République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université de 8 Mai 1945-Guelma-
Faculté des Mathématiques, d'Informatique et des Sciences de la matière
Département d'Informatique

Mémoire de fin d'études Master

Filière : Informatique

Option :

Système informatique

Thème

Nouvelle approche de reconnaissance multimodale des émotions

Présenté par : *M^r* BELHIRECHE MOHAMED

Membres du jury :

N	NOM & PRENOM	Qualité
1	*D^r* DERDAR SALAH	Président
2	*P^r* KOUAHLA MOHAMED NADJIB	Encadreur
3	*D^r* BORJIBA YAMINA	Examinateur

juin 2023

Remerciements

C'est bien de célébrer le succès, mais le plus important est de tirer des leçons de l'échec.

Avant tout, je voudrais remercier Dieu de m'avoir donné la force, la connaissance, la capacité et l'opportunité d'entreprendre ce travail, de persévérer et de la terminer de manière satisfaisante.

Je tiens à remercier sincèrement mon encadreur Professeur ***KOUAHLA MOHAMED NADJIB*** *pour m'avoir dirigé et ses encouragements dans ce travail. Merci de m'avoir fait confiance et pour tout ce que vous avez fait pour m'amener jusqu'ici.*

Je remercie également mes parents pour tous les sacrifices qu'ils ont consentis pour me permettre de poursuivre mes études dans les meilleures conditions .

Enfin, je remercie tous les amis, professeurs et travailleurs que j'ai connus tout au long du parcours universitaire.

Dédicace

Je dédie ce travail

À mes très chers parents

À mes parents, qui a travaillé sur ma réussite par leur amour et leur soutien tout au long de ma carrière scolaire et tous les sacrifices faits et ses précieux conseils, pour toute son aide et sa présence dans ma vie, elle reçoit à travers cette humble œuvre comme elle est, l'expression de mes sentiments et ma gratitude éternelle.

À mes très chers MALAK,KAWTHER & BOUTHAYNA

Je voudrais exprimer ma gratitude à mes belles sœurs MALAK,KAWTHER et ma jolie BOUTHAYNA pour son amour, son soutien et sa présence dans ma vie pour votre gentillesse, votre patience et votre compréhension. Tu es quelqu'un de bien et j'ai de la chance de t'avoir comme mes soeurs.

À ma famille

Mes chers cousins AYOUB ,ABDEL-JALIL,ISSAM ,OUSSAMA ,HOUSSEM,KHALIL,HAYTHEM et mes oncles, je tiens à vous remercier pour votre amour inconditionnel, votre soutien indéfectible et votre patience infinie. Vous êtes ma famille et je suis tellement fière de vous avoir dans ma vie.

Sans oublier mes très chers amis ***ADEL, SAYED, LOUFI,YACINE,SOFIANE CHIPA,HOZDA,ZAKI,AKREM,ANIS,*** *et tous ceux qui m'ont soutenu et encouragé durant toute la période de mon travail...*

MOHAMED

Résumé

Au cours de la dernière décennie, les systèmes intelligents ont fourni des services publics, ce qui a entraîné un changement important dans le milieu de vie des gens, mais il reste encore de nombreux problèmes à régler pour s'assurer qu'ils travaillent ensemble de façon sécuritaire, fiable et efficace. La recherche dans le domaine l'IA mettra dorénavant l'accent sur la résolution de ces problèmes. Nous avons axé nos recherches sur la reconnaissance des émotions afin d'élaborer une stratégie intelligente pour proposer un système multimodal de reconnaissance des émotions.

Pour construire notre propre système, nous avons d'abord créer un modèle basé sur des techniques d'apprentissage profond où l'audio et le texte ont été sélectionnés dans le processus de reconnaissance des émotions multimodale. Nous avons utilisé les caractéristiques du domaine des fréquences comme entrées à la source audio et nous utilisons la procédure de tokanisation pour traiter les textes et les utilités comme entrées à la source du texte. Pour réaliser ce système, nous avons introduit une technique d'intégration appelée "Fusion Network " qui vise à obtenir de meilleures performances en incluant des modèles et en combinant plusieurs résultats de réseaux neuronaux indépendants.

Pendant la phase expérimentale, le système de reconnaissance des émotions multimodale a été développé dans l'environnement colab avec l'ensemble de données IEMOCAP, permettant la création d'un modèle de prédiction sophistiqué avec un taux de précision de 72%.

Mot clés : *reconnaissance des emotion, multimodale, technique de fusion, modalité,plongement lexical.*

Abstract

In the last decade, intelligent systems have provided public services, that has resulted in a major change in the environment of people's lives, but it still remains There are many problems to solve in order to ensure that they work together safely. reliable and effective. Research on AI will focus on resolving these problems. and problems. We focused our research on the recognition of emotions in order to develop a Intelligent strategy to propose a multimodal emotion recognition system.

To build our own system, we must first create a model based on deep learning techniques where audio and text were selected in the process of Recognition of Multimodal Emotions. We used the characteristics of the frequencies as input to the audio source and we support the tokanization procedure to process the texts and use them as input to the source of the text. In order to create a system, We have introduced a technique of integration called "Fusion Network" which aims to obtain Improved performance by including models and combining multiple network results and independent neurons.

During the experimental phase, the multi-emotional recognition system modale was developed in the collaborative environment with the IEMOCAP data set, This allows the creation of a sophisticated prediction model with a rate of accuracy of 71

Key words : *emotional recognition, multimodal, fusion technique, modality , word embedding.*

Table des matières

Table des figures

Liste des tableaux

Abréviations et Acronymes

<BERT> <Bidirectional Encoder Representations from Transformers>

<GLOVE> < Global Vectors >

<IEMOCAP> <Interactive Emotional Dyadic Motion Capture>

<LPC> <Linear prediction coefficients>

<MFCC> <Mel-frequency cepstral coefficients>

<NLP> <Natural Language Processing>

<NLTK> <Natural Language Tool Kit>

<NUMPY> <NUMerical PYthon>

<PLP> < Perceptual linear prediction>

Introduction générale

La reconnaissance des émotions est un dilemme pour les chercheurs depuis longtemps et devient maintenant une réalité où la machine informatique est plus capable que jamais de comprendre et d'interpréter les comportements humains.Il existe six expressions émotionnelles universelles selon Ekman (dégoûts, colère, joie, tristesse, surprise et peur)[1].Ces émotions peuvent être comprises à partir de plusieurs sources comme les expressions faciales,mouvement oculaire,le langage corporel et le son. Pour que la machine puisse comprendre ces concepts-là, ces derniers devront être convertis en une forme qu'elle peut comprendre,et ceci est en extrayant des caractéristiques spécifiques puis les représenter sous forme de vecteurs par exemple.

De nombreux travaux ont introduit des systèmes de reconnaissance des émotions unimodale , qui se concentrerait sur l'analyse d'une seule modalité, telle que la voix ou le visage, comporte certaines limites. Ces limites peuvent être que les signaux dans une seule modalité sont vagues et difficiles à interpréter,de plus ces systèmes unimodale sont sensibles au bruit et aux variations dans les données d'une seule modalité. Les systèmes multimodale ont donc été utilisés pour établir ces limites.

L'objectif de ce projet est de proposer une approche pour la reconnaissance des émotions multimodale basée sur les techniques de fusion .Un module principal a été développé pour la reconnaissance des émotions à partir deux sources de données.
Ce mémoire est réparti sur quatre chapitres différents :
Une introduction générale situant le contexte et explicitant la problématique et les objectifs à atteindre ainsi que la structure du mémoire .
Chapitre 1 : Ce chapitre est dédié au terme de reconnaissance des émotions,nous présentons les concepts de base pour les reconnaissances des émotions tels que les définitions, les modélisations,les sources et les caractéristiques des données émotionnelles.
Chapitre 2 : Nous avons consacré ce chapitre aux systèmes de reconnaissances des émotions multimodales, où nous avons commencé par la définition de la multimodalité et les techniques de fusions, puis nous avons décrit les moyens et les méthodes qui permettent le bon fonctionnement de ces systèmes. Enfin, nous citons des travaux pertinents pour ces systèmes afin de bénéficier et exploiter les avantages de leurs approches.
Chapitre 3 : Présente notre conception, qui est décrite par la modélisation des modules pour les deux sources textuelle et vocale , suivie par la modélisation de module du fusion des deux sources utilisant une technique de fusion appelez "fusion network".
Chapitre 4 : Ce chapitre est consacré à la mise en œuvre et l'expérimentation de notre ap-

proche. Nous présentons d'abord les outils de développement et les environnements de travail, puis l'expérimentation et les évaluations des résultats du module de reconnaissance d'émotion multimodale sur les données de "IEMOCAP".

Une conclusion générale : Présente les résultats finals et notre point de vue sur le domaine de reconnaissance d'émotions multimodale et les perspectives.

Première partie

État de l'art

Chapitre 1

La reconnaissance de l'émotion

1 Introduction

La reconnaissance d'émotion est un domaine d'étude visant à comprendre et à détecter les émotions humaines dans les expressions faciales et le son. La reconnaissance des émotions est importante pour améliorer la communication entre les gens, promouvoir les relations sociales et faciliter la prise de décision. Les progrès technologiques dans la reconnaissance des émotions ont conduit au développement d'applications telles que la reconnaissance vocale, la détection des mensonges et l'analyse des sentiments sur les réseaux sociaux pour faciliter les interactions sociales.

Dans ce chapitre, nous proposons un aspect théorique sur la reconnaissance des émotions en expliquant différents concepts, depuis la définition du terme émotion et la reconnaissance d'émotions en utilisant toutes les sources d'information.

2 Cadre théoriques du projet

2.1 Emotion :

Une réaction complexe qui engage à la fois le corps et un processus psychophysiologique déclenché par la perception consciente et/ou inconsciente d'un objet ou d'un événement. Elle est souvent associée à l'humour, au tempérament, à la personnalité, à la disposition et à la motivation. Le rôle des émotions dans la communication humaine, la prise de décision, l'interaction et les processus cognitifs est crucial. L'émotion est basée sur des expériences subjectives, que les gens représentent avec une variété de termes sémantiques[2].

2.1.1 L'émotion en psychologie :

En psychologie, Piéron définit l'émotion comme une réaction émotionnelle d'intensité modérée qui dépend des centres diencéphales et se manifeste généralement comme des symptômes végétatifs. Il y a une conscience de l'émotion, bien que son intensité semble varier[3].

2.1.2 L'émotion en philosophie :

En philosophie, l'émotion peut être généralement définie comme une expression de la vie émotionnelle qui est généralement accompagnée d'un état de conscience agréable ou désagréable. L'émotion est une perturbation à court terme et une rupture de déséquilibre. Parfois la question est violente et entraîne une augmentation du mouvement (enthousiasme, colère), ou d'autre part, une cessation du mouvement (peur ou "coup de foudre" dans l'amour)[4].

2.2 Modélisation des émotions :

2.2.1 Modélisation catégorielle :

Les approches discrètes consistent à considérer les émotions comme des caractéristiques universelles et épisodiques qu'est représentent groupe d'émotions comme un ensemble discret dans lequel chaque type d'émotion est désigné par une étiquette spécifique. Elle est basée sur un ensemble d'émotions qui sont considérées comme fondamentales, universelles, irréductibles et innées. Quelques émotions de base (peur, tristesse, etc.) qui peuvent être observées chez les gens de toutes nationalités et cultures sont définies par la nature universelle des émotions , néanmoins, il y a désaccord sur le nombre et la nature de ces soi-disant émotions fondamentales. Le principal avantage est qu'une fois que les émotions sont clairement identifiées pour le traitement, elles sont plus faciles à manipuler[5].

Pour comprendre les émotions, des chercheurs comme Ekman et d'autres ont proposé différents modèles qui tentent de les expliquer et de les classer,illustré dans le tableau suivant.

Izard (1977)	Plutchik (1980)	Tomkins (1980)	Panksepp(1989)	Ekman (1992)
Colère	Colère	Colère	Colère	Colère
Détresse	Dégoût	Dégoût	Peur	Dégoût
Joie	Joie	Joie	Attente	Joie
Peur	Peur	Peur	Panique	Peur
Surprise	Surprise	Surprise		Surprise
Tristesse	Tristesse	Mépris		Tristesse
Mépris	Acceptation	Honte		
Honte	Anticipation	Intérêt		
Intérêt		Détresse		
Culpabilité				
Amour				

TABLE 1.1 – Les différentes émotions catégorielles [1].

2.2.2 Modélisation dimensionnelle :

L'approche dimensionnelle propose de représenter les émotions dans un espace multidimensionnel sur la base qu'elles sont produites par un nombre fixe de concepts illustré dans la figure 1.1. Ces dimensions varient selon les exigences du modèle, mais le modèle de Russell avec ses dimensions de valence et d'activation est le plus populaire. Les émotions positives comme la

joie et les émotions négatives comme la colère peuvent être distinguées par valence. L'activation représente le niveau d'excitation corporelle.Le principal avantage est de représenter les émotions sans l'utilisation d'étiquettes et possible d'associer certaines zones à des étiquettes émotionnelles[5].

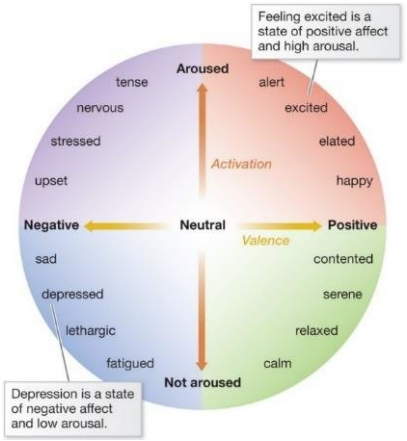

FIGURE 1.1 – Modélisation dimensionnelle [6].

2.3 Systèmes de reconnaissance d'émotions (unimodale) :

Afin de reconnaître l'émotion, il y a plusieurs étapes de base à suivre illustré dans la figure 1.2. La première phase consiste à recueillir et à traiter des données à l'aide de divers capteurs ou outils techniques. Ensuite, l'extraction des caractéristiques depuis ces données pour obtenir autant d'informations que possible de l'entrée. De nombreuses méthodes utilisées pour extraire les caractéristiques, qui peuvent dépendre des caractéristiques géométriques, des caractéristiques statistiques, les caractéristiques de texture,et dans la dernière phase sont classées via ces caractéristiques extraites pour déterminer les règles de classification des objets en catégories basées sur les variables qualitatives ou la quantité qui caractérisent ces objets[7].

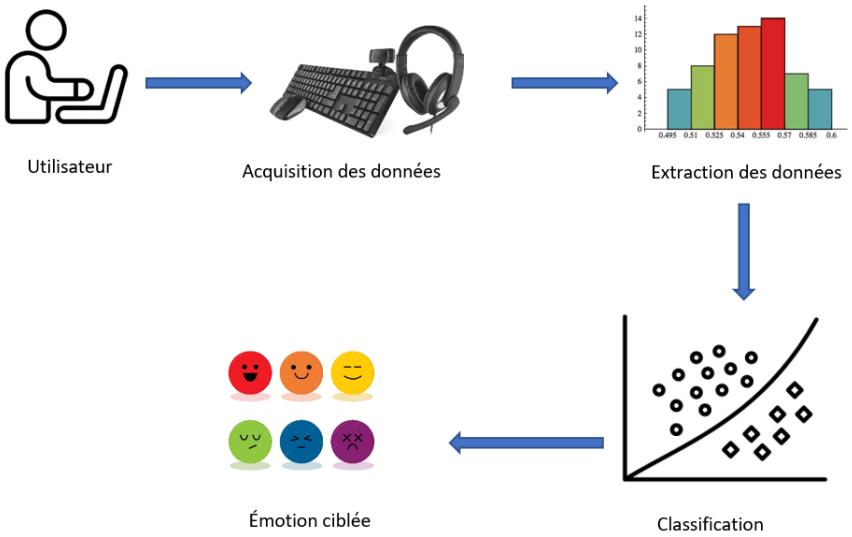

FIGURE 1.2 – Le processus de reconnaissance d'émotion.

2.4 Principaux sources des données émotionnelles :

Les émotions peuvent être identifiées par de nombreuses sources illustrées dans la figure 1.3 , notamment visuelles, textuelles, audio ou physiologiques, mais ces sources diffèrent dans la façon dont les émotions sont capturées. Dans les paragraphes ce qui suit nous présentons la description des différentes Sources.

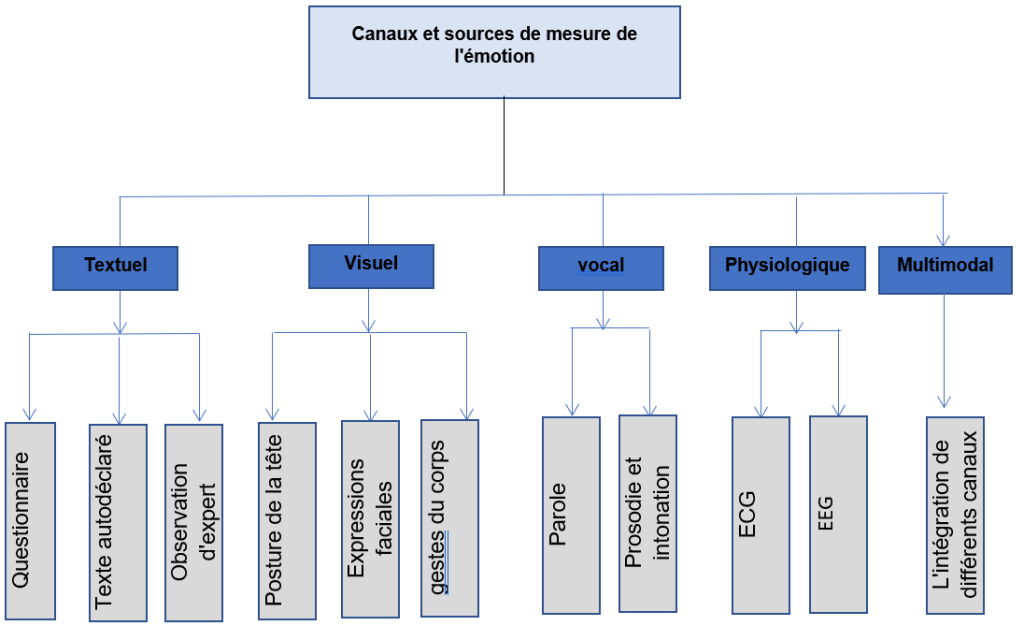

FIGURE 1.3 – Sources d'informations émotionnelles.

2.4.1 Source Textuel :

Une méthode de reconnaissance des émotions au moyen de textes consiste à utiliser le questionnaire où une série de questions sont posées et les réponses sont distribuées en plusieurs degrés d'émotions selon la nature de la réponse. On peut demander aux participants de répondre à des questions au moyen de classifications numériques ou de réponses pré sélectionnées, ou de fournir des réponses ouvertes à des questions plus qualitatives ce qu'ils peuvent être conçus pour évaluer différents aspects des émotions[8].On peut reconnaître les émotions à travers des textes pures où références linguistiques telles que les mots, les phrases et la structure des phrases aident à identifier les émotions à leur sujet dans le texte. Mais les émotions peuvent être implicites dans le contexte dans lequel le texte est écrit qui affectent le sens des mots et des expressions[9]. Pour les caractéristiques, nous pouvons citer :

- Plongement lexical.
- Caractéristique de sémantique.

2.4.2 Source visuel :

L'émotion peut être reconnue dans une large proportion par des sources visuelles en raison de son contact direct avec les relations extérieures, la position de la tête peut symboliser

de nombreuses connotations où différents mouvements de la tête peuvent toujours provoquer des changements dans l'apparence du visage [10].Bien que les humains communiquent non seulement par les mots, mais aussi par les mouvements du corps. Ces mouvements non verbaux peuvent fournir des renseignements importants sur l'état émotionnel d'une personne. Par exemple, lorsque le corps est joyeux, il traverse une gamme de mouvements impliquant les mains la tête et d'autres parties[11].

FIGURE 1.4 – Les expressions faciales[12].

La composante la plus cruciale du processus de communication est les expressions faciales illustrées dans la figure 1.4, qui sont également une source importante d'information sur les états émotionnels des gens et la transmission de signaux non verbaux lors de rencontres sociales. En fait, certains muscles sont déclenchés lorsqu'une personne éprouve des émotions, ce qui permet des actions comme froncer les sourcils ou sourire. De plus, la recherche d'Eckman dans le domaine de l'analyse automatique de l'expression faciale se concentre maintenant plus sur six expressions faciales émotionnelles primaires (joie, surprise, peur, tristesse, dégoût et colère)[13].

En plus des mouvements oculaires qui représentent une autre source de reconnaissance des émotions, il y a d'autres sources comme le fait de dynamique des frappes et les mouvements de souris qui nécessitent des outils techniques pour faire le suivi. Pour les caractéristiques, nous pouvons citer :

- Caractéristiques géométriques.
- Caractéristiques d'apparence.

2.4.3 La source vocale :

Le discours ou la voix est un signal complexe qui transmet de l'information sur le locuteur, le message, la langue, l'humeur, etc. L'information importante comme l'intention du locuteur peut être transmise par une communication non verbale. La façon dont les mots sont parlés envoyés des informations non linguistiques essentielles en plus du message qui est communiqué

par le texte. Le même message texte sera transmis avec de multiples implications en ajoutant les émotions pertinentes. Plusieurs interprétations du texte opérationnel peuvent être possibles selon la façon dont il est exprimé[14]. Pour les caractéristiques, nous pouvons citer :

- Mel-Frequency Cepstral Coefficients.
- Linear Prediction Coefficients.
- Perceptual Linear Prediction.

2.4.4 La source Physiologique :

L'émotion est une condition psychophysiologique qui est à l'origine basée dans le cerveau. Ainsi, il est responsable de l'interprétation émotionnelle. Lorsque l'émotion est exprimée, il y a de l'activité et des stimuli dans le cerveau qui peuvent être captés par l'électroencéphalogramme (EEG), qui est représentée par un signal électrique[15].Bien que,l'Électrocardiographie (ECG) est utilisée comme signal physiologique comme méthode traditionnelle d'interprétation non chirurgicale de l'activité électrique du cœur en temps réel. Il peut également être utilisé pour reconnaître l'émotion. Lorsque l'émotion peut être suivie d'une augmentation ou d'une diminution du rythme cardiaque ou de la pression artérielle[16].Afin de capturer et de collecter ces stimuli cérébraux et cardiaques, des outils techniques sont utilisés pour être ensuite représentés comme des signaux électriques.

2.5 Représentations des caractéristiques des modalités :

Le système de reconnaissance émotionnelle dans lequel les sources que nous avons précédemment présentées sont utilisées est compris pour les humains plutôt que pour la machine, de sorte que ces sources sont représentées sous la forme d'un concept de la machine en utilisant des techniques spéciales selon la nature de chaque source.

2.5.1 Caractéristique textuel :

- **Plongement lexical** :C'est la méthode utilisée pour représenter des mots à un vecteur numérique caractérisé par la densité, la longueur constante et sa représentation dépend de deux bases soient sur la base de la prédiction ou sur la base de la fréquence[17]. Il existe plusieurs techniques pour la vectorisation de texte bien que :

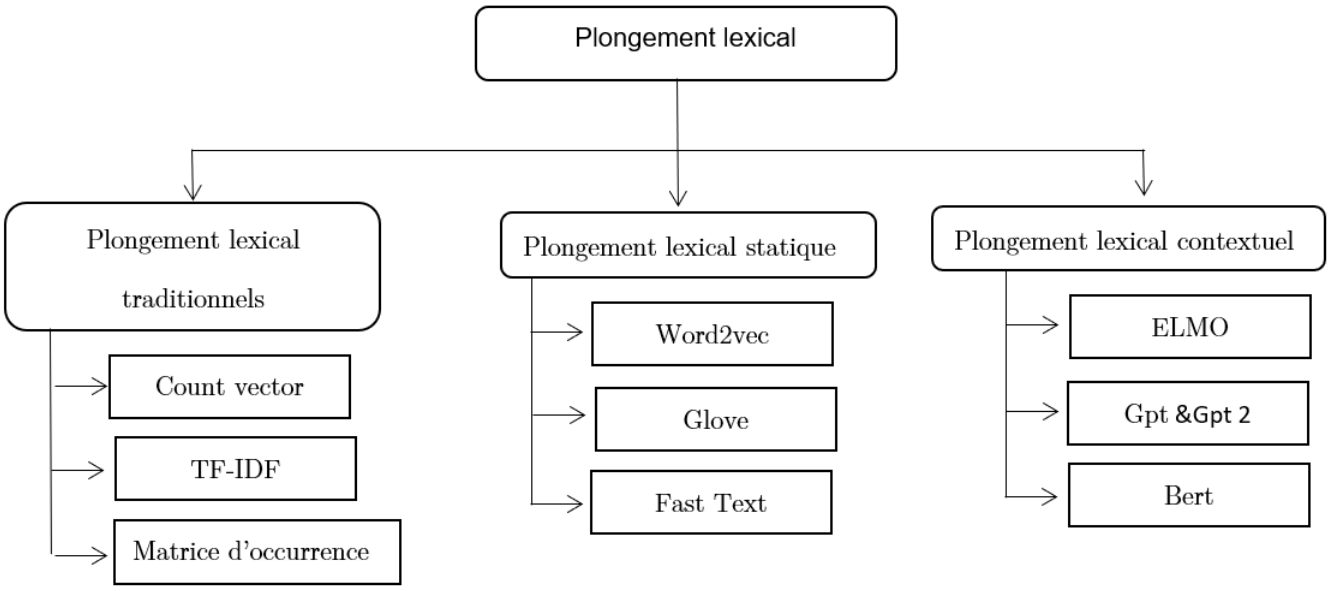

FIGURE 1.5 – Type des techniques de plongement lexical[18].

(a) TF-IDF (term frequency-inverse document frequency) : Elle est déterminée comme le produit de la fréquence d'un mot dans un échantillon spécifique et sa fréquence sur tout le texte. Insiste sur la signification d'une phrase spécifique sur la base de sa rareté dans tous les documents et rend facile de calculer à quel point deux textes sont similaires[18].

(b) WORD2VEC :C'est une approche qui utilise une représentation dense pour produire l'intégration de mots. C'est un modèle de prédiction qui attribue la probabilité aux mots qui font bien dans les défis de similarité de mots. En mappant le mot cible à son mot de contexte, il peut convertir un corpus non étiqueté en informations étiquetées[18]. Par exemple : v("queen") − v("woman") + v("man") ≈ v("king") se signifie qu'ensemble de vecteurs queen women et man sont similaires avec le vecteur de king .

(c) BERT : Un encodeur de transformateur bidirectionnel multicouche qu'est interne se compose de deux parties. Tout d'abord, une intégration initiale pour chaque jeton est créée en combinant un texte préformé avec des informations de position et de segment. Ensuite, cette séquence initiale des cartes du texte passe par plusieurs couches de transformateurs, produisant une nouvelle séquence des cartes contextuelles [19].Bien qu'il ait appris les relations contextuelles entre les mots ou les sous-mots. Il contient des significations syntaxiques et sémantiques d'un texte.

2.5.2 Caractéristique visuel :

Les traits du visage sont principalement liés aux composantes du visage comme les yeux, la bouche, les sourcils, le nez et le menton. Lorsque les expressions faciales sont formées accompagnées de contractions faciales qui affectent l'emplacement de ces composants de la hauteur des sourcils ou de la fermeture des yeux,on peut se résumer comme des caractéristiques géométriques ou des caractéristiques d'apparence.

- **Caractéristiques géométriques :** Les traits mesurent les mouvements de traits particuliers du visage, y compris les coins des lèvres ou des sourcils illustré dans la figure 1.6. Pour créer un vecteur caractéristique qui reflète la géométrie du visage, les composantes faciales, ou points de caractéristique faciale, sont extraits. En surveillant le mouvement de certains points faciaux, l'expression faciale sous-jacente peut être vérifiée. Selon la technique basée sur la géométrie, les émotions du visage influencent l'emplacement relatif et la taille de nombreuses caractéristiques. Le travail de mesure des caractéristiques géométriques implique souvent l'analyse de la région de taille, en particulier la localisation et le suivi des points clés dans la région de taille[20]. Différentes méthodes existent qui peuvent extraire ce type de caractéristique tels que :
  - Active Appearance Model (AAM).
  - Modèle de forme active (ASM).
- **Caractéristiques d'apparence :** Ce sont les caractéristiques physiques ou les attributs de l'apparence extérieure d'une personne qui peut être observée visuellement. Quand une certaine action est prise, les caractéristiques, telles que les rides, les bosses, le front, et les zones autour des lèvres et des yeux, changent la texture du visage. Afin d'extraire un vecteur caractéristique d'une image de visage, des filtres d'image sont utilisés et appliqués à l'ensemble du visage ou à certaines parties[20]. Différentes méthodes existent qui peuvent extraire ce type de caractéristique tels que :
  - Gabor wavelets.
  - Local Binary Pattern (LBP).

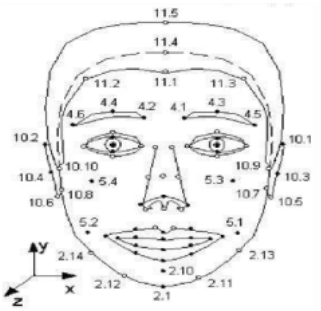

FIGURE 1.6 – Les points d'intérets visage [21].

2.5.3 Caractéristique vocal :

Les caractéristiques acoustiques sont utilisées dans la reconnaissance émotionnelle pour extraire et analyser les propriétés acoustiques de la parole, telles que le son, le ton, le rythme et la fréquence vocale. Ces caractéristiques peuvent être divisées en deux majorités domaine. Le domaine temporel représente les changements de signal dans le temps (la forme d'onde),toutes les caractéristiques temporelles du signal audio ont le trait d'être récupérées directement à partir de la source audio originale sans traitement préalable. Les caractéristiques de domaine de fréquence basées sur le contenu en fréquence du signal sonore[22]. De nombreuses techniques d'extraction des caractéristiques sont disponibles, notamment :

(a) Mel-Frequency Cepstral Coefficients (MFCC) : La représentation à court terme de la force permet de reconnaître des mots monosyllabiques dans des phrases prononcées de façon constante, mais non parlées. Utilisé logarithmiques à haute fréquence pour maintenir les propriétés de base des signaux vocaux acoustiquement, qui se composent souvent de tons avec des fréquences variables. MFCC s'appuie sur la désintégration du signal à l'aide d'une banque de filtres [23].

(b) Linear Prediction Coefficients (LPC) : C'est une technique qui mesure la fréquence et la concentration des résidus au-delà du signal vocal après avoir arrondi les formations et éliminé leurs impacts du signal vocal. Les fréquences de formation sont les fréquences auxquelles apparaissent les pics de résonance. En conséquence, en calculant les transactions prédictives linéaires à travers la fenêtre coulissante et en identifiant les pics dans le spectre de filtrage de prédiction linéaire suivant, l'emplacement des problèmes dans le signal vocal peut être anticipé à l'aide de cette méthode[23].

(c) Perceptual Linear Prediction (PLP) : PLP est une technique utilisée pour extraire les informations essentielles de la parole, combine les bandes cruciales, la compression intensité-son et la préconcentraient à son égal. La technique est basée sur une banque de filtration auditive et présente une précision réduite aux hautes fréquences. Il combine l'analyse de prédiction linéaire avec la spectroscopie[23].

3 Conclusion

Dans ce chapitre, nous avons présenté les aspects théoriques en termes reconnaissance des émotions et en expliquant différents concepts, nous avons défini l'émotion, ensuite, les méthodes utilisées pour identifier l'émotion dont toutes les sources d'information et l'avons scellée avec quelques aspects techniques.

Dans le prochain chapitre, nous expliquerons en détail les aspects techniques de la mise en place de systèmes de reconnaissance des émotions multimodale.

Chapitre 2

Système de reconnaissance des émotions multimodales

1 Introduction

Les systèmes multimodales sont des systèmes informatiques qui utilisent plusieurs modalités pour interagir avec les utilisateurs, telles que la voix, le texte et les images. Ces systèmes sont de plus en plus importants dans le domaine de l'informatique car ils permettent aux utilisateurs de communiquer avec les machines de manière plus efficace et plus fiable.

Dans ce chapitre, nous aborderons la définition des systèmes de reconnaissance des émotions multimodale et les techniques de fusion et présenter certaines architectures utilisées dans ces systèmes,on termine par une synthèse de quelques travaux connexes sur la reconnaissance d'émotion multimodale.

2 Reconnaissance d'émotion multimodale

La reconnaissance d'émotion multimodalede est le processus de reconnaître et de comprendre les émotions humaines à travers des nombreux types d'informations illustrés dans la figure 2.1, telles que les expressions faciales, le langage corporel et les modèles de parole, conduit à une compréhension plus approfondie des émotions humaines, d'une manière ou d'une autre offrant une preuve particulière de l'état émotionnel d'une personne. Ces sources sont intégrées à l'aide d'approches d'intégration au niveau des caractéristiques ou au niveau décisionnel[24].

3 Modalité

Une modalité est assimilée à un médium, parfois appelé canal. Il existe des modes verbaux, auditif, visuel, physiologique et textuel. Plusieurs sources d'information pouvant offrir divers types d'information et points de vue sont reflétées dans les modalités. On peut catégoriser les techniques de reconnaissance émotionnelle comme unimodales ou multimodales. Le terme

"reconnaissance émotionnelle unimodale" désigne l'identification des émotions humaines en utilisant une seule modalité, comme un visage, un texte, un EEG, une voix ou une image. Le principal inconvénient de l'approche unimodale est le potentiel de faiblesses modales dans certains contextes, par exemple, l'audio est préférable à la vidéo dans les environnements à faible luminosité, mais le texte surpasse parfois l'audio lorsqu'il décrit la dimension valence[25].

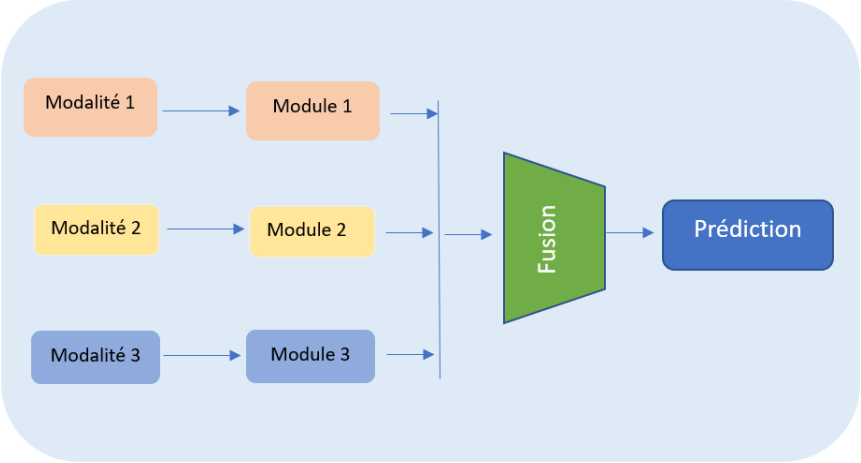

FIGURE 2.1 – Reconnaissance des émotions multimodales.

4 Technique de fusion

Les techniques de fusion multimodales sont utilisées pour combiner les informations provenant de différentes modalités (comme la parole, les expressions faciales et les textes) afin d'améliorer la précision de la reconnaissance émotionnelle. Il existe plusieurs approches de fusion multimodale, telles que la fusion de décision, la fusion de caractéristiques. Dans ce qui suit nous présentons des définitions pour chaque technique de fusion.

4.1 Fusion au niveau caractéristiques :

Concaténée les caractéristiques de chaque modalité en un seul vecteur de caractéristiques illustré dans la figure 2.2, qui est ensuite utilisé pour l'apprentissage du modèle de reconnaissance des émotions. Simple à mettre en œuvre, mais peut donner lieu à un espace de caractéristiques de grande dimension [26]. Le processus d'intégration avec cette technique se fait par la simple concaténation des caractéristiques de différentes modalités [27] ou par la combinaison selon « Weighted concatenation-based » où les poids sont appris pendant l'apprentissage du modèle pour donner à ces modalités différents degrés d'importance [28].

- **concaténation pondérée :** La concaténation pondérée est une technique utilisée pour combiner plusieurs entrées ou caractéristiques dans un réseau neuronal. Dans la concaténation pondérée, des poids sont appliqués aux caractéristiques avant de les concaténer. Cela permet au réseau d'apprendre quelles fonctionnalités sont les plus importantes[29].Pendant l'entraînement, les paramètres de poids sont ajustés de sorte que les caractéristiques avec

des poids plus élevés ont plus d'influence sur la sortie. Cela aide le réseau à apprendre quelles fonctionnalités sont les plus importantes pour la tâche.

- **concaténation simple :** La concaténation simple est une technique utilisée pour combiner plusieurs entrées ou caractéristiques dans un réseau neuronal. [30]. La principale différence avec la concaténation pondérée est que la concaténation simple n'applique aucun poids aux entrées individuelles avant de les concaténer. Tous les intrants sont traités également. Pendant l'entraînement, le réseau doit apprendre quelles entrées sont plus importantes indirectement, en ajustant les poids dans les couches suivantes.

4.2 Fusion au niveau décisionnel :

Une technique de combinaison de différents classificateurs après qu'ils ont été entraînés individuellement, et les prédictions sont combinées au stade de la prise de décision[26] illustré dans la figure 2.2. Il existe plusieurs stratégies de fusion notamment :

4.2.1 Fusion de décisions :

Une stratégie de fusion qui consiste d'abord à construire un classifieur utilisant chaque source de données séparément avant de fusionner les prédictions faites par les différents classifieurs en utilisant ces techniques [31].

(a) Majority Voting : L'ensemble choisit une classe lorsque tous les classifieurs s'entendent sur la classe spécifique ou bien prédit par au moins un plus de la moitié du nombre de classifieurs ou reçoit le plus grand nombre de votes, que la somme de ces votes dépasse ou non 50%[32].

(b) Weighted Voting : L'ensemble de classsifieurs n'a généralement pas de performances identiques, donc les classifieurs qualifiés ont plus de pouvoir pour porter la décision finale[32].Cette méthode doit déterminer les poids des classifieurs à l'aide d'un ensemble de validation. Chaque classifieur génère une décision indiquant l'étiquette de classe prévue d'une seule instance, puis ces décisions sont évaluées pour mettre à jour les poids. À la fin, combiner les produits de chaque classifieur en tenant compte de leur poids[33].

4.2.2 Fusion de score :

La décision finale est prise en utilisant le score global qui est créé après que les scores ont été combinés. Pour ce faire, un certain nombre de stratégies peuvent être utilisées. Il existe plusieurs méthodes basées sur les règles telles que la somme simple, la règle du maximum, la règle du minimum et la règle du produit [34].

4.3 Fusion hybride :

Combine les techniques de fusion au niveau décisionnel et fusion au niveau caractéristiques, où les caractéristiques sont d'abord extraites de chaque modalité, puis un modèle de reconnais-

sance des émotions séparé est formé pour chaque modalité. Les résultats de chaque modèle sont ensuite combinés au stade de la prise de décision[26] illustré dans la figure 2.2.

4.3.1 Fusion au niveau caractéristiques suivie par fusion au niveau décisionnel :

Cette fusion implique d'extraire les caractéristiques de chaque modalité et de les concaténer en un seul vecteur de caractéristiques. Le vecteur de caractéristique concaténé est ensuite utilisé pour former des modèles de reconnaissance émotionnelle distincts pour chaque modalité, et les prédictions de chaque modèle sont combinées à l'étape de la prise de décision[35].

4.3.2 Fusion au niveau caractéristiques suivie par fusion de score

Consiste à combiner les caractéristiques des différents modèles en un seul vecteur de caractéristique avant la partie décisionnelle et après que les modèles ont fait leurs prédictions indépendamment. Les scores de confiance ou probabilités de chaque modèle sont alors combinés, par exemple en prenant la moyenne[36].

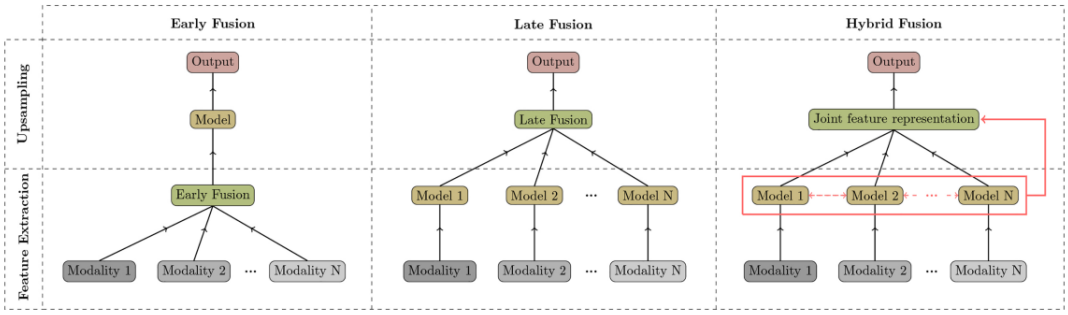

FIGURE 2.2 – Structure de late fusion,early fusion,hybrid fuison[37].

4.4 Fusion à plusieurs niveaux :

Implique l'utilisation de plusieurs niveaux de fusion, où les caractéristiques sont d'abord fusionnées au niveau bas, comme le niveau de caractéristique, puis fusionnées au niveau supérieur, comme le niveau de décision. Peut capturer les relations intermodales entre les différentes modalités et peut améliorer les performances du système de reconnaissance des émotions[26].

4.5 Fusion basée sur l'attention :

Utilise un mécanisme d'attention pour pondérer la contribution de chaque modalité à la prédiction finale. Le mécanisme d'attention peut apprendre à se concentrer sur les modalités les plus informatives et peut améliorer les performances du système de reconnaissance des émotions[38].Il existe plusieurs techniques adoptées sur attention fusion notamment :

4.5.1 Cross-modal Attention :

Une technique utilisée dans l'apprentissage automatique pour améliorer les performances des modèles qui traitent de multiples modalités d'entrée, telles que le texte, les images et l'audio. Il s'agit d'utiliser des mécanismes d'attention pour permettre au modèle de se concentrer sélectivement sur des parties spécifiques de chaque modalité afin de faciliter l'intégration de l'information entre les modalités[39].

4.5.2 L'attention modality-spécifique

Une technique utilisée pour la reconnaissance des émotions multimodales qui consiste à calculer les poids d'attention au niveau de la caractéristique ou du segment pour chaque modalité séparément. Dans cette approche, chaque modalité est traitée séparément, et un mécanisme d'attention est utilisé pour pondérer les caractéristiques ou les segments les plus informatifs de chaque modalité[40].

5 Comparaison des techniques de fusion

Après avoir expliqué les techniques de fusion, le tableau suivant présente les avantages et les inconvénients de chaque technique.

Technique	Avantage	inconvénient
Fusion au niveau caractéristiques	- Peut saisir les relations intermodales entre différentes modalités. - Simple à mettre en œuvre. - Peut améliorer les performances du système [41].	- Peut conduire à un espace de caractéristiques de grande dimension. - Peut ne pas saisir les relations intermodales entre les différentes modalités[41].
Fusion au niveau décisionnel	- efficace sur le plan informatique. - Peut saisir les relations intermodales entre différentes modalités. - Peut améliorer les performances du système [41].	- Peut ne pas saisir les relations intermodales entre les différentes modalités[41].
Fusion hybride	- Peut saisir les relations intermodales entre différentes modalités. - Peut améliorer les performances du système [42].	- Coûteux en calcul[42].
Fusion à plusieurs niveaux	- Peut saisir les relations intermodales entre différentes modalités. - Peut améliorer les performances du système [26].	- Coûteux en calcul [26].
Fusion basée sur l'attention	- Peut saisir les relations intermodales entre différentes modalités. - Peut améliorer les performances du système de reconnaissance des émotions. - Peut apprendre automatiquement à se concentrer sur les modalités les plus informatives[38].	- Coûteux en calcul. - Nécessite une grande quantité de données pour former le mécanisme d'attention[38].

Table 2.1 – Comparaison des techniques de fusion.

6 Apprentissage en profondeur (Deep Learning)

6.1 Définition

Le Deep Learning (ou apprentissage profond) est un ensemble de méthodes d'apprentissage automatique conçues sur la base de réseaux de neurones profonds, visant à mimer la profondeur des couches d'un cerveau : le cerveau humain est profond, dans le sens où chaque action est le résultat d'une longue chaîne de communications synaptiques avec de nombreuses couches de traitement. Le deep learning réunit une classe d'algorithmes d'apprentissage correspondants à ces architectures profondes. Il est souvent utilisé pour un apprentissage «de bout en bout», c'est-à-dire l'apprentissage simultané des caractéristiques utiles des données, et de la meilleure façon de les utiliser.[43]

6.2 Réseaux neurones

Les réseaux de neurones artificiels sont des algorithmes d'apprentissage automatique inspirés très librement du fonctionnement des neurones biologiques,bien que :

6.2.1 Long short-term memory

La mémoire à court terme (LSTM) est une variation d'un modèle RNN.Un modèle LSTM peut rappeler des données antérieures de séries chronologiques à long terme et dispose d'un contrôle automatique pour conserver les caractéristiques pertinentes ou éliminer les caractéristiques non pertinentes dans l'état de la cellule. Un modèle LSTM a trois portes pour contrôler les caractéristiques, c'est-à-dire, la porte d'entrée, oublier porte, et porte de sortie illustrée dans la figure 2.3. La porte d'entrée contrôle les nouvelles informations pour qu'elles circulent dans l'état de la cellule. La porte d'oubli supprime les informations antérieures sans importance de l'état de la cellule. La porte de sortie régule l'information extraite de l'état de la cellule, puis décide de l'état caché suivant. Un modèle LSTM peut automatiquement enregistrer ou supprimer la mémoire stockée en utilisant ces portes.[44]

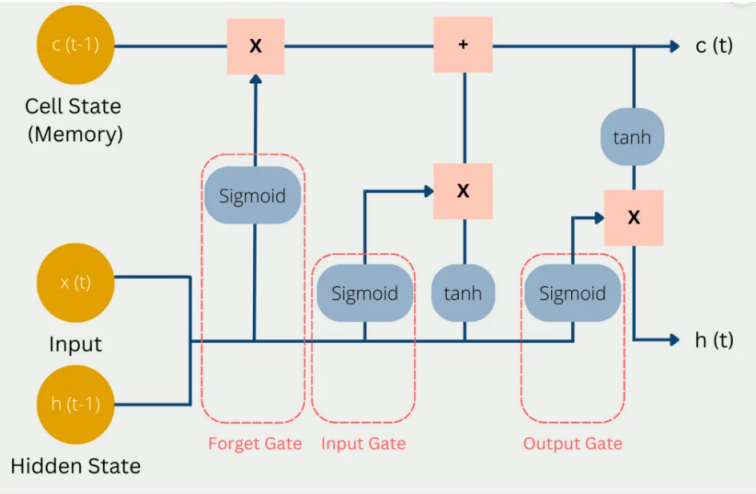

FIGURE 2.3 – Réseaux neurones LSTM [45].

6.2.2 Convolutional Neural Network

Les réseaux neuronaux convolutifs (CNN), sont un type de réseaux neuronaux à anticipation bien adaptés aux tâches liées au domaine de la vision par ordinateur. Une architecture CNN typique comprend généralement des couches alternées de convolution et de mise en commun (pooling), suivies d'une ou plusieurs couches entièrement connectées(fully connected) à la fin.En plus des différentes fonctions de mappage, différentes unités de régulation telles que la normalisation et le dropout des lots sont également incorporées pour optimiser les performances du CNN. La disposition des composants du CNN joue un rôle fondamental dans l'amélioration de meilleures performances.[46]

7 Quelques architectures multimodales

Il existe de nombreuses architectures pour les systèmes multimodales qui ont introduit différents modèles d'apprentissage profond,y compris des architectures basé sur CNN, LSTM et d'autres basé sur le mécanisme "Attention".

- **Deep Canonical Correlation Analysis (DCCA) :** Est une technique d'apprentissage des transformations non linéaires complexes de deux ensembles de données pour produire des représentations linéaires hautement associées. Pour maximiser la corrélation régularisée ou totale, les paramètres des deux transformations sont simultanément appris [47].
- **Multimodal Deep Boltzmann Machines (MDBM) :** Est un réseau d'unités binaires stochastiques couplées symétriquement. Il contient un ensemble d'unités visibles et une séquence de couches d'unités cachées. Il n'y a de connexions qu'entre les unités cachées dans les couches adjacentes et il n'y a pas de connexions non plus dans la couche visible ou dans la couche cachée. il peut être en modélisant chaque modalité de données en utilisant des couches séparées de DBM. L'idée clé est d'apprendre un modèle de densité conjointe sur l'espace d'entrées multimodales [48].
- **Multimodal Convolutional Neural Networks (MCNNs) :** Les réseaux neuronaux qui peuvent gérer les entrées provenant de sources multiples et de modalités multiples, en utilisant plusieurs couches convolutionnelles partagées pour extraire les caractéristiques de chaque modalité et en ajustant les poids du réseau, puis les combiner pour une classification conjointe[49].
- **Multimodal Long Short-Term Memory Networks (MLSTMs) :** LSTM est en mesure de modéliser les dépendances à long terme dans les données séquentielles parce qu'à chaque étape de temps peut sélectivement « se souvenir » ou « oublier » l'information passée, il peut explicitement modéliser les dépendances à long terme à la fois dans la même modalité et entre les modalités en une seule multimodal LSTM . L'idée est de partager sélectivement les poids entre les différentes modalités pendant la passe avant [50].
- **Multimodal Graph Convolutional Network :** Traitement des données et modélisation sous forme de graphique avec de nombreuses modalités. Un graphique est un ensemble de nœuds reliés par des bords, où chaque nœud représente un élément de données et

chaque bord indique une relation entre les nœuds. Chaque couche du réseau consiste en des processus de transmission de messages entre les nœuds. Lorsqu'il s'agit de données multimodales, chaque modalité peut être représentée sous forme de graphique distinct, et le MGCN est utilisé pour apprendre une représentation combinée de toutes les modalités[51].

- **Multimodal Attention-based Recursive Neural Network :** Les réseaux de neurones récurrents fondés sur le mécanisme "attention"(RNN) peuvent représenter efficacement les relations spatio-temporelles dynamiques entre les séries exogènes et les séries cibles, mais ils ne fonctionnent bien qu'en prédiction temporelle en une étape et en prédiction temporelle à court terme, l'idée de mécanisme "attention" et RNN basé sur le comprendre les corrélations spatiales dans la prédiction de séries chronologiques, et peut apprendre la stratégie de distribution de poids des données brutes [52].
- **Multimodal Deep Learning Framework for Cross-Modal :** Cross-modal utilise un certain type de données comme requête pour obtenir du matériel pertinent d'un type différent. Ce modèle est utilisé pour tester si des représentations partagées peuvent être utilisées pour apprendre des informations communes à de nombreuses autres modalités. La majorité des techniques de recherche intermodale actuellement utilisées ont été conçues pour apprendre conjointement un sous-espace commun, exigeant que les données de toutes les modalités soient incluses pendant tout le processus de formation. Cela permet une récupération flexible entre les différentes modalités [53].
- **Multimodal Deep Autoencoder with Attention (MDAE-Att) :** L'erreur de reconstruction est utilisée par Autoencoder, qui se compose d'un encodeur et d'un décodeur, pour trouver les anomalies de plusieurs modalités. Les données d'entrée de ces modalités sont compressées par l'encodeur et stockées dans un vecteur caché, qui est utilisé par le décodeur pour reconstruire les données d'entrée. La méthode d'attention de l'autocodeur profond multimodale permet au modèle de se concentrer sélectivement sur les données pertinentes de chaque modalité tout au long des phases d'encodage et de décodage. Ceci est accompli en donnant un poids à chaque élément d'entrée[54].
- **Multimodal Deep Neural Networks with Attention (MDNNA) :** La combinaison d'un mécanisme "attention" avec une méthode de réseau neuronal profond comme CNN pour se concentrer sur des caractéristiques plus importantes et interagir l'information entre les branches. Mettre l'accent sur des caractéristiques plus représentatives qui sont pertinentes tout en limitant l'information non essentielle. D'autre part, pour réaliser l'interaction d'information entre les branches d'extraction des caractéristiques en attribuant un poids à chaque élément de cette modalité, l'objectif principal est de sélectionner l'information qui est plus critique à partir de nombreuses informations [55].

7.1 Quelques approches de reconnaissance multimodale

Plusieurs approches ont été proposées pour la reconnaissance des émotions multimodale sur différentes sources comme indiqué précédemment en utilisant divers techniques de fusion tels qu'au niveau de caractéristique , niveau décisionnel .Dans ce qui suit, nous décrivons quelques

travaux récemment publiés dans le domaine de la reconnaissance des émotions multimodale, ces travaux sont ordonnés selon type de source (parole-expression faciale, parole-texte, parole-texte-expression facial, EEG-EOG).On peut citer :

Wan Ding et ses collaborateurs ont proposé un modèle Deep Learning de reconnaissance d'émotions qui utilise deux différentes modalités : la voix (audio) et les expressions faciales (vidéo) afin de faire la classification en une de ces 6 émotions (Angy, sad, happy, surprise, fear et disgust) sur EmotiW dataset. Pour le visage, cette méthode est composée en trois étapes extraction les caractéristiques du visage en utilisant DCNN Deep Convolutional neural network, ensuite l'extraction des caractéristiques de la vidéo à l'aide la méthode " image set modeling " pour trouver les vecteurs propres , la matrice de covariance et la distribution gaussienne multidimensionnelle et la dernière étape la classification utilisant PLS partial least-square. Pour l'extraction des caractéristiques auditives, ils ont utilisé openSMILE à travers la version Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) pour couvrir les informations de qualité spectrale,prosodique et vocale et ils utilisent long short terme memory (LSTM) pour la classification de la modalité audio. À la fin ils ont décidé de fusionner les prédictions score de chaque classification par la méthode « Score-level fusion» pour obtenir à résultat final. La validation des résultats avec le dataset EmotiW montre que le système donne de bons résultats pour les classes comme la peur et la tristesse, mais inversement pour les classes comme le dégoût et la surprise. Le taux de reconnaissance final obtenue de cette expérience est de 53.9%. Il est meilleur que la méthode baseline qui a atteint un taux de 40,47% [56].

Jian Huang et al. [57] ont proposé une nouvelle approche basée sur transformer learning pour établir la reconnaissance des émotions utilisent AVEC 2017 dataset. Ils utilisent pour cela, deux modalités : audio, Visuel (expressions faciales). Pour l'extraction des caractéristiques auditives, ils ont utilisé la version Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) . Les caractéristiques géométriques sont considérées comme des caractéristiques des expressions faciales, y compris les landmarks. Cette équipe décide de séparer ces modalités en deux différent modèle, chaque modèle contient plusieurs couches. Les sorties de ces deux modèles sont combinées avec la technique « Model level fusion » et ils ajoutent de plus une couche de LSTM pour augmenter la performance du résultat. Ils ont atteint une précision de 62,9% pour arousal et 59,3% pour la valence avec la méthode de fusion « Model level fusion ».

Shizchen et ses collaborateurs ont proposé nouvelle stratégie multimodale de fusion appelée "attention fusion". Ils utilisent pour cela deux modalités : audio, Visuel (expressions faciales) depuis AVEC2015 dataset. Les caractéristiques acoustiques ont été extraites par OpenSmile sur 76 dimensions. Pour les caractéristiques visuelles, deux ensembles sont extraits de l'expression faciale : caractéristiques basées sur l'apparence (avec Local Gabor Binary Patterns) puis compressée avec ACP à 84 dimensions et caractéristiques géométriques. Les deux vecteurs de caractéristiques sont fusionnés par condition attention fusion modèle . Cette approche a donné le deuxième meilleur résultat après la méthode early fusion pour la performance de l'arousal 85%. Pour la valence ils ont atteint 73% [28] .

Yagya Raj et ses collaborateurs ont proposé un système pour la classification des émotions (Excitant, Peur, Neutre, Détente, Triste et Tension) par adopter le transfert learning de certains réseaux neuronaux profonds bien connus pour la classification de la musique et de la vidéo. Ils ont utilisé pour cela deux modalités : audio, visuel (expression faciale) avec RECOLA dataset . Pour l'extraction des caractéristiques et la classification de ces deux modalités ils ont appliqué un 3DCNN (three-dimensional convolutional neural networks) pour chaque modalité. Les deux résultats de classification obtenus ont été combinés par la méthode decision-level feature et la décision finale est prédite par la couche softmax. Le modèle prédictif par intégration de toutes les structures multimodales atteint un taux de reconaissance (accuracy) de 88,56% et 98,7% pour la mesure de performance Area Under the Curve (AUC)[2].

Shiqing Zhang et al.[58] ont proposé une approche pour la reconnaissance d'émotion multimodale, les auteurs utilisent pour cela, deux modalités : expressions faciales et audio afin de classifier l'émotion en : anger, disgust, fear, joy, sadness, et surprise avec RML audio-visual dataset. Ils ont utilisé le DCNN pré-entrainé de la classification d'image à grande échelle pour initialiser le DCNN . Pour l'extraction des caractéristiques des modalités audio et visual individuellement, ils utilisent AlexNet en raison de ses performances prometteuses et de sa capacité de généralisation dans diverses tâches de vision. À la fin, la fusion de ces modalités, se faire sur la couche « fusion network » qui contient deux couches fully connected et une couche softmax par concaténer les deux couches « fully connected » de chaque modalité et ils ont utilisé SVM pour compléter la classification. La performance de cette approche atteint une précision de 74.32%. De plus les émotions « joie » et « tristesse » sont plus difficiles à identifier que « colère » et « surprise ».

Pen Shixin et ses collaborateurs ont proposé une architecture basée sur l'autoencoder pour la reconnaissance des émotions. Ils utilisent deux modalités : texte et audio, afin de faire la classification d'émotions (Angy, sad, happy,neutral). Ils commencent par l'extraction des caractéristiques des deux modalités. Pour l'audio, ils ont proposé une structure appelée Temporal Global Feature Extractor (TGFE) pour extraire les caractéristiques haut niveau à partir des caractéristiques LOG-MEL de la voix. Pour le texte, ils sont appliqués la technique BERT (Bidirectional Encoder Representations from Transformers) afin de transformer ces mots en une séquence de vecteurs. Dans les deux cas et en raison de la nature séquentielle de ces données vocales et textuelles, ils ont appliqué le Bidirectional LSTM. Dans les deux modalités, la couche attention est utilisée pour améliorer la précision de la reconnaissance des émotions. Les deux vecteurs obtenus sont fusionnés par un autoencoder au niveau des caractéristiques hauts niveaux pour apprendre la représentation partagée des caractéristiques émotionnelles vocales de haut niveau des modalités acoustiques et textuelles. Cette expérience a été évaluée avec trois datasets (IEMOCAP, CMU_MOSI, et MELD). Ils ont obtenu avec IEMOCAP dataset une précision de 74,8% à partir de fusion de modalité (audio+ texte). Pour les datasets MELD et CMU-MOSI ils ont obtenu 63.64% et 79.85% respectivement[59].

Davamanyu et al. [60] ont proposé un système pour la reconnaissance des émotions qui utilisent deux modalités : texte et audio avec « IEMOCAP » dataset. Pour extraire des caractéristiques du texte,ils ont appliqué une technique appelée Fast text afin de transformer ces mots en une séquence de vecteurs pour nourrir à CNN. De l'autre côté, ils ont utilisé OpenSmile pour extraction ces caractéristique auditive. Les deux vecteurs sont ensuite combinés par Multimodal attention fusion qui est basé sur les mécanismes de self-attention effectuant une addition pondérée. Les performances de l'approche sont calculées avec trois indicateurs : Précision 71,4%, F-1 mesure 71,3% et UAR (Unweighted Recall) 72,1%. D'après les auteurs, les mécanismes self-attention surpassent les autres méthodes de fusion tels que features level fusion.Cette expérience montre l'importance de la modalité audio qui fournit des informations supplémentaires.

Suraj Tripathi et ses collaborateurs ont proposé une approche pour la reconnaissance des émotions. Ils utilisent pour cela, deux modalités : audio, texte afin de classifier les émotions en : neutral, happiness, sadness, anger, surprise, fear, disgust frustration, excited et autres, avec IEMOCAP dataset . Pour l'extraction des caractéristiques textuelles, ils ont adopté la technique de plongement lexical appelez Word2Vec. Pour l'extraction des caractéristiques auditives ils ont appliqué un modèle CNN à partir de deux différentes entrées : Spectrogramme et Mel-frequency Cepstral Coefficients (MFCC). Dans cette approche, ils ont proposé plusieurs techniques de combination des modalités afin de les comparer entre eux. La première fusion les deux vecteurs normalisés textuel et auditif (MFCC) de chaque couche fully connected dans un seul vecteur et ensuite utilise la couche softmax pour compléter la classification. La combinaison du texte-audio donne de meilleurs résultats avec de précision égale à 76.1%, 75,1% pour la combinaison du Texte Spectrogramme et 73,6% pour le spectrogramme et MFCC [61].

Soujanya et al. [62] ont proposé une nouvelle approche pour la reconnaissance des émotions et l'analyse des sentiments en utilisant trois différentes modalités (texte, video et audio) à partir de dataset IEMOCAP, afin de faire la classification d'émotions (Angy, sad, happy,neutral). Ils ont utilisé pour cela CNN pour extraction des caractéristiques à partir d'une séquence de vidéo, ensuite la classification avec RNN. La vectorisation de l'audio se fait à l'aide de OPENSmile tools. Pour extraction des caractéristiques d'un texte, ils ont proposé un CNN basique (convolution, max pooling ,fully connected) et le texte est traité par une technique appelez word2vec. Tous ces vecteurs obtenus de ces modalités sont fusionnés par la méthode features level fusion sous forme de fonctions de concaténation des trois vecteurs afin d'utiliser ce dernier comme entrée du modèle Multi Karnel Learning pour effectuer la classification. Ils ont atteint un taux de reconnaissance de de 96,55% avec IEMOCAP dataset . Le classifieur textuel reconnaît bien les cas de colère, de bonheur et de neutralité, les cas de colère et de tristesse sont très difficiles à distinguer. Dans le cas de la modalité audio, ils observent une meilleure précision que la modalité textuelle pour les classes tristes et neutres mais pas pour les classes heureuses et en colère. La modalité visuelle donne de bonne précision par rapport aux deux autres modalités.

Asokan et al.[63] ont proposé une approche pour la reconnaissance d'émotions (happy,

sad, neutral,anger ,excited et frustration) en utilisant trois modalités : audio, texte, et vidéo avec IEMOCAP dataset . Ils utilisent word2vec puis CNN pour extraire les caractéristiques textuelles. L'extraction des caractéristiques pour l'entrée visuelle se fait à l'aide d'un 3D-CNN qui est capable d'apprendre les caractéristiques pour chaque image de la vidéo. L'extraction des caractéristiques audio se fait à l'aide du logiciel openSMILE qui est capable d'extraire plusieurs caractéristiques de bas niveau telles que la hauteur, l'intensité et MFCC. La fusion des modalités,se faire de manière hiérarchique, ils adaptent pour cela avec un Contextual-LSTM. A la fin ils appliquent la méthode Concept Activation Vectors (TCAV) pour la prédiction finale. Ce travail a été évalué sur trois concepts de précision : Variations in Physiognomy (VP), Voice Pitch (PT) et Utterance Polarity (UP). Pour la métrique VP par exemple, ils voient que les scores les plus élevés sont observés pour les classes heureuses et excitées, tandis que les scores pour les classes tristes, en colère et de frustration sont relativement plus faibles avec 78% de précision.

Bahar Hatipoglu et ses collaborateurs ont proposé une approche de reconnaissance d'émotions avec le signal EEG et EOG en utilisant DEAP dataset . Ils ont converti ces signaux aux images de deux dimensions de type AAG (Angle-Amplitude Graph) à l'aide de la technique AAT (angle amplitude transformations) qui permet de détecter les points maximal et minimal sur un signal du canal et ils ont ensuite utilisé l'algorithme SIFT pour éliminer les caractéristiques inutiles. Ensuite les auteurs ont choisi deux méthodes de fusion au niveau des caractéristiques, la première permet le calcul du vecteur final à l'aide des vecteurs SIFT calculés à partir des signaux EOG et la deuxième permet le calcul du vecteur final à l'aide des vecteurs SIFT calculés à partir des signaux EOG . À la fin, la classification de ces vecteurs est obtenue avec SVM. Les résultats de la première méthode de fusion sont à 91,53% pour arousal et à 90,31% pour la valence. Pour la deuxième méthode de fusion les résultats sont 89.71% pour l'arousal et 88.59% pour la valence[64].

Young et ses collaborateurs ont proposé une approche basée sur CNN pour la reconnaissance des émotions. Pour cela, ils ont effectué des expériences de classification binaire sur les dimensions de la valence et de l'activation en utilisant deux ensembles de données (Manhob-hci/Deap). Ils ont choisi les signaux EEG et PPS comme modalités. L'extraction des caractéristiques de l'EEG et du PPS se fait manuellement afin d'entraîner le modèle HFCNN (Hierarchical Fusion Convolutional Neural Network). Chacune des couches de convolution et de pooling dans ce modèle HFCNN produit des vecteurs qui doivent être combinés au niveau de caractéristiques selon leurs poids pour obtenir un vecteur global de ces vecteurs. Ensuite, une fusion de trois vecteurs (vecteur global, vecteur EEG, vecteur PPS) afin d'effectuer la classification avec Random Forest. Les auteurs ont obtenu un taux de reconnaissance de 84,71% pour le jeu de données DEAP et de 89% pour le jeu de données MAHNOB-HCI [65].

Yongrui Huang et Jianhao Yang utilisent deux modalités : le signal EEG et les expressions faciales afin de classifier quatre types d'émotions (happiness, neutral, sadness ,fear). Le vecteur des caractéristiques d'expressions faciales a été classé par un réseau de neurones à propagation avant. Pour le signal EEG, l'extraction des caractéristiques est basée sur le PSD (Poststroke

Depression) et SVM pour la classification. À la fin, les résultats de classification sont fusionnés par deux méthodes de fusion. Dans la première, ils ont appliqué une technique de fusion au niveau de décision en calcule la somme des scores de classifieur d'expression faciales et des scores de classifieur EEG pour chacun des quatre états émotionnels et ensuite trouver le maximum des quatre valeurs additionnées. Dans la deuxième méthode, ils ont appliqué la stratégie decision marking fondée sur les règles de production pour le résultat final. Le taux de reconnaissance de cette technique de fusion est de 81,25% sur les données online et 82,75% sur les données offlines [66] .

Wei Liu et al.[67] ont proposé une approche de reconnaissance d'émotion à partir du signal EEG et le mouvement des yeux en proposant deux différents modèles : Deep AutoEncode (DAE) et Bimodal DAE (BDAE) avec l'utilisation deux datasets SEED et DEEP. Les caractéristiques de densité spectrale de puissance (PSD) et l'entropie différentielle (DE) ont été extraites des données EEG et les mêmes pour celles des mouvements des yeux. Les auteurs décident de combiner des deux modalités afin d'améliorer la précision de la reconnaissance des émotions, en utilisant le BDAE pour extraire une représentation conjointe des données de l'EEG et des mouvements des yeux. S'il n'y a qu'une seule modalité, appliquez un encodeur unimodal (DAE) pour extraire les représentations partagées. Enfin, la classification est faite avec le SVM. L'approche atteint des taux de reconnaissance de 91,01% et 83,25% pour les datasets SEED et DEAP.

réf	Datasets	Sources (technique d'extraction)	Technique de classification	Etiquette d'émotion	Technique de fusion	Stratégie de validation	Performance
[56]	EmotiW	Parole (openSMILE) Expressions faciales (DCNN)	LSTM / partial least squares	Angry -disgust-fear happy-sad-surpise-neutral	« Score-level fusion»	/	53.9%
[57]	AVEC 2017	Parole (openSMILE) Expressions faciales (manuelle)	Transfermer LSTM	Valence -arousal (dimensionnelle)	« Model level fusion »	/	Arousal(65,4%) Valance (70,8%)
[28]	AVEC2015	Parole (openSMILE) Expressions faciales (Local Gabor Binary Patterns)	LSTM	Valence -arousal (dimensionnelle)	« condition attention fusion model »	Out cross validation	Arousal(85%) Valance (73%)
[2]	RECOLA	Parole (3DCNN) Expressions facials(3DCNN)	DCNN	Excitant, Peur, Neutre, Détente, Triste et Tension	« Decision-level feature »	cross-validation	88,56%
[58]	RML audio-visual database	Parole (DCNN) Expressions facials(DCNN)	SVM	anger,disgust, fear, joy, sadness, et surprise	« fusion network »	/	74,32%
[59]	IEMOCAP,	texte (BERT) parole (TGFE)	Bidirectional LSTM autoencoder	Angy, sad, happy,neutral	niveau des caractéristiques	/	IEMOCAP 74,8% MELD 63.64% CMU_MOSI 79.85%
[60]	IEMOCAP	texte (CNN) parole (openSMILE)	CNN Self-Attention	Happy,Sad,Neutral,Angry	« Multimodal attention fusion»	Hold-out validation	71,4%
[61]	IEMOCAP	texte (CNN) parole (CNN))	CNN	neutral, happiness, sadness, anger, surprise, fear, disgust frustration, excited	« features level fusion »	K-folds cross validation	76.1%
[62]	IEMOCAP	Texte(CNN) Parole (openSMILE) Expression facial (CNN)	RNN - Multi Karnel Learning	Angy, sad, happy,neutral	« features level fusion »	10-fold cross-validation	96,55%
[63]	IEMOCAP	Texte(CNN) Parole (openSMILE) Expression facial (3D-CNN)	BI-LSTM	happy, sad,neutral,anger ,excited et frustration	« Contextual-LSTM »	/	78% (métrique VP)
[64]	DEAP	EEG (SIFT) EOG (SIFT)	SVM	Valence -arousal (dimensionnelle)	feature level fusion	11-folds cross validation	Arousal(91,53%) Valance (90,30%)
[65]	Manhobhci Deap	EEG (manuelle) PPS(manuelle)	HFCNN	Valence -arousal (dimensionnelle)	feature level fusion	10-folds cross validation	DEAP (84,71%) Manhob(89,00%)
[66]	online offline	EEG (manuelle) expression facial (manuelle)	feedforward NN SVM	happiness, neutral, sadness ,fear	décision level fusion	hold-out validation	81.25%(online) 82.75% (offline)
[67]	SEED DEEP	EEG (PSD) yeux mouvement(PSD)	SVM	Valence -arousal (dimensionnelle)	BDAE	/	SEED 91,01% DEEP 83,25%

TABLE 2.2 – Comparaison qualitative des travaux reliés.

8 Conclusion

Dans ce chapitre,nous avons présenté des définitions sur le terme modalité et le système de reconnaissance multimodale,les types de fusion et les stratégies de fusion des systèmes sur lesquels nous nous concentrons entièrement pour créer notre système et on a présenté les différentes architectures multimodale. Pour conclure,nous avons présenté et analyser quelques travaux les plus récents sur la reconnaissance d'émotion multimodale. L'étude des projets de recherche susmentionnés nous a orientés vers une nouvelle approche multimodale de la reconnaissance des émotions .

Dans le chapitre suivant, nous expliquons la conception du système proposé.

Deuxième partie

Conception et Implémentation

Chapitre 3

Conception et modélisation de l'approche proposée

1 Introduction

L'objectif de ce projet est de développer un modèle pour la reconnaissance des émotions multimodale, où étaient les sources de parole et de texte adoptées dans le processus de reconnaissance des émotions.

Ce chapitre est consacré à la présentation détaillée de la conception du modèle de reconnaissance d'émotion multimodale proposé. Nous le commençons par la présentation de l'architecture générale du système proposé avec ces différentes phases. Puis on passe à une explication détaillée des différentes parties du système.

2 Système multimodal de reconnaissance des émotions

Nous avons proposé trois différentes approches pour réaliser un système multimodal de reconnaissance des émotions . Dans la première approche nous avons utilisé la technique de fusion "fusion network" illustré dans la figure 3.1, la deuxième nous utilisons la technique de fusion au niveau caractéristique illustré dans la figure 3.8 et dans la dernière approche nous avons utilisé la technique "fusion network" avec un mécanisme d'attention illustré dans la figure 3.9.

3 L'approche fusion network

Le système proposé est basé principalement quand un humain interagit avec une machine pour définir les émotions. L'utilisateur doit entrer deux sources afin de reconnaître l'émotion, qui est représentée comme parole (voix) et texte.La reconnaissance de son état émotionnel est faite en utilisant deux modules basés sur le deep learning en utilisant un CNN et LSTM pour ces modules,voir la figure 3.1 .

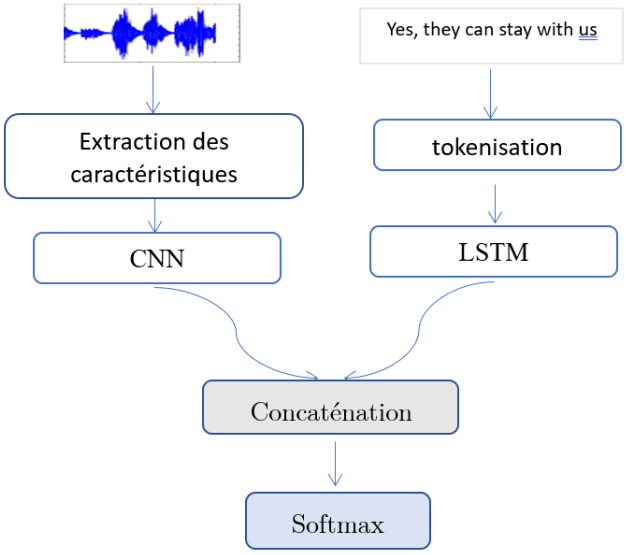

FIGURE 3.1 – conception globale de fusion network.

3.1 Module de reconnaissance des expressions vocales

3.1.1 préparation des données vocales

Ce module consiste à reconnaître d'émotions en se basant sur la voix de l'utilisateur. Ce module fonctionne en traitant le signal vocal capté par un microphone qui va s'exprimer sous forme d'enregistrement et en le convertissant en représentation numérique pouvant être analysée .Les enregistrements vocaux d'entrée doivent d'abord être transformés en un ensemble de vecteurs de paramètres.

3.1.2 L'architecture du modèle de l'audio

Nous avons utilisé "Convolutional Neural Network" pour l'architecture du modèle de l'audio illustré dans la figure 3.2 qu'est un type de réseau neuronal d'apprentissage profond.

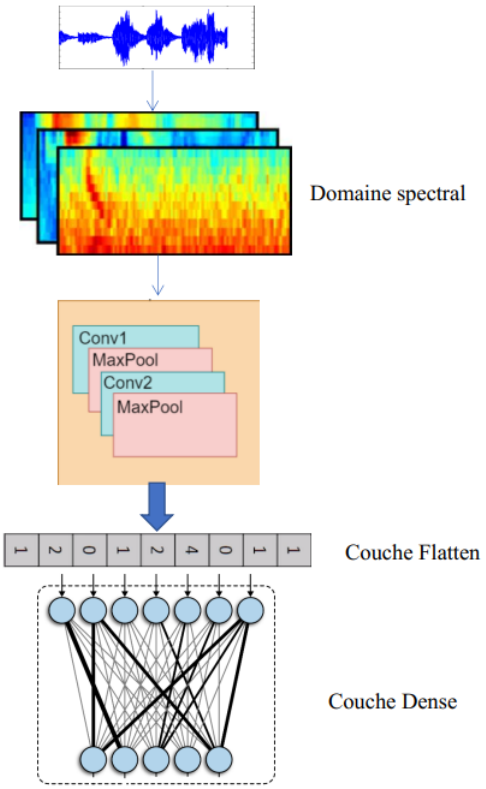

FIGURE 3.2 – L'architecture du module audio basé sur CNN.

- **Couche d'entrée**
  La couche d'entrée de CNN contiennent les valeurs du domaine spectral extrait de l'audio , ces données sont représentées par une matrice bidimensionnelle (ligne, colonne) de taille 100 x 34.
- **Couche convolution**
  On a utilisé deux couches de convolution 2D pour analyser les caractéristiques de l'audio comme déjà mentionné. Tout d'abord, une partie de la matrice (100 x 34) est connectée à la couche de convolution pour effectuer une opération de convolution et calculer le produit de point entre le champ réceptif et le filtre. Le résultat de l'opération est un entier unique du volume de sortie. Puis le filtre glisse sur le champ réceptif suivant de même matrice d'entrée et effectue la même opération à nouveau ,figure 3.3. Ce processus est répété jusqu'à la fin de la matrice. Le résultat sera l'entrée pour la couche suivante.

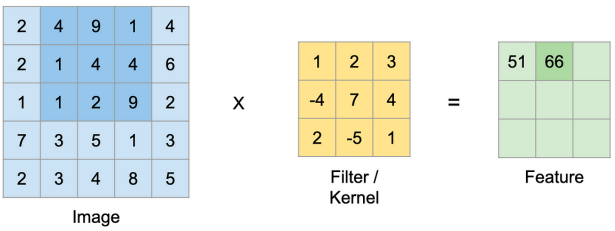

FIGURE 3.3 – Exemple de couche convolution.

- **Couche Pooling**
  Chaque couche de convolution dans ce modèle est suivie par une couche de Pooling pour réduire la taille des cartes de caractéristiques (feature maps) tout en préservant les informations les plus importantes , figure 3.4.Cette couche prend une carte de caractéristiques (feature map) de taille N x N et divise la carte de caractéristiques en régions non-recouvrantes de taille du filtre de pooling. Pour chaque région, la couche MaxPooling ne conserve que la valeur maximale et ignore les autres valeurs.

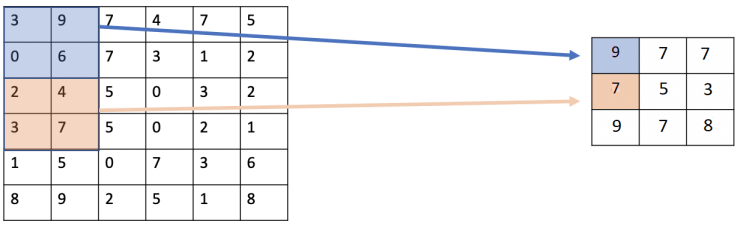

FIGURE 3.4 – Exemple la couche max-pooling.

- **Couches Flatten**
  Chaque couche de convolution produit une sortie multi-dimensionnelle. La couche Flatten prend cette sortie et se transforme vers un espace unidimensionnel en empilant toutes les caractéristiques ensembles. Le résultat obtenu peut ensuite être introduit dans une couche entièrement connectée, qui peut effectuer des tâches de classification.
- **Couche Dense**
  Dans une couche dense, chaque neurone calcule une somme pondérée des entrées, ajoute un terme de biais et applique une fonction d'activation au résultat. Les poids et les biais sont les paramètres apprenable de la couche dense, et ils sont mis à jour au cours du processus de formation en utilisant la rétropropagation.

3.2 Module de reconnaissance des expressions textuelles

Nous avons utilisé "long short term memory " pour l'architecture du modèle de texte qu'est un type de réseau neuronal d'apprentissage profond et nous avons adopté la procédure tokenisation et plongement lexical pour modéliser les données textuelles.

3.2.1 Préparation des données textuelles

Ce module se concentre sur la compréhension et l'analyse des émotions humaines par le langage écrit de l'utilisateur, Nous traitons le texte au moyen des procédures de tokenisation et de techniques de plongement lexical où nous utilisons un modèle préformé appelez GLOVE pour obtenir des représentations vectorielles des mots.

- **Tokenization :** La tokénisation est la représentation d'une seule notion en utilisant un seul mot ou une seule phrase. Le processus de tokenisation d'un texte implique de le diviser en

jetons qui sont séparés par des espaces. Par exemple, la phrase "je suis heureux " résultats dans la liste des jetons "je, suis, heureux " puisqu'un jeton peut inclure un nombre spécifique d'expressions, la pertinence technique de séparer une chaîne de caractères en espaces n'est pas pertinente[68].

3.2.2 Plongement lexical(GLOVE)

L'idée principale derrière GloVe est que le rapport des probabilités de co-occurrence de deux mots prédit à quel point leurs significations sont similaires. Sur cette base, GloVe construit une matrice de co-occurrence mot-mot à partir d'un corpus et apprend ensuite les représentations vectorielles (Plongement lexical).

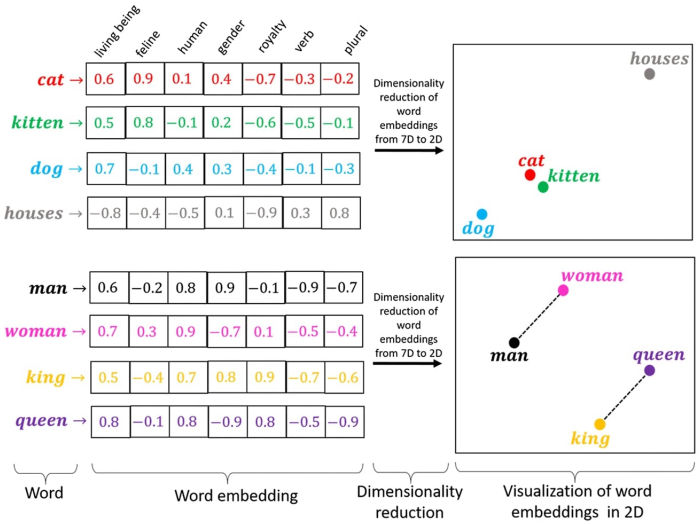

FIGURE 3.5 – Plongement lexical (GLOVE).

3.2.3 L'architecture du modèle de texte

LSTM est utilisé dans le modèle lexical (texte) qui est une forme de réseau de neurones récurrents (RNN) pour faciliter la modélisation de séquences de données dans lesquelles des dépendances à long terme sont présentes . Nous avons converti chaque texte en une séquence de valeurs utilisant les procédures de tokenisation et déterminé la dimension de ces textes à un vecteur où ce dernier serait comme une entrée pour la couche plongement lexical, cette couche utilise les 2196018 mots de GLOVE. Ces vecteurs sont introduits dans deux couches de réseaux neuronaux longs short term memory avec des cellules de mémoire à court terme, chacune ayant 256 cellules. Ceci suit une couche dense avec 256 unités,voir la figure 3.6.

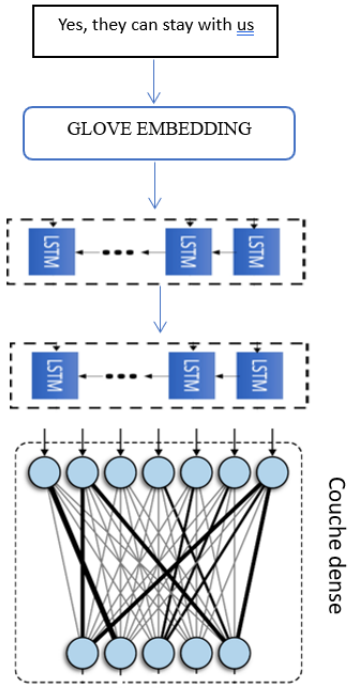

FIGURE 3.6 – Architecture de module texte basé sur LSTM.

3.3 L'architecture du modèle Multimodale

Ce modèle utilise les résultats des deux modalités, audio et texte. Les vecteurs sont concaténés on appliquant une fusion appelez « fusion network ».

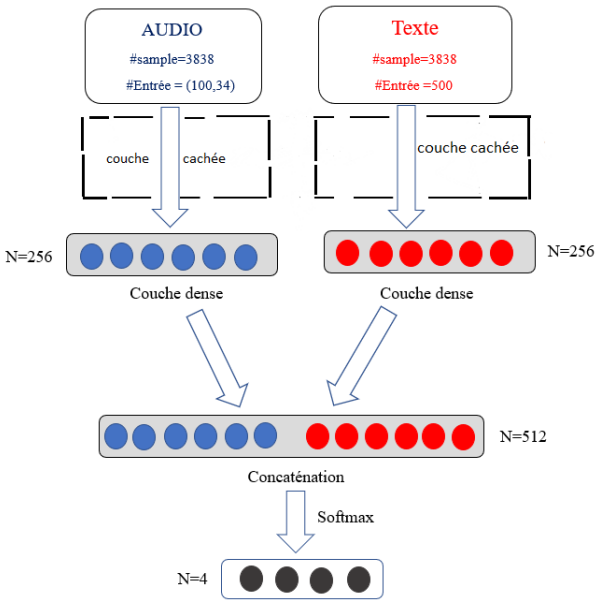

FIGURE 3.7 – modèle multimodale "fusion network".

3.3.1 Fusion Network

L'objectif de ce type de fusion est de tirer parti des informations complémentaires fournies par de multiples modalités afin d'améliorer les performances de modèle, chaque modalité dans notre approche est traitée par un ensemble distinct de couches de réseau neuronal comme déjà mentionné, qui extraient des caractéristiques de haut niveau qui capturent les caractéristiques uniques de chaque modalité. Ces vecteurs de caractéristiques spécifiques de taille 256d à la modalité sont ensuite combinés dans une couche de fusion , qui apprend à pondérer chaque modalité en fonction de sa pertinence pour la tâche à accomplir. A la fin on a introduit le vecteur du fusion de taille 512 dans deux couches dense avec 256 unités, et passé la sortie finale dans une couche SoftMax pour la classification de quatre émotions.

4 L'approche fusion au niveau caractéristique

Nous avons opté la technique de fusion au niveau caractéristique dans cette architecture illustrée dans la figure 3.8 pour intégrer les modalités audio et texte en un seul vecteur, nous avons utilisé "convolution neural network" dans le processus de l'apprentissage et pour classer les quatre émotions, nous avons utilisé la couche "SOFTMAX".

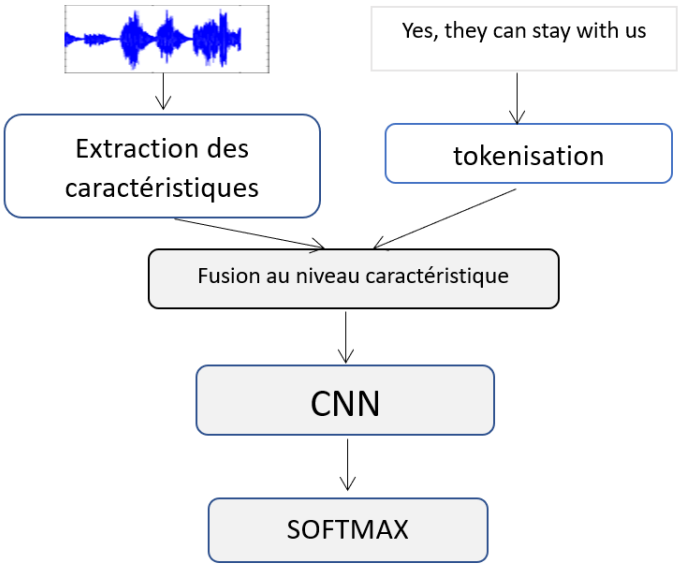

FIGURE 3.8 – Fusion au niveau caractéristique.

5 L'approche fusion network avec attention

Dans cette approche, nous avons modifié le modèle principal en ajoutant un mécanisme de "attention" dans chaque module (vocal et texte) voir la figure 3.9, qui permet aux réseaux neuronaux de se concentrer de manière adaptative sur les parties les plus pertinentes de l'entrée pour produire chaque sortie. Ce mécanisme aide les modèles à traiter de longues séquences .

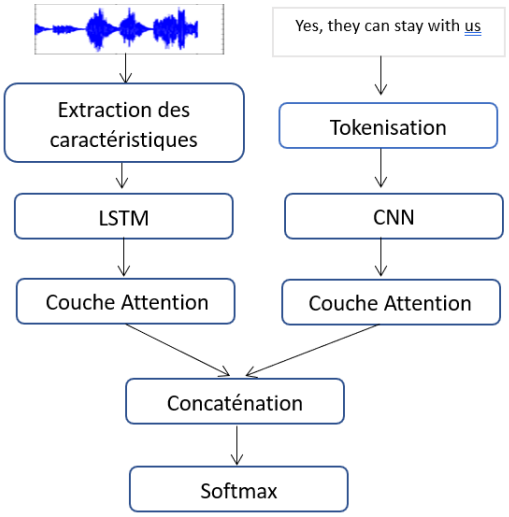

FIGURE 3.9 – Fusion avec mécanisme "attention".

6 Conclusion

Dans ce chapitre, nous avons présenté la conception avec les détails de modélisation, nous avons expliqué chaque module en expliquant les couches utilisées dans ces modules et la nature de l'ensembles des données trouvées dans chaque partie. Maintenant on peut passer à la phase d'implémentation pour réaliser cette conception dans le chapitre suivant.

Chapitre 4

Mise en oeuvre et résultat de l'approche proposée

1 Introduction

Après avoir terminé la phase de conception et formalisation de notre approche, nous passons à la phase de mise en œuvre, qui constitue la dernière partie de ce mémoire et qui tente de mettre en œuvre et de valider notre approche.

Notre projet consiste à réaliser un modèle multimodale basé sur le deep learning pour la reconnaissance d'émotion de l'utilisateur en utilisant deux sources textuelles et vocales , dans ce chapitre nous allons présenter les étapes d'implémentation et les environnements matériel et logiciel abordés pour la réalisation de ce système,et enfin on termine par une présentation des résultats expérimentaux des tests.

2 Présentation des outils de développement

Nous utilisons une collection d'outils pour nous aider à concevoir notre approche de manière appropriée pendant les étapes de développement et de mise en œuvre de notre système. Dans les sections ci-dessous, nous allons examiner ces outils en détail.

2.1 Environnent de travail :

- **Google collaboratory :** Google a créé cet environnement, souvent connu comme "Colab". Quiconque utilise Colab peut créer et exécuter n'importe quel code Python à l'aide d'un navigateur Web. C'est un cadre particulièrement adapté à l'enseignement, l'analyse de données et l'apprentissage automatique. Colab est un service de notebook Jupyter hébergé qui ne nécessite aucune configuration et offre un accès gratuit aux ressources informatiques, y compris les GPU, en termes plus techniques. Selon certaines définitions, COLAB est une plateforme de cloud computing basée sur Jupyter Notebooks qui permet la création d'applications Deep Learning basées sur Python. Il fournit un processeur GPU gratuit, 12 Go de RAM, et plus de 100 Go de stockage. Tout ce que nous devons faire pour utiliser

ce service est d'avoir un compte Google[69].

2.2 Langage de programmation

- **Python :**Pour le langage de développement, nous avons choisi Python, un langage de programmation interprété, multi-paradigme et multi-plateforme, c'est aussi un langage plus courant et populaire pour l'apprentissage automatique et l'intelligence artificielle grâce à sa flexibilité et aussi parce qu'il y a un nombre important de bibliothèques de logiciels libres disponibles. Active ses bibliothèques utilisées dans notre projet : pandas, Numpy TensorFlow,sklearn, google-Colab, matplotlib. . , etc. Les cadres TensorFlow et Keras ont été choisis pour la mise en œuvre des méthodes d'apprentissage profond proposées[70].

2.3 Bibliothèques

- **Numpy :** Numerical Python est une bibliothèque python pour les structures de données, l'algèbre linéaire et la manipulation des matrices. Sa syntaxe est similaire à celle de matlab, tout comme la façon dont elle gère les structures de données et les matrices. Elle contient les structures de données, les méthodes et les bibliothèques nécessaires à la plupart des applications scientifiques basées sur python et nécessitant des données numériques [71].
- **Pandas :** Pandas est une bibliothèque python qui fournit des structures de données rapides, polyvalentes et expressives pour traiter des données "relationnelles" ou "étiquetées". Son objectif est de servir comme base pour entreprendre des analyses de données réalistes et concrètes en python. En outre, il aspire à devenir l'outil d'analyse et de manipulation de données open source le plus puissant et le plus polyvalent accessible dans n'importe quel langage [72].
- **Matplotlib :** Est la bibliothèque la plus célèbre pour la visualisation de données avec python.Il permet la création de littéralement chaque type de graphique avec un grand niveau de personnalisation , qui est étroitement intégrée à numpy et pandas, est devenue un composant important de la pile python pour la science des données[73].
- **Scikit-learn :** Scikit-learn est une bibliothèque Python libre qui contient une variété d'algorithmes pour la classification, la régression et le regroupement, y compris les machines avec des vecteurs de support et des forêts aléatoires. Il est conçu pour fonctionner en conjonction avec les bibliothèques scientifiques et mathématiques de Python, NumPy et SciPy [74].
- **TensorFlow :** Est une bibliothèque de deep-learning open-source créée par Google qui est utilisée pour modéliser les architectures de deep learning en effectuant des calculs numériques sophistiqués et un certain nombre d'autres tâches. Il peut déployer des calculs facilement sur plusieurs plates-formes, y compris les processeurs et les GPU[75].
- **Keras :** Python a été utilisé pour développer l'API réseau neuronal connue sous le nom de Keras. C'est une bibliothèque open-source qui fonctionne sur des frameworks comme Theano et TensorFlow. Bien que le fait que TensorFlow offre des opérations de plus en

plus sophistiquées pour obtenir un bon contrôle pour développer un type spécifique de modèle et nous permet de mieux comprendre ce qui se passe à l'intérieur d'un réseau DL, Keras n'offre pas autant que TensorFlow, qui fournit toutes les fonctionnalités générales requises pour construire des modèles d'apprentissage profond[76].

- **Seaborn :** Python a un module appelé Seaborn qui est utilisé pour faire des graphiques statistiques. Il est basé sur zMatplotlib et intègre les structures de Pandas. Cette bibliothèque offre simplicité et nouvelles capacités tout en étant aussi puissante que Matplotlib. Elle permet une exploration et une compréhension rapides des données. [77].
- **NLTK :** Natural Language Tool Kit est une bibliothèque Python dédiée au traitement naturel du langage ou Natural Language Processing. Il comprend une suite des bibliothèques de traitement de texte pour la classification, tokenization, stemming, l'analyse syntaxique et le raisonnement sémantique, des wrappers pour des bibliothèques NLP industrielles, un forum de discussion actif, ainsi que des interfaces faciles à utiliser pour plus de 50 corpus et ressources lexicales comme WordNet [78].

3 Expérimentation du module reconnaissance d'émotion multimodale

Dans cette partie nous allons présenter la mise en œuvre et le développement du module de reconnaissance d'émotion multimodale. Nous commençons par la présentation de la base, puis l'expérimentation et les résultats de chaque étape, et enfin, une évaluation des résultats.

3.1 La base de donnée IEMOCAP

"Interactive Emotional Dyadic Motion Capture (IEMOCAP)" est une base de données multimodale et multispeaker, collectée au laboratoire SAIL de l'USC.Elle contient environ de 10039 des données audiovisuelles, y compris la vidéo, la parole, les transcriptions de texte. Elle se compose de sessions dyadiques où les acteurs exécutent des improvisations , spécifiquement sélectionnés pour susciter des expressions émotionnelles. La base de données IEMOCAP est annotée par de multiples d'étiquettes catégoriques, telles que colère, excitation, tristesse, neutralité[79].

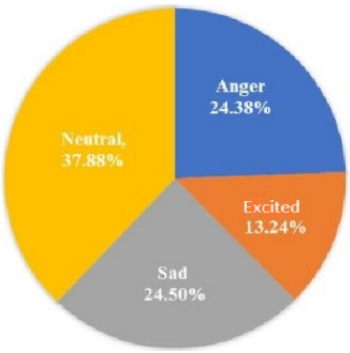

FIGURE 4.1 – Étiquettes de IEMOCAP.

3.2 Sélection des caractéristiques

Ce module est le responsable sur l'extraction des caractéristiques des textes et de vocales. Nous avons traité les textes en utilisant la fonction prédéfinie "tokenizer" où le filtre est d'abord appliqué à ces textes pour supprimer les mots et les symboles vides, ensuite chaque texte est divisé en mots dans la méthode de tokenization.

```
[6]  text = []

     for ses_mod in data2:
         text.append(ses_mod['transcription'])

     MAX_SEQUENCE_LENGTH = 500

     tokenizer = Tokenizer()
     tokenizer.fit_on_texts(text)

     token_tr_X = tokenizer.texts_to_sequences(text)
     x_train_text = []

     x_train_text = pad_sequences(token_tr_X, maxlen=MAX_SEQUENCE_LENGTH)
```

FIGURE 4.2 – Tokenization du texte.

```
shape of features text  (3838, 500)
premiere element dans la base [  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1  16 154 177
 528 300  21 114   1  16   9  73  18   2  20   3  50  84]
```

FIGURE 4.3 – Exemple d'un texte après Tokenization.

Les enregistrements vocaux sont divisés en trames. Le total de 34 caractéristiques est extrait pour chaque vecteur se compose de 3 caractéristiques de domaine temporel (taux de croisement nul, énergie et entropie d'énergie), 5 caractéristiques de domaine spectral (centroïde spectral, propagation spectrale, entropie spectrale, flux spectral, roll-off spectral), 13 MFCC et 13 chromas,voir la figure 4.4.

```
# read speech feature data
x_train_speech = np.load('/content/drive/MyDrive/voiced_feat_without_sil_removal.npy')
print("shape of features speech",x_train_speech.shape)
print("premiere element dans la base ",x_train_speech[2])
```

```
shape of features speech (4936, 100, 34)
premiere element dans la base  [[9.06533292e-02 1.38479130e-04 3.23105605e+00 ... 0.00000000e+00
  0.00000000e+00 0.00000000e+00]
 [1.70990935e-01 2.75597566e-04 3.12120923e+00 ... 0.00000000e+00
  0.00000000e+00 0.00000000e+00]
 [2.40387621e-01 4.22948718e-04 3.29571514e+00 ... 0.00000000e+00
  0.00000000e+00 0.00000000e+00]
```

FIGURE 4.4 – Vecteur numérique de l'audio.

3.3 Implémentation du modèle multimodale

De la conception que nous avons présentée dans le chapitre précédent, et en utilisant des outils nécessaires , nous avons obtenu une réalisation de modèle de reconnaissance des émotions ci-dessous.

Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	[(None, 500)]	0	[]
input_2 (InputLayer)	[(None, 100, 34)]	0	[]
embedding (Embedding)	(None, 500, 128)	350336	['input_1[0][0]']
flatten (Flatten)	(None, 3400)	0	['input_2[0][0]']
lstm (LSTM)	(None, 500, 256)	394240	['embedding[0][0]']
dense_1 (Dense)	(None, 128)	435328	['flatten[0][0]']
lstm_1 (LSTM)	(None, 256)	525312	['lstm[0][0]']
dropout (Dropout)	(None, 128)	0	['dense_1[0][0]']
dense (Dense)	(None, 256)	65792	['lstm_1[0][0]']
dense_2 (Dense)	(None, 256)	33024	['dropout[0][0]']
concatenate (Concatenate)	(None, 512)	0	['dense[0][0]', 'dense_2[0][0]']
dense_3 (Dense)	(None, 256)	131328	['concatenate[0][0]']
dense_4 (Dense)	(None, 4)	1028	['dense_3[0][0]']

```
Total params: 1,936,388
Trainable params: 1,936,388
Non-trainable params: 0
```

FIGURE 4.5 – Implémentation du modèle "Fusion network".

3.4 Évaluation et discussion des résultats

La classification par apprentissage automatique peut être résumée en deux étapes :

1. Apprentissage du modèle à l'aide de l'ensemble de données d'entraînement.
2. Appliquer le modèle formé à l'ensemble de données de test.

3.4.1 Apprentissage du modèle

Nous avons appris le modèle sur les données de l'apprentissage et les valider sur les fichiers de matrices de données de test. Une fois que la performance du modèle sur l'ensemble de validation est conforme à nos normes, nous pouvons l'utiliser pour faire des prédictions sur des données inconnues. Après avoir expérimenté avec une variété d'options, nous avons finalement décidé sur 25 époques (époques sont le nombre d'itérations que le processus de l'apprentissage de modèle est répété), on a obtenu une précision de 77% illustré dans la figure 4.6. Nous avons ensuite archivé le modèle pour une utilisation future.

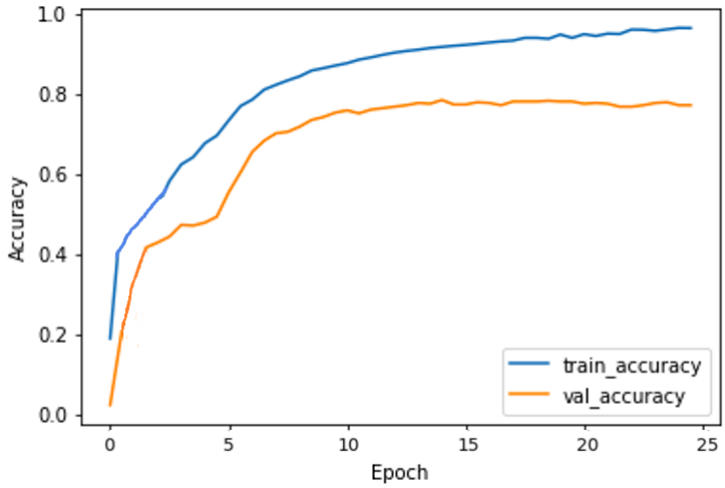

FIGURE 4.6 – Accuracy de l'apprentissage .

3.4.2 Test du modèle

Dans cette partie nous allons tester les performances du modèle avec plusieurs métriques d'évaluation (accuracy,précision,recall).Nous avons fait plusieurs expériences sur la base IEMOCAP , les meilleurs résultats du modèle avec une précision de 72%.

- **matrice de confusion**
  La matrice de confusion est un tableau qui est souvent utilisé pour évaluer la performance d'un modèle de classification. La matrice de confusion contient quatre paramètres :
  Vrais positifs (TP) : observations qui ont été correctement prévues comme positives.
  Faux positifs (FP) : observations qui étaient prévues positives, mais qui sont en fait négatives.
  Faux négatifs (FN) : observations qui étaient prévues négatives, mais qui sont en fait positives.
  Vrais négatifs (TN) : observations qui ont été correctement prédites comme négatives.

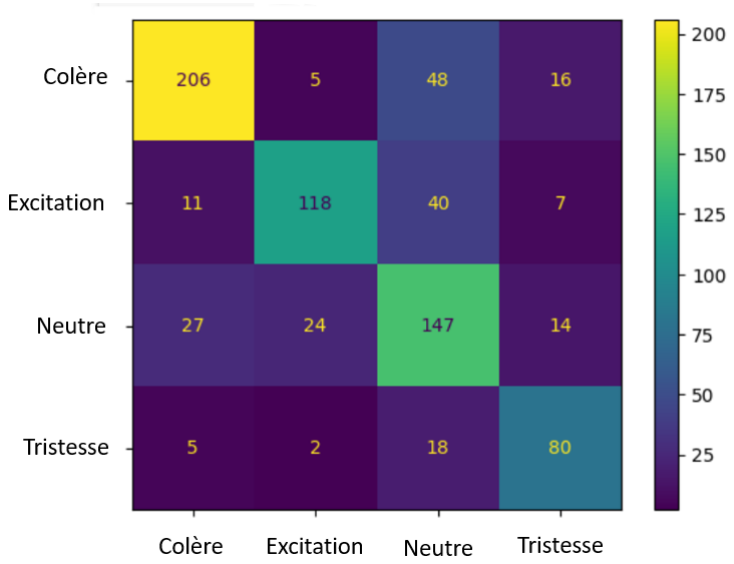

FIGURE 4.7 – Matrice de confusion de modèle "Fusion network".

- **Rapport de classification**
  Ce rapport montre la précision, le rappel, le score f1 et le support pour chaque classe ainsi que l'exactitude globale, la moyenne macro et la moyenne pondérée. Il fournit un résumé détaillé de la performance du modèle sur les données de test. Ce type de rapport de classification vous permet d'identifier :
  -Pour quelles classes le modèle est-il le plus/le moins performant.
  -Tout biais ou déséquilibre dans les résultats lorsqu'il manque des échantillons positifs ou négatifs dans le modèle .

```
24/24 [==============================] - 33s 1s/step
              precision    recall  f1-score   support

         ang       0.83      0.75      0.79       275
         exc       0.79      0.67      0.73       176
         neu       0.58      0.69      0.63       212
         sad       0.68      0.76      0.72       105

    accuracy                           0.72       768
   macro avg       0.72      0.72      0.72       768
weighted avg       0.73      0.72      0.72       768
```

FIGURE 4.8 – Rapport de classification de modèle "fusion network".

- **Accuracy**
  Représente le nombre d'instances de données correctement classées par rapport au nombre total d'instances de données,Il est défini comme :

$$ACC = \frac{TP + TN}{TN + FP + TP + FN}$$

- **Precision**
  Mesure la capacité du classificateur à identifier uniquement les échantillons pertinents. Il est défini comme :
$$PRE = \frac{TP}{FP + TP}$$
- **Recall**
  Mesure la capacité du classificateur de trouver tous les échantillons positifs. Il est défini comme suit :
$$REC = \frac{TP}{FN + TP}$$
- **F1-score**
  Est une moyenne pondérée de la précision et du recall. Elle mesure l'exactitude du modèle en tenant compte de la précision et du rappel. Elle est définie comme suit :
$$Fscore = 2 * \frac{(Recall * Precision)}{(Recall + Precision)}$$

Les résultats obtenus lors de l'essai de ces modèles utilisant ces schémas de sélection de caractéristiques sont présentés dans le tableau suivant :

Mesures / Algorithmes	Accuracy(%)	Recall(%)	Precision(%)	F1-score(%)
Fusion network	72	72	72	72
Fusion au niveau caractéristique	42	42	42	42
Fusion network+attention	43	43	43	43

TABLE 4.1 – Résultats des modèle.

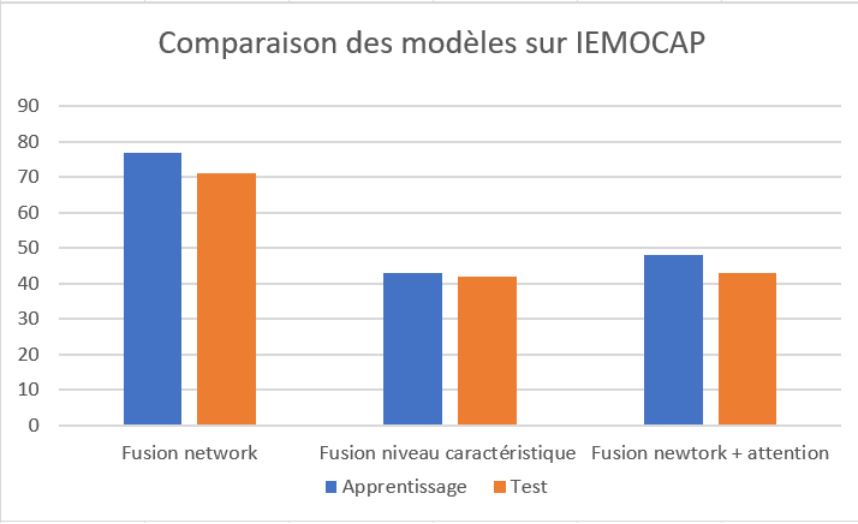

FIGURE 4.9 – Apprentissage et test des modèles.

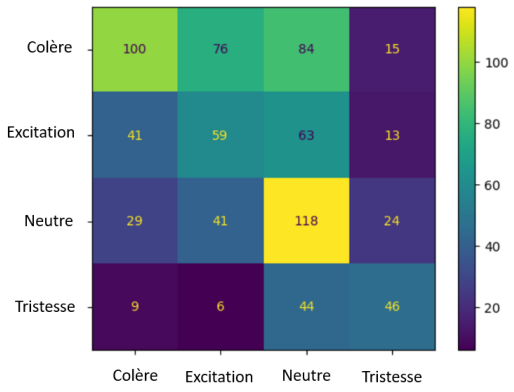

FIGURE 4.10 – Confusion matrice de modèle fusion au niveau caractéristique .

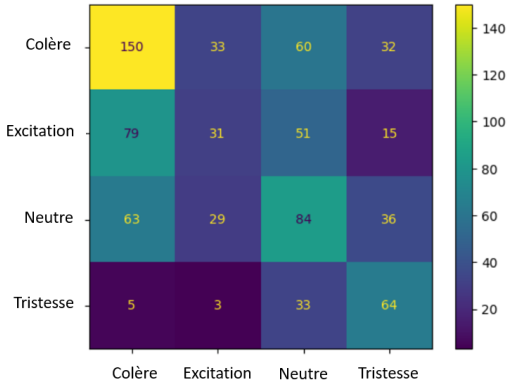

FIGURE 4.12 – Confusion matrice de modèle fusion network+attention.

```
24/24 [==============================] - 1s 24ms/step
              precision    recall  f1-score   support

         ang       0.56      0.36      0.44       275
         exc       0.32      0.34      0.33       176
         neu       0.38      0.56      0.45       212
         sad       0.47      0.44      0.45       105

    accuracy                           0.42       768
   macro avg       0.43      0.42      0.42       768
weighted avg       0.44      0.42      0.42       768
```

FIGURE 4.11 – Rapport de classification fusion au niveau caractéristique.

```
24/24 [==============================] - 4s 136ms/step
              precision    recall  f1-score   support

         ang       0.51      0.55      0.52       275
         exc       0.32      0.18      0.23       176
         neu       0.37      0.40      0.38       212
         sad       0.44      0.61      0.51       105

    accuracy                           0.43       768
   macro avg       0.41      0.43      0.41       768
weighted avg       0.42      0.43      0.41       768
```

FIGURE 4.13 – Rapport de classification de modèle fusion network+attention.

Approche	Modalité	Modèle	Accuracy
Pen Shixin et al[59]	text - vocal	CNN, BLSTM	74%
Davamanyu et al [60]	text - vocal	CNN	71,4%
Suraj Tripathi et al[61]	text - vocal	CNN	76%
Notre approche	text - vocal	CNN , LSTM	**72%**

TABLE 4.2 – Comparaison accuracy des modèles avec notre approche .

3.5 Discussion :

Les meilleurs résultats du modèle ont été obtenus grâce à l'utilisation de LSTM et CNN dans les modules de reconnaissances des émotions de la parole et du texte et avec l'utilisation du modèle GLOVE où nous avons obtenu un taux de 77% sur l'apprentissage et 72% sur le test, le résultat de matrice de confusion de 4 classes de notation pour ce modèle est illustré dans la figure 4.7,depuis cette matrice on peut dire :

-Les échantillons de la classe 1 (colère) sont généralement classés correctement, avec un nombre élevé de vrais positifs (206) et des nombres relativement faibles dans les autres cellules. Le modèle fonctionne donc bien pour la classe 1 .

-Les échantillons de la classe 2 (excitation) sont également un nombre élevé de vrais positifs (118) mais aussi un nombre important de faux négatifs (40). Le modèle a donc une bonne

performance pour la classe 2 .
-Les échantillons de la classe 3 (neutre) ont un nombre assez élevé de vrais positifs (147) mais aussi un nombre raisonnablement élevé de faux positifs (27).
-Les échantillons de classe 4 (tristesse) contiennent le plus faible nombre de vrais positifs (80) et des nombres plus élevés dans d'autres cellules. Le modèle a donc la moins performance de la classe 4.

Quant aux résultats de rapport de classification , ils nous fournissent un rapport de la précision du modèle à chaque classe , où nous voyons que le modèle s'attend très bien à reconnaître l'émotion de la colère et de l'excitation avec une précision de 83% et 79% respectivement. Quant à l'émotion de la tristesse, il s'attend d'une façon acceptable qui atteint la précision 68%, de la neutralité qui souffre à reconnaître avec une précision 58%.

Au cours les résultats indiqués précédemment dans le tableau 4.1 et la figure 4.9 on peut dire l'approche " fusion network " donne le meilleur résultat par rapport aux autres approches avec les précisions 72% et 46% , 45% pour les approches "fusion + attention" et fusion au niveau de caractéristique respectivement.De plus, à partir les résultats de confusion matrice illustré dans les figures 4.7 4.10 4.12 et les rapports de classifications illustré dans les figures 4.8 4.11 4.13 nous concluons que l'approche fondée sur "fusion network" a donné les meilleurs résultats par rapport aux d'autres approches de reconnaissance d'émotion multimodale.

Au cours les résultats indiqués précédemment dans le tableau 4.2 on peut dire que nous avons obtenu le meilleur troisième résultat par rapport aux approches montrées avec un taux de 72% et 76% ,74% ,71,4% pour les approches Suraj Tripathi et al[61] ,Pen Shixin et al[59] et Davamanyu et al [60] respectivement, cela est dû aux types des caractéristiques et les méthodes de traitement qui influent sur les résultats et les performances du système en général, ainsi que la méthode d'intégration de ces modalités.

4 Conclusion

Dans ce chapitre, nous avons décrit les étapes de la réalisation de notre système de reconnaissance d'émotion multimodale basé sur le deep learning , les outils utilisés et les différents résultats obtenus lors du test le modèle . Les résultats sont acceptables et satisfaisants dans une certaine mesure, et nous pouvons affirmer qu'ils répondent aux critères fondamentaux d'un système de reconnaissance d'émotion multimodale.

Conclusion générale

La reconnaissance d'émotions à partir d'une seule modalité (visuelle, audio ou textuelle) reste limitée en termes de performance et de robustesse. L'utilisation de multiples modalités permet d'obtenir de meilleurs résultats.

Dans ce projet de fin d'études, nous nous sommes intéressés au domaine de reconnaissance d'émotion multimodale en proposant une approche et de réaliser un système de reconnaissance émotionnelle où nous avons utilisé les sources de la parole et du texte en utilisant les réseaux neuronaux CNN et LSTM pour la parole et le texte. La technique "fusion network" a été optée pour fusionner les deux sources qui donnaient 72% de précision et qui sont jugées satisfaisantes dans ces certains critères.

La reconnaissance d'émotions multimodale est une approche prometteuse, mais qui nécessite encore des progrès pour faire face à la complexité et la subtilité des émotions humaines. Des recherches supplémentaires, associant apprentissage automatique et compréhension du fonctionnement émotionnel humain sont nécessaires pour faire encore progresser le domaine.

Les contributions de ce travail peuvent être résumées comme suit :

1. Modélisation des données pour les modules de reconnaissance d'émotion multimodales .
2. Modéliser et mettre en œuvre la technique de fusion pour le système multimodale sur l'environnement Colab, y compris toutes les phases de traitement du texte et vocale.
3. Implémentation et expérimentation de l'approche globale proposée sous l'environnement colab avec l'évaluation de ces résultats finale.

Une des limites du travail que nous avons présentées et dans une perspective de recherches futures, Nous espérons de faire ce qui suit :

1. Réalisation un système de reconnaissance d'émotion multimodale utilisant d'autres modalités.
2. Réalisation un système de reconnaissance d'émotion multimodale en temps réel.
3. Intégré le système de reconnaissance d'émotion multimodale dans une plate-forme, y compris e-learning.

Bibliographie

[1] Moïra Mikolajczak. Les compétences émotionnelles : historique et conceptualisation. 2009.

[2] Yagya Raj Pandeya and Joonwhoan Lee. Deep learning-based late fusion of multimodal information for emotion classification of music video. *Multimedia Tools and Applications*, 80 :2887–2905, 2021.

[3] Philippe Claudon 1 and Margot Weber 2. L'émotion : contribution à l'étude psychodynamique du développement de la pensée de l'enfant sans langage en interaction. *Devenir*, 21(1) :61–99, 2009.

[4] la-philosophie. `https://la-philosophie.com/emotion-philosophie#:~:text=Qu'est%2Dce%20qu',%2C%20une%20rupture%20d'%C3%A9quilibre.` Accessed on Février 20, 2023.

[5] Imen Tayari Tayari Meftah. *Modélisation, détection et annotation des états émotionnels à l'aide d'un espace vectoriel multidimensionnel*. PhD thesis, Université Nice Sophia Antipolis, 2013.

[6] Sofie Pringle, Mirko Guaralda, and Severine Mayere. Urban environment characteristics and their implications on emotional happiness and well-being : Proposal of a theoretical and conceptual framework. In *Proceedings of the 12th Liveable Cities Conference*, pages 33–66. Association for Sustainability in Business Inc., 2019.

[7] Wamidh K Mutlag, Shaker K Ali, Zahoor M Aydam, and Bahaa H Taher. Feature extraction methods : a review. In *Journal of Physics : Conference Series*, volume 1591, page 012028. IOP Publishing, 2020.

[8] Dhong Fhel K Gom-os and Kelvin Y Yong. An empirical study on the use of a facial emotion recognition system in guidance counseling utilizing the technology acceptance model and the general comfort questionnaire. *Applied Computing and Informatics*, (ahead-of-print), 2022.

[9] Chung-Hsien Wu, Ze-Jing Chuang, and Yu-Chung Lin. Emotion recognition from text using semantic labels and separable mixture models. *ACM transactions on Asian language information processing (TALIP)*, 5(2) :165–183, 2006.

[10] path partner tech. `https://www.pathpartnertech.com/challenges-faced-by-facial-recognition-system/`. Accessed on Février 20, 2023.

[11] Fatemeh Noroozi, Ciprian Adrian Corneanu, Dorota Kamińska, Tomasz Sapiński, Sergio Escalera, and Gholamreza Anbarjafari. Survey on emotional body gesture recognition. *IEEE transactions on affective computing*, 12(2) :505–523, 2018.

[12] Ahmad R Naghsh-Nilchi and Mohammad Roshanzamir. An efficient algorithm for motion detection based facial expression recognition using optical flow. *International Journal of Computer and Information Engineering*, 2(8) :2724–2729, 2008.

[13] Khadija Lekdioui. *Reconnaissance d'états émotionnels par analyse visuelle du visage et apprentissage machine*. PhD thesis, Université Bourgogne Franche-Comté ; Université Ibn Tofail. Faculté des ..., 2018.

[14] Shashidhar G Koolagudi and K Sreenivasa Rao. Emotion recognition from speech : a review. *International journal of speech technology*, 15 :99–117, 2012.

[15] Danny Oude Bos et al. Eeg-based emotion recognition. *The influence of visual and auditory stimuli*, 56(3) :1–17, 2006.

[16] Andrius Dzedzickis, Artūras Kaklauskas, and Vytautas Bucinskas. Human emotion recognition : Review of sensors and methods. *Sensors*, 20(3) :592, 2020.

[17] Felipe Almeida and Geraldo Xexéo. Word embeddings : A survey. *arXiv preprint arXiv :1901.09069*, 2019.

[18] S Selva Birunda and R Kanniga Devi. A review on word embedding techniques for text classification. *Innovative Data Communication Technologies and Application : Proceedings of ICIDCA 2020*, pages 267–281, 2021.

[19] Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. Visualizing and measuring the geometry of bert. *Advances in Neural Information Processing Systems*, 32, 2019.

[20] Nidhi N Khatri, Zankhana H Shah, and Samip A Patel. Facial expression recognition : A survey. *International Journal of Computer Science and Information Technologies (IJCSIT)*, 5(1) :149–152, 2014.

[21] Rebahi Ghediri Imane, Semri KhawLa, and Belhouchette Kenza. La reconnaissance des emotions de base par les reseaux de neurones : application de deep learning. 2021.

[22] Dalibor Mitrović, Matthias Zeppelzauer, and Christian Breiteneder. Features for content-based audio retrieval. In *Advances in computers*, volume 78, pages 71–150. Elsevier, 2010.

[23] Sabur Ajibola Alim and Nahrul Khair Alang Rashid. Some commonly used speech feature extraction algorithms. In Ricardo Lopez-Ruiz, editor, *From Natural to Artificial Intelligence*, chapter 1. IntechOpen, Rijeka, 2018.

[24] Luefeng Chen, Min Wu, Witold Pedrycz, and Kaoru Hirota. *Emotion Recognition and Understanding for Emotional Human-Robot Interaction Systems*, volume 926. Springer Nature, 2020.

[25] Naveed Ahmed, Zaher Al Aghbari, and Shini Girija. A systematic survey on multimodal emotion recognition using learn-ing algorithms. *Intelligent Systems with Applications*, 17 :200171, 2023.

[26] Soujanya Poria, Amir Hussain, and Erik Cambria. *Multimodal sentiment analysis*. Springer, 2018.

[27] Kashif Ahmad, Khalil Khan, and Ala Al-Fuqaha. Intelligent fusion of deep features for improved waste classification. *IEEE access*, 8 :96495–96504, 2020.

[28] Shizhe Chen and Qin Jin. Multi-modal conditional attention fusion for dimensional emotion prediction. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 571–575, 2016.

[29] Muhammad Naveed Iqbal Qureshi, Jooyoung Oh, Dongrae Cho, Hang Joon Jo, and Boreom Lee. Multimodal discrimination of schizophrenia using hybrid weighted feature concatenation of brain functional connectivity and anatomical features with an extreme learning machine. *Frontiers in neuroinformatics*, 11 :59, 2017.

[30] Garam Lee, Byungkon Kang, Kwangsik Nho, Kyung-Ah Sohn, and Dokyoon Kim. Mildint : deep learning-based multimodal longitudinal data integration framework. *Frontiers in genetics*, 10 :617, 2019.

[31] Wilson Chango, Juan A Lara, Rebeca Cerezo, and Cristobal Romero. A review on data fusion in multimodal learning analytics and educational data mining. *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery*, 12(4) :e1458, 2022.

[32] Utthara Gosa Mangai, Suranjana Samanta, Sukhendu Das, and Pinaki Roy Chowdhury. A survey of decision fusion and feature fusion strategies for pattern classification. *IETE Technical review*, 27(4) :293–307, 2010.

[33] Alican Dogan and Derya Birant. A weighted majority voting ensemble approach for classification. In *2019 4th International Conference on Computer Science and Engineering (UBMK)*, pages 1–6. IEEE, 2019.

[34] Cherifi Dalila, Hafnaoui Imane, and Nait-Ali Amine. Multimodal score-level fusion using hybrid ga-pso for multibiometric system. *Informatica*, 39(2), 2015.

[35] Hamza Osman Ilhan, Gorkem Serbes, and Nizamettin Aydin. Decision and feature level fusion of deep features extracted from public covid-19 data-sets. *arXiv e-prints*, pages arXiv–2011, 2020.

[36] Smriti Srivastava. Accurate human recognition by score-level and feature-level fusion using palm–phalanges print. *Arabian Journal for Science and Engineering*, 43(2) :543–554, 2018.

[37] Yifei Zhang, Désiré Sidibé, Olivier Morel, and Fabrice Mériaudeau. Deep multimodal fusion for semantic image segmentation : A survey. *Image and Vision Computing*, 105 :104042, 2021.

[38] Peng Liu, Lemei Zhang, and Jon Atle Gulla. Dynamic attention-based explainable recommendation with textual and visual fusion. *Information Processing & Management*, 57(6) :102099, 2020.

[39] Jon Driver and Charles Spence. Crossmodal attention. *Current opinion in neurobiology*, 8(2) :245–253, 1998.

[40] Stephanie Badde, Karen T Navarro, and Michael S Landy. Modality-specific attention attenuates visual-tactile integration and recalibration effects by reducing prior expectations of a common source for vision and touch. *Cognition*, 197 :104170, 2020.

[41] Cees GM Snoek, Marcel Worring, and Arnold WM Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402, 2005.

[42] Pilar Manchón Portillo, Guillermo Pérez García, and Gabriel Amores Carredano. Multimodal fusion : a new hybrid strategy for dialogue systems. In *Proceedings of the 8th international Conference on Multimodal interfaces*, pages 357–363, 2006.

[43] dataanalyticspost. DEEP LEARNING . `https://dataanalyticspost.com/Lexique/deep-learning/`. [Online ; accessed 25-mai-2023].

[44] Che-Lun Hung. Deep learning in biomedical informatics. In *Intelligent Nanotechnology*, pages 307–329. Elsevier, 2023.

[45] databasecamp.de. Long Short-Term Memory Networks (LSTM)- simply explained ! . `https://databasecamp.de/en/ml/lstms`. [Online ; accessed 25-mai-2023].

[46] Asifullah Khan, Anabia Sohail, Umme Zahoora, and Aqsa Saeed Qureshi. A survey of the recent architectures of deep convolutional neural networks. *Artificial intelligence review*, 53 :5455–5516, 2020.

[47] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255. PMLR, 2013.

[48] Nitish Srivastava and Russ R Salakhutdinov. Multimodal learning with deep boltzmann machines. *Advances in neural information processing systems*, 25, 2012.

[49] Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. Multimodal convolutional neural networks for matching image and sentence. In *Proceedings of the IEEE international conference on computer vision*, pages 2623–2631, 2015.

[50] Jimmy Ren, Yongtao Hu, Yu-Wing Tai, Chuan Wang, Li Xu, Wenxiu Sun, and Qiong Yan. Look, listen and learn—a multimodal lstm for speaker identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

[51] Youze Wang, Shengsheng Qian, Jun Hu, Quan Fang, and Changsheng Xu. Fake news detection via knowledge-driven multimodal graph convolutional networks. In *Proceedings of the 2020 international conference on multimedia retrieval*, pages 540–547, 2020.

[52] Yeqi Liu, Chuanyang Gong, Ling Yang, and Yingyi Chen. Dstp-rnn : A dual-stage two-phase attention-based recurrent neural network for long-term and multivariate time series prediction. *Expert Systems with Applications*, 143 :113082, 2020.

[53] Peng Hu, Liangli Zhen, Dezhong Peng, and Pei Liu. Scalable deep multimodal learning for cross-modal retrieval. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 635–644, 2019.

[54] Chaoyue Ding, Shiliang Sun, and Jing Zhao. Mst-gat : A multimodal spatial–temporal graph attention network for time series anomaly detection. *Information Fusion*, 89 :527–536, 2023.

[55] Houhong Lu, Yangyang Zhu, Ming Yin, Guofu Yin, and Luofeng Xie. Multimodal fusion convolutional neural network with cross-attention mechanism for internal defect detection of magnetic tile. *IEEE Access*, 10 :60876–60886, 2022.

[56] Wan Ding, Mingyu Xu, Dongyan Huang, Weisi Lin, Minghui Dong, Xinguo Yu, and Haizhou Li. Audio and face video emotion recognition in the wild using deep neural networks and small datasets. In *Proceedings of the 18th ACM international conference on multimodal interaction*, pages 506–513, 2016.

[57] Jian Huang, Jianhua Tao, Bin Liu, Zheng Lian, and Mingyue Niu. Multimodal transformer fusion for continuous emotion recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3507–3511. IEEE, 2020.

[58] Shiqing Zhang, Shiliang Zhang, Tiejun Huang, and Wen Gao. Multimodal deep convolutional neural network for audio-visual emotion recognition. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 281–284, 2016.

[59] Peng Shixin, Chen Kai, Tian Tian, and Chen Jingying. An autoencoder-based feature level fusion for speech emotion recognition. *Digital Communications and Networks*, 2022.

[60] Devamanyu Hazarika, Sruthi Gorantla, Soujanya Poria, and Roger Zimmermann. Self-attentive feature-level fusion for multimodal emotion detection. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 196–201. IEEE, 2018.

[61] Suraj Tripathi, Abhay Kumar, Abhiram Ramesh, Chirag Singh, and Promod Yenigalla. Deep learning based emotion recognition system using speech features and transcriptions. *arXiv preprint arXiv :1906.05681*, 2019.

[62] Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *2016 IEEE 16th international conference on data mining (ICDM)*, pages 439–448. IEEE, 2016.

[63] Ashish Ramayee Asokan, Nidarshan Kumar, Anirudh V Ragam, and SS Shylaja. Interpretability for multimodal emotion recognition using concept activation vectors. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 01–08. IEEE, 2022.

[64] Bahar Hatipoglu Yilmaz and Cemal Kose. A novel signal to image transformation and feature level fusion for multimodal emotion recognition. *Biomedical Engineering/Biomedizinische Technik*, 66(4) :353–362, 2021.

[65] Yong Zhang, Cheng Cheng, and Yidie Zhang. Multimodal emotion recognition using a hierarchical fusion convolutional neural network. *IEEE access*, 9 :7943–7951, 2021.

[66] Yongrui Huang, Jianhao Yang, Pengkai Liao, and Jiahui Pan. Fusion of facial expressions and eeg for multimodal emotion recognition. *Computational intelligence and neuroscience*, 2017, 2017.

[67] Wei Liu, Wei-Long Zheng, and Bao-Liang Lu. Multimodal emotion recognition using multimodal deep learning. *arXiv preprint arXiv :1602.08225*, 2016.

[68] tokenex. Tokenization . `https://www.tokenex.com/blog/what-is-tokenization/`. [Online ; accessed 25-mai-2023].

[69] google. Colaboratory. https://research.google.com/colaboratory/faq.html. [Online ; accessed 25-mai-2023].

[70] datascientest. Python : Focus sur le langage le plus populaire . https://datascientest.com/python-tout-savoir#:~:text=C'est%20quoi%20le%20langage,programmation%20et%20de%20d%C3%A9veloppement%20logiciel. [Online ; accessed 25-mai-2023].

[71] datascientest. NumPy : la bibliothèque Python la plus utilisée en Data Science . https://datascientest.com/numpy. [Online ; accessed 25-mai-2023].

[72] datascientest. Pandas : la bibliothèque Python dédiée à la Data Science . https://datascientest.com/pandas-python-data-science. [Online ; accessed 25-mai-2023].

[73] datacamp. Introduction to Plotting with Matplotlib in Python . https://www.datacamp.com/tutorial/matplotlib-tutorial-python. [Online ; accessed 25-mai-2023].

[74] nvidia. Scikit-learn . https://www.nvidia.com/en-us/glossary/data-science/scikit-learn/. [Online ; accessed 25-mai-2023].

[75] pypi. tensorflow . https://pypi.org/project/tensorflow/. [Online ; accessed 25-mai-2023].

[76] datascientest. Keras : tout savoir sur l'API de Deep Learning . https://datascientest.com/keras. [Online ; accessed 25-mai-2023].

[77] datacamp. Seaborn : all about the Data Visualization tool in Python . https://datascientest.com/en/seaborn_and_data_visualization. [Online ; accessed 25-mai-2023].

[78] datacamp. NTLK : guide de l'outil de Traitement Naturel du Langage en Python . https://datascientest.com/nltk. [Online ; accessed 25-mai-2023].

[79] University of Southern California. IEMOCAP . https://sail.usc.edu/iemocap/. [Online ; accessed 25-mai-2023].

.