. Affective computing in virtual reality : emotion recognition from brain and heartbeat dynamics using wearable sensors. *Scientific reports*. 2018, t. 8, n° 1, p. 13657.
5. EKMAN, Paul ; FRIESEN, Wallace V. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*. 1978.
6. DU, Shichuan ; TAO, Yong ; MARTINEZ, Aleix M. Compound facial expressions of emotion. *Proceedings of the national academy of sciences*. 2014, t. 111, n° 15, E1454-E1462.
7. COWIE, Roddy ; DOUGLAS-COWIE, Ellen ; TSAPATSOULIS, Nicolas ; VOTSIS, George ; KOLLIAS, Stefanos ; FELLENZ, Winfried ; TAYLOR, John G. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*. 2001, t. 18, n° 1, p. 32-80.
8. LI, Wu ; ZHANG, Yanhui ; FU, Yingzi. Speech emotion recognition in e-learning system based on affective computing. In : *Third international conference on natural computation (ICNC 2007)*. IEEE, 2007, t. 5, p. 809-813.
9. SCHULLER, Björn ; RIGOLL, Gerhard ; LANG, Manfred. Hidden Markov model-based speech emotion recognition. In : *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03)*. Ieee, 2003, t. 2, p. II-1.
10. LALITHA, S ; MUDUPU, Anoop ; NANDYALA, Bala Visali ; MUNAGALA, Renuka. Speech emotion recognition using DWT. In : *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*. IEEE, 2015, p. 1-4.

11. MILTON, A ; ROY, S Sharmy ; SELVI, S Tamil. SVM scheme for speech emotion recognition using MFCC feature. *International Journal of Computer Applications*. 2013, t. 69, n° 9.

12. CHANDRASEKAR, Purnima ; CHAPANERI, Santosh ; JAYASWAL, Deepak. Emotion recognition from speech using discriminative features. *International Journal of Computer Applications*. 2014, t. 101, n° 16, p. 31-36.

13. HAN, Kun ; YU, Dong ; TASHEV, Ivan. Speech emotion recognition using deep neural network and extreme learning machine. In : *Interspeech 2014*. 2014.

14. ZHANG, Shiqing ; TAO, Xin ; CHUANG, Yuelong ; ZHAO, Xiaoming. Learning deep multimodal affective features for spontaneous speech emotion recognition. *Speech Communication*. 2021, t. 127, p. 73-81.

15. ZIELONKA, Marta ; PIASTOWSKI, Artur ; CZYŻEWSKI, Andrzej ; NADACHOWSKI, Paweł ; OPERLEJN, Maksymilian ; KACZOR, Kamil. Recognition of Emotions in Speech Using Convolutional Neural Networks on Different Datasets. *Electronics*. 2022, t. 11, n° 22, p. 3831.

16. MAHMOOD, Mayyadah R ; ABDULRAZZAQ, Maiwan B ; ZEEBAREE, S ; IBRAHIM, Abbas Kh ; ZEBARI, Rizgar R ; DINO, Hivi Ismat. Classification techniques' performance evaluation for facial expression recognition. *Indonesian Journal of Electrical Engineering and Computer Science*. 2021, t. 21, n° 2, p. 176-1184.

17. GHIMIRE, Deepak ; LEE, Joonwhoan. Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines. *Sensors*. 2013, t. 13, n° 6, p. 7714-7734.

18. PRANAV, E ; KAMAL, Suraj ; CHANDRAN, C Satheesh ; SUPRIYA, MH. Facial emotion recognition using deep convolutional neural network. In : *2020 6th International conference on advanced computing and communication Systems (ICACCS)*. IEEE, 2020, p. 317-320.

19. QUINN, Minh-An ; SIVESIND, Grant ; REIS, Guilherme. Real-time emotion recognition from facial expressions. *Standford University*. 2017.

20. MOLLAHOSSEINI, Ali ; CHAN, David ; MAHOOR, Mohammad H. Going deeper in facial expression recognition using deep neural networks. In : *2016 IEEE Winter conference on applications of computer vision (WACV)*. IEEE, 2016, p. 1-10.

21. ZHAO-YI, Peng ; YAN-HUI, Zhu ; YU, Zhou. Real-time facial expression recognition based on adaptive canny operator edge detection. In : *2010 Second International Conference on MultiMedia and Information Technology*. IEEE, 2010, t. 2, p. 154-157.

22. ABD, Raghad Ghalib ; IBRAHIM, Abdul-Wahab Sami ; NOOR, Ameen A. Facial Emotion Recognition Using HOG and Convolution Neural Network. *Ingénierie des Systèmes d'Information*. 2023, t. 28, n° 1.

23. LOPES, André Teixeira ; DE AGUIAR, Edilson ; DE SOUZA, Alberto F ; OLIVEIRA-SANTOS, Thiago. Facial expression recognition with convolutional neural networks : coping with few data and the training sample order. *Pattern recognition*. 2017, t. 61, p. 610-628.

24. PITALOKA, Diah Anggraeni ; WULANDARI, Ajeng ; BASARUDDIN, Tjan ; LILIANA, Dewi Yanti. Enhancing CNN with preprocessing stage in automatic emotion recognition. *Procedia computer science*. 2017, t. 116, p. 523-529.

25. BREUER, Ran ; KIMMEL, Ron. A deep learning perspective on the origin of facial expressions. *arXiv preprint arXiv :1705.01842*. 2017.

26. RISTEA, Nicolae-Cătălin ; DUŢU, Liviu Cristian ; RADOI, Anamaria. Emotion recognition system from speech and visual information based on convolutional neural networks. In : *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. IEEE, 2019, p. 1-6.

27. WANG, Yutai ; YANG, Xinghai ; ZOU, Jing. Research of emotion recognition based on speech and facial expression. *TELKOMNIKA Indonesian Journal of Electrical Engineering*. 2013, t. 11, n° 1, p. 83-90.

28. PAN, Jiahui ; FANG, Weijie ; ZHANG, Zhihang ; CHEN, Bingzhi ; ZHANG, Zheng ; WANG, Shuihua. Multimodal Emotion Recognition based on Facial Expressions, Speech, and EEG. *IEEE Open Journal of Engineering in Medicine and Biology*. 2023.

29. SONG, Kyu-Seob ; NHO, Young-Hoon ; SEO, Ju-Hwan ; KWON, Dong-soo. Decision-level fusion method for emotion recognition using multimodal emotion recognition information. In : *2018 15th International Conference on Ubiquitous Robots (UR)*. IEEE, 2018, p. 472-476.

30. KUDIRI, Krishna Mohan ; SAID, Abas Md ; NAYAN, M Yunus. Human emotion detection through speech and facial expressions. In : *2016 3rd International Conference on Computer and Information Sciences (ICCOINS)*. IEEE, 2016, p. 351-356.

31. WANG, Yongjin ; GUAN, Ling. Recognizing human emotional state from audiovisual signals. *IEEE transactions on multimedia*. 2008, t. 10, n° 5, p. 936-946.

32. TZIRAKIS, Panagiotis ; TRIGEORGIS, George ; NICOLAOU, Mihalis A ; SCHULLER, Björn W ; ZAFEIRIOU, Stefanos. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of selected topics in signal processing*. 2017, t. 11, n° 8, p. 1301-1309.

33. ALLUHAIDAN, Ala Saleh ; SAIDANI, Oumaima ; JAHANGIR, Rashid ; NAUMAN, Muhammad Asif ; NEFFATI, Omnia Saidani. Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network. *Applied Sciences*. 2023, t. 13, n° 8, p. 4750.

34. ISSA, Dias ; DEMIRCI, M Fatih ; YAZICI, Adnan. Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control*. 2020, t. 59, p. 101894.

35. NEMA, Bashar M ; ABDUL-KAREEM, Ahmed A. Preprocessing signal for speech emotion recognition. *Al-Mustansiriyah Journal of Science*. 2018, t. 28, n° 3, p. 157-165.

36. VIOLA, Paul ; JONES, Michael. Rapid object detection using a boosted cascade of simple features. In : *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*. Ieee, 2001, t. 1, p. I-I.

37. CANNY, John. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*. 1986, n° 6, p. 679-698.

39. DALAL, Navneet ; TRIGGS, Bill. Histograms of oriented gradients for human detection. In : *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*. Ieee, 2005, t. 1, p. 886-893.

40. FUKUSHIMA, Kunihiko. Neocognitron : A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*. 1980, t. 36, n° 4, p. 193-202.

41. LECUN, Yann ; BOSER, Bernhard ; DENKER, John S ; HENDERSON, Donnie ; HOWARD, Richard E ; HUBBARD, Wayne ; JACKEL, Lawrence D. Backpropagation applied to handwritten zip code recognition. *Neural computation*. 1989, t. 1, n° 4, p. 541-551.

43. PICHORA-FULLER, M. Kathleen ; DUPUIS, Kate. *Toronto emotional speech set (TESS)*. Borealis, 2020. VERSION PROVISOIRE. Disp. à l'adr. DOI : 10.5683/SP2/E8H2MF.

44. LUCEY, Patrick ; COHN, Jeffrey F. ; KANADE, Takeo ; SARAGIH, Jason ; AMBADAR, Zara ; MATTHEWS, Iain. The Extended Cohn-Kanade Dataset (CK+) : A complete dataset for action unit and emotion-specified expression. In : *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*. 2010, p. 94-101. Disp. à l'adr. DOI : 10.1109/CVPRW.2010.5543262.

46. LIVINGSTONE, Steven R. ; RUSSO, Frank A. *The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)*. Zenodo, 2018. Version 1.0.0. Disp. à l'adr. DOI : 10.5281/zenodo.1188976. Funding Information Natural Sciences and Engineering Research Council of Canada : 2012-341583 Hear the world research chair in music and emotional speech from Phonak.

47. CNN, Convolutional Neural Network. Speech emotion recognition using convolutional neural network (CNN). *International Journal of Psychosocial Rehabilitation*. 2020, t. 24, n° 8, p. 1-20.

48. BHANDARI, Arkaprabha ; PAL, Nikhil R. Can edges help convolution neural networks in emotion recognition ? *Neurocomputing*. 2021, t. 433, p. 162-168.

49. M.K., Linga ; MODEPALLI, Divyanjali ; SHAIK, Maibu ; BUSI, Madhuri ; VENKATAIAH, C. ; Y., Mallikarjuna ; ALKHAYYAT, Ahmed ; RAWAT, Divya. Efficient Feature Extraction for Recognition of Human Emotions through Facial Expressions Using Image Processing Algorithms. *E3S Web of Conferences*. 2023, t. 391. Disp. à l'adr. DOI : 10.1051/e3sconf/202339101182.

50. KUMAR, Pranav ; HAPPY, SL ; ROUTRAY, Aurobinda. A real-time robust facial expression recognition system using HOG features. In : *2016 International Conference on Computing, Analytics and Security Trends (CAST)*. IEEE, 2016, p. 289-293.

51. SANDERSON, Conrad ; PALIWAL, Kuldip. Information Fusion and Person Verification Using Speech Face Information. *Technical Report IDIAP-RR 02-33, IDIAP*. 2004.

52. LUNA-JIMÉNEZ, Cristina ; GRIOL, David ; CALLEJAS, Zoraida ; KLEINLEIN, Ricardo ; MONTERO, Juan M ; FERNÁNDEZ-MARTÍNEZ, Fernando. Multimodal emotion recognition on ravdess dataset using transfer learning. *Sensors*. 2021, t. 21, n° 22, p. 7665.

Webographie

38. MA, Rui. *Parametric speech emotion recognition using neural network*. 2014.

42. JACKSON, Philip ; HAQ, Sanaul. *SAVEE : Surrey Audio-Visual Expressed Emotion Database*. 2015. Aussi disponible à l'adresse : http://kahlan.eps.surrey.ac.uk/savee/. Last update : 2 April 2015.

45. *Karolinska Directed Emotional Faces (KDEF)* [Online Database]. n.d. Aussi disponible à l'adresse : https://kdef.se/.

53. LDLC. *Fiche produit du Raspberry Pi 4* [https://www.ldlc.com/fiche/PB00343113.html]. 2023. Consulté le 17 juin 2023.