République Algérienne Démocratique et Populaire
Ministère de l'enseignement supérieur et de la recherche scientifique

UNIVERSITE Dr. TAHAR MOULAY SAIDA

FACULTE : TECHNOLOGIE

DEPARTEMENT : INFORMATIQUE

Mémoire de Master

Option :

Modélisation Informatique des Connaissances et du Raisonnement (MICR)

Thème

Speech Emotion Recognition using deep learning

Présenté par:

MAZOUNI Soumaia Saâdia.

KHAROUBI Chaimaâ.

Encadré par:

Mr. BOUDIA Mohamed Amine.

Co-encadré par :

BENALIOUA Ghania.

Promotion : *septembre 2020*

REMERCIEMENTS :

Tout d'abord autant que binôme nous tenons à remercier Allah pour tout le bien, la santé et le courage que nous a donné pour accomplir ce travail .

Un grand merci pour nos pères et mères Mohamed, Bachir, Halima et Radia ; pour tous et spécialement leurs encouragements, leurs soutiens, leurs amours ...un merci pour nos frères et sœurs(ABDELKRIM, NADJET, ABDELMALEK , NOUR EL HOUDA et HIBA).

Nos sincère remerciements pour notre encadreur D_r Mohamed Amine BOUDIA pour son soutien, son aide, ses précieux conseils pour élaborer ce travail.

Un très grand merci pour Ghania BENALIOUA pour sa patience, sa disponibilité et surtout ses judicieux conseils, qui ont contribué à alimenter notre réflexion et pour le partage de ses connaissances et expériences.

Sans oublier de remercier notre chère Houda pour sa présence surtout ,Un merci également à notre chère Madjda.

Et tous ceux qui nous ont aidés de prés ou de loin.

ملخص

في السنوات الأخيرة ، استثمر الباحثون بكثافة في مجال الذكاء الاصطناعي والروبوتات حتى تتمكن هذه الروبوتات من التواصل بشكل طبيعي قدر الإمكان مع البشر. تحتاج هذه الروبوتات إلى التعرف على المشاعر البشرية لفهمها والاستجابة لرغباتهم. يمكن للمرء أن يتعرف على عاطفة الشخص إما من خلال تعابير وجهه أو صوته أو إيماءات جسده أو من خلال كلماته. في هذه الأطروحة ركزنا فقط على الخطاب ، وعملنا مع مجموع $RAVDESS$ ، التي تحتوي على ٨ مشاعر (محايدة ، سعيدة ، حزينة ، خوف ، غضب ، مفاجأة ، هدوء واشمئزاز). للتعرف على المشاعر ، استخدمنا التعلم العميق الذي اختبرنا فيه العديد من البنى المختلفة بناءً على CNN و RNNوقارننا تفاصيلها.

الكلمات المفتاحية: التعلم العميق ، التعرف على المشاعر الكلامية ، الشبكة العصبية التلافيفية ، الشبكة العصبية المتكررة.

Abstract

In recent years, researchers have been investing intensively in the field of AI and robotics so that these robots can communicate as normally as possible with humans. These robots need to recognize human emotions to understand them and respond to their desires. We can recognize the emotion of the person either by his facial expressions, his voice, the gestures of his body or by his words. In this memory we just focused on the speech, we worked with the RAVDESS corpus, which contains 8 emotions (neutral, happy, sad, fear, angry, surprise, calm and disgust). For emotion recognition, we have used deep learning which we tested two different architectures based on CNN, RNN and compared their accuracies.

Keywords: Deep learning, speech emotion recognition, convolutional neuronal network, recurrent neural network.

Résumé

Ces dernières années, les chercheurs ont investi intensivement dans le domaine de l'IA et de la robotique afin que ces robots puissent communiquer le plus normalement possible avec les humains. Ces robots ont besoin de reconnaître les émotions humaines pour les comprendre et répondre à leurs désirs. On peut reconnaître l'émotion de la personne soit par ses expressions faciales, sa voix, les gestes de son corps ou par ses mots. Dans ce mémoire nous nous sommes juste concentrés sur le discours, nous avons travaillé avec le corpus RAVDESS, qui contient 8 émotions (neutre, heureuse, triste, peur, colère, surprise, calme et dégoût). Pour la reconnaissance des émotions, nous avons utilisé l'apprentissage en profondeur (deep learning) que nous avons testé plusieurs architectures différentes basées sur CNN, RNN et comparé leurs précisions.

Mots clés : apprentissage profond, reconnaissance des émotions de la parole, réseau neuronal convolutif, réseau neuronal récurrent.

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

ACRONYMES

SER : Speech Emotion Recognition.
DL : Deep Learning.
ML : Machine Learning .
AI : Artificiel Inteligence.
HMM : Hidden Markov Model.
DNN : Deep Neuronal Network.
SVM : Support Vector Machine.
GMM : Gaussien Mixture Model.
ANN : Artificiel Neuronal Network.
CNN : Convolutional Neuronal Network.
RNN : Recurrent Neuronal Network.
LSTM : Long Short Term Memory.
GRU : Gated Recurrent Unit .
BLSTM : Bidirectional Long Short Term Memory.
RAVDESS : Ryerson Audio-Visual Database of Emotional Speech and Song.
MFCC : Mel-Frequency Cepstral Coefficients.
ReLU : Rectified Linear Unit.

INTRODUCTION GÉNÉRALE.

L'informatique est un domaine qui a vu le jour dans les années 1945, dont il tente d'automatiser l'information pour faciliter la vie de l'homme. Mais au fil du temps ce terme a marqué des avancées technologiques énormes, qu'ils ont permis aux chercheurs d'oser s'approprié des idées innovantes comme « le test d'Alan Turing » en 1950 qui porte sur est-ce qu'une machine peut maintenir une discussion normale avec un humain sans qu'il se rende compte. Et c'est depuis cela, que les chercheurs se sont focalisés sur l'idées de rendre la machine intelligente pour qu'elle puisse comprendre, communiquer et raisonner. Et pour réaliser cela, la reconnaissance des émotions reste une faculté très importante que la machine doit avoir pour interagir le plus normalement possible avec l'humain.

L'émotion peut être reconnu de plusieurs manières soit par le visage et les expressions faciales, par les gestes du corps ou bien par les intonations vocales ou la voix ; car ces manières expriment un état interne qui révèle nos sentiments, nos préférences ou bien nos réactions .

Ce mémoire s'intéressera particulièrement à la partie reconnaissance des émotions à partir les intonations vocales c'est-à-dire la voix. Ce domaine devient de plus en plus important car beaucoup de domaines peuvent bénéficier de cette technologie par exemple le domaine de la médecine ou la télé-médecine , le marketing , Éducation. Cependant, détecter l'émotion d'un locuteur dans des conditions de la vie réelle demeure un défi.

Objectif

Dans ce mémoire notre objectif est d'explorer le domaine de la reconnaissance des émotions par la parole (speech emotion recognition SER) et de donner un aperçu sur ce problème comment est apparu voir les techniques utilisées avant leurs résultats également. Comme nous allons aussi parler en détail sur l'approche du deep learning, et sur notre contribution

pour résoudre ce problème qui est le développement de notre propre model.

Organisation du mémoire .

Notre mémoire est constitué de trois chapitres :

Le chapitre 1

Dans le premier chapitre nous allons présente la technique utilisée dans ce mémoire pour résoudre ce problème qui est le deep leaning son principe, ses différentes architectures et ses différentes domaines d'application.

Le chapitre 2

Dans le second chapitre nous allons présenter le problème de la reconnaissance des émotions par la parole un petit historique, les approches utilisées auparavant pour résoudre ce problème et leurs inconvénients.

Le chapitre 3

Le troisième chapitre sera une description de nos modèles proposés, les outils utilisés et les résultats obtenus.

CHAPITRE 1

DEEP LEARNING (L'APPRENTISSAGE PROFOND) .

Sommaire

1.1 Introduction

Le deep Learning (DL) un sous domaine de l'apprentissage automatique(ML) qui est un sous domaine de l'intelligence artificielle(AI). Le DL se base sur les réseaux de neurones artificielles qui sont inspirés du fonctionnement du cerveau humain. C'est un domaine de recherche qui à évoluer ces dernières années et qui a apporté beaucoup de nouveautés au domaine de la recherche en générale et à l'intelligence artificielle spécialement et qui l'a rendu plus efficace et plus accessible.

1.2 Historique

Les débuts de l'IA remontent à Alan Turing dans les années 1950 [1], c'est une branche de l'Informatique fondamentale s'est développée avec pour objectif la simulation des comportements du cerveau humain. Les premières tentatives de modélisation du cerveau sont anciennes et précèdent même l'ère informatique. En effet, dans l'imaginaire commun, lorsqu'on parle d'intelligence artificielle on désigne par là un programme qui peut effectuer des tâches d'humain, en apprenant toute seule. Or, l'IA telle que définie dans l'industrie est plutôt « des algorithmes plus ou moins évolués qui imitent des actions humaines » [1]. L'IA a beaucoup évolué grâce notamment à l'émergence du Cloud Computing et du Big Data, soit d'une puissance de calcul peu coûteuse et de l'accessibilité à un grand nombre de données. Ainsi, les machines ne sont plus programmées ; elles apprennent.[2]

En 1980 apparait le terme d'apprentissage automatique (machine learning en anglais ML), est un sous-domaine de l'IA qui s'intéresse en particulier aux capacités d'apprentissage. Le principe est de reproduire un comportement non pas en le programmant à la main dans un ordinateur, mais en concevant un système plus général capable d'apprendre à partir d'exemples à résoudre votre problème[3].

En 2010 un nouveau terme fait son apparition et c'est l'apprentissage profond" (en anglais deep Learning DL) et c'est l'abréviation du terme "apprentissage dans les réseaux de neurones profonds".[3] et comme son nom l'indique il se base sur les réseaux de neurones artificielles.

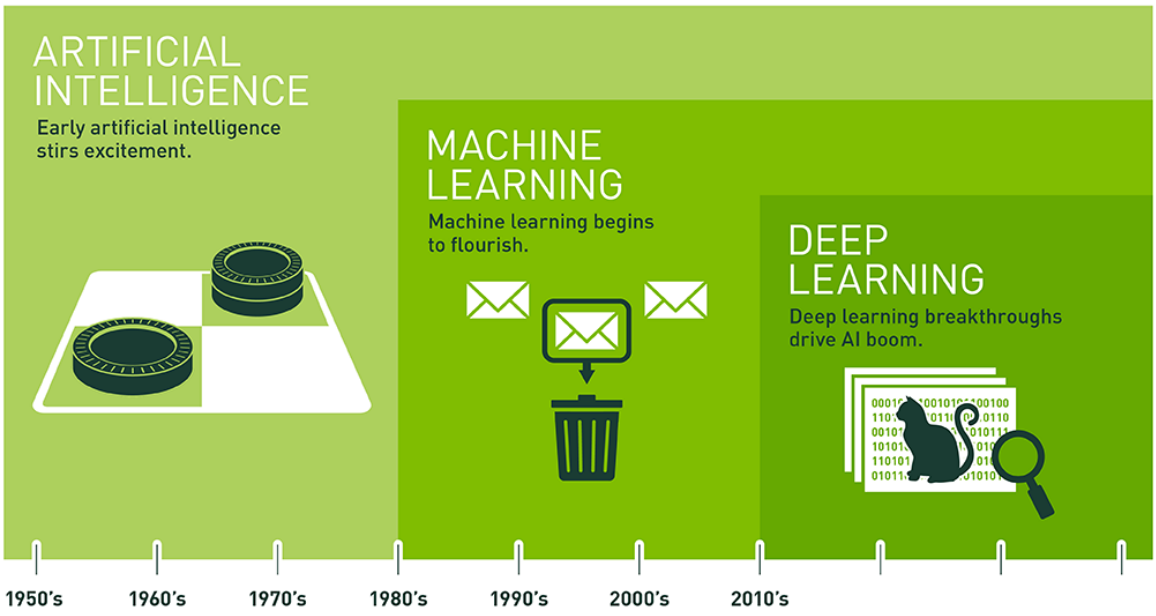

Figure 1.1: Échelle de temps pour l'intelligence artificielle, apprentissage automatique et l'apprentissage profond [3].

1.3 Définition de l'apprentissage profond (Deep Learning DL)

Dans l'apprentissage automatique, un programme analyse un ensemble de données afin de tirer des règles qui permettront de tirer des conclusions sur de nouvelles données. L'apprentissage profond est basé sur les « réseaux de neurones artificiels », constituer de milliers de neurones qui effectuent chacune de simples opérations. Les sorties d'une première couche servent comme entrée d'une deuxième couche et ainsi de suite. Les avancées du deep leaning ont été possibles grâce à la grande puissance des ordinateurs et au développement de vastes ensemble de données « big data ».[4]

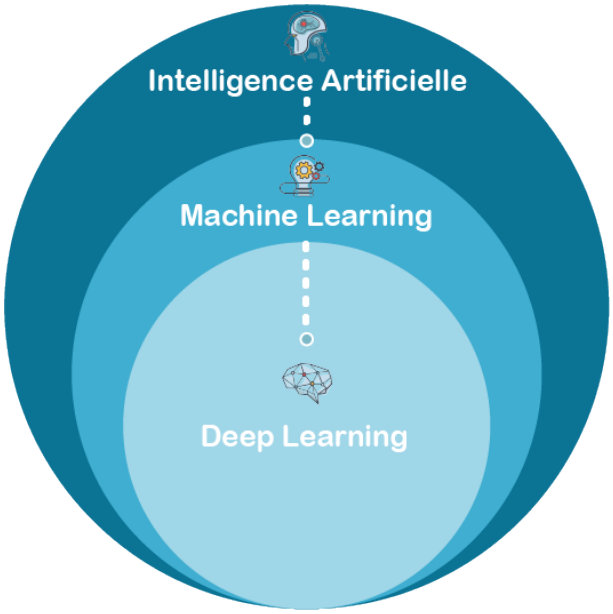

Figure 1.2: La relation entre l'intelligence artificielle le machine learning et le deep learning.[5].

1.4 Les réseaux de neurones

1.4.1 Définition

Les réseaux de neurones artificiels sont des réseaux fortement connectés de processeurs élémentaires fonctionnant en parallèle. Chaque processeur élémentaire calcule une sortie unique sur la base des informations qu'il reçoit. Toute structure hiérarchique de réseaux est évidemment un réseau[6].

Les réseaux de neurones artificiels consistent en des modèles plus ou moins inspirés du fonctionnement cérébral de l'être humain en se basant principalement sur le concept de neurone[7].

1.4.2 Les neurones biologiques

Le cerveau humain est constitué de près de 100 milliards de neurones reliés entre eux, au cœur du neurone on retrouve son noyau, ou « corps cellulaire ». Ces neurones sont connectés à des branches appelées « dendrites » qui transmettent les signaux depuis l'extérieur vers le corps cellulaire ce dernier traite l'information et la transmet via son axone, à la fin de l'axone se trouve les synapses[8].

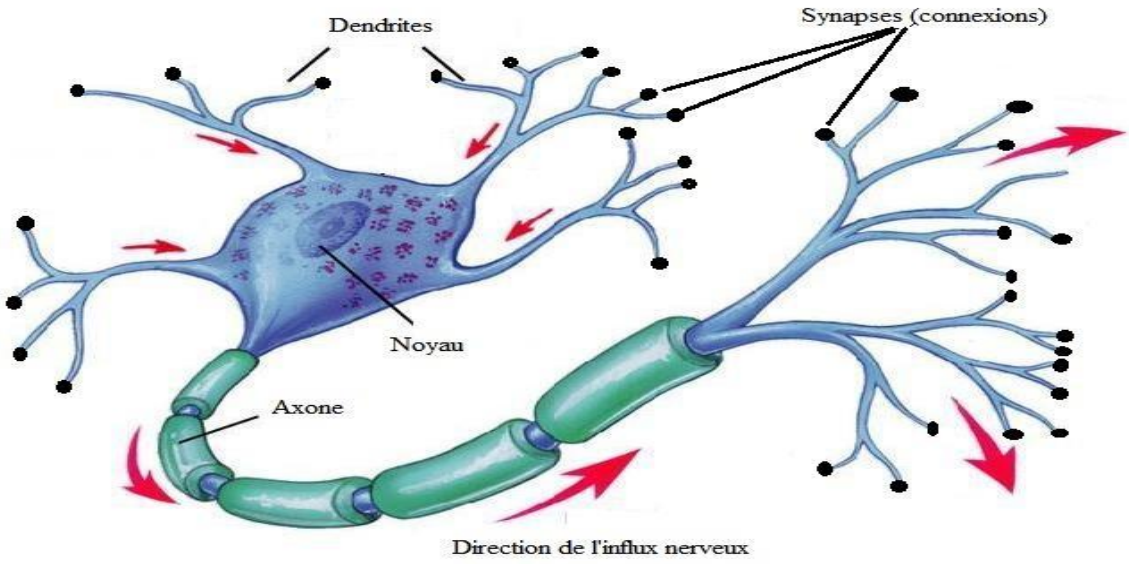

Figure 1.3: Modèle d'un neurone biologique. [7].

1.4.3 Les neurones formels

Un neurone formel est une fonction algébrique non linéaire et bornée, dont la valeur dépend des paramètres appelés poids. Les variables de cette fonction sont habituellement appelées "entrées" du neurone, et la valeur de la fonction est appelée sa "sortie" [9].

On peut représenter graphiquement un neurone comme indiqué sur la figure 4

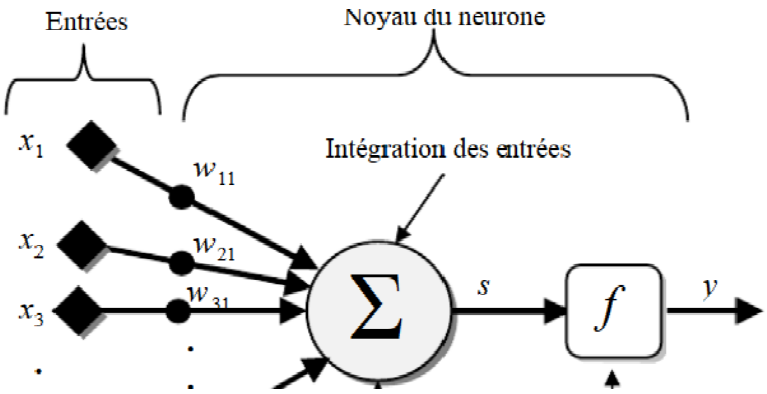

Figure 1.4: Modèle d'un neurone artificiel. [10].

Le modèle du neurone formel est proposé par $W.M.Culloch$ et $W.Pitts$ en 1943 on en trouve [9] :

Les x_i c'est les vecteurs d'entrées.

Les w_{ij} c'est les poids synaptiques du neurone j.

Ces poids pondèrent les entrées et peuvent être modifiés par apprentissage[9].

Biais : permet d'ajouter de la flexibilité au réseau [9].

Noyau : intègre toutes les entrées et le biais et calcul la sortie du neurone selon une fonction d'activation [9].

Un neurone est essentiellement constitué d'un intégrateur qui effectue la somme pondérée de ses entrées. Le résultat s de cette somme est ensuite transformé par une fonction de transfert f qui produit la sortie y du neurone. les n entrées du neurone correspondent au vecteur$x = [x_1, x_2, \ldots, x_n]^T$, alors que $w = [w_{11}, w_{21}, \ldots, w_{n1}]^T$ représente le vecteur des poids du neurone[9]. La sortie s de l'intégrateur est donnée par l'équation suivante [9] :

$$s = \sum_{i=1}^{n} w_{i1} x_1 \mp b \tag{1.1}$$

Cette sortie correspond 'a une somme pondérée des poids et des entrées plus ce qu'on nomme le biais b du neurone. Le résultat s de la somme pond´er´ee s'appelle le niveau d'activation du neurone. Le biais b s'appelle aussi le seuil d'activation du neurone. Lorsque le niveau d'activation atteint ou dépasse le seuil b, alors l'argument de f devient positif (ou nul). Sinon, il est négatif[9].

Les réseaux de neurones sont caractérisés par l'architecture (l'organisation des neurones), l'apprentissage (méthode de détermination des poids de connexions), et par leur fonction d'activation [11].

1.4.3.1 Fonctions d'activations

Différentes fonctions d'activation peuvent exister pour activer le neurone dont elles sont énumérées dans le tableau 1. Les fonctions les plus utilisés sont « seuil » (en anglais « hard limit »), « linéaire » et « sigmoïde ». [9]

Nom de la fonction	Relation entrée/sortie	Icône
Seuil	y = 0 Si (s<0) $y = 1Si(s \geq 0)$	
Seuil symétrique	y =-1 Si(s<0) $y = 1Si(s \geq 0)$	
Linéaire	y=s	
Linéaire saturée	$y = 0Si(s \leq 0)$ $y = sSi(0 \leq s \leq 1)$ y = 1 si s>1	
Linéaire saturée symétrique	y = -1 si s < -1 $y = sSi(-1 \leq s \leq 1)$ y =1 si s > 1	
Linéaire positive (ReLU)	$y = 0Si(s \leq 0)$ $y = sSi(s \geq 0)$	
Sigmoïde	$y = \frac{1}{1+e^{-s}}$ (1.2)	
Tangente hyperbolique (Tanh)	$y = \frac{e^{s}-e^{-s}}{e^{s}+e^{-s}}$ (1.3)	

Table 1.1: Différentes fonctions d'activations[9].

Pour mettre en œuvre un système de réseaux de neurones artificiel il suffit de correspondre chaque élément du neurone biologique au neurone formel [9] le tableau 2 suivant pourra résumer cette modélisation :

Neurone biologique	Neurone formel
Synapses	Poids des connexions
Axones	Signal de sortie
Dendrites	Signal d'entrée
Noyau ou Somma	Fonction d'activation

Table 1.2: Analogie entre le neurone biologique et le neurone formel [9].

1.4.3.2 Architecture des réseaux de neurones classiques

Il existe deux grands types d'architectures de réseaux de neurones : les réseaux de neurones non bouclés et les réseaux de neurones bouclés. On peut reconnaitre une architecture tout dépend de la façon dont les neurones sont organisés[9].

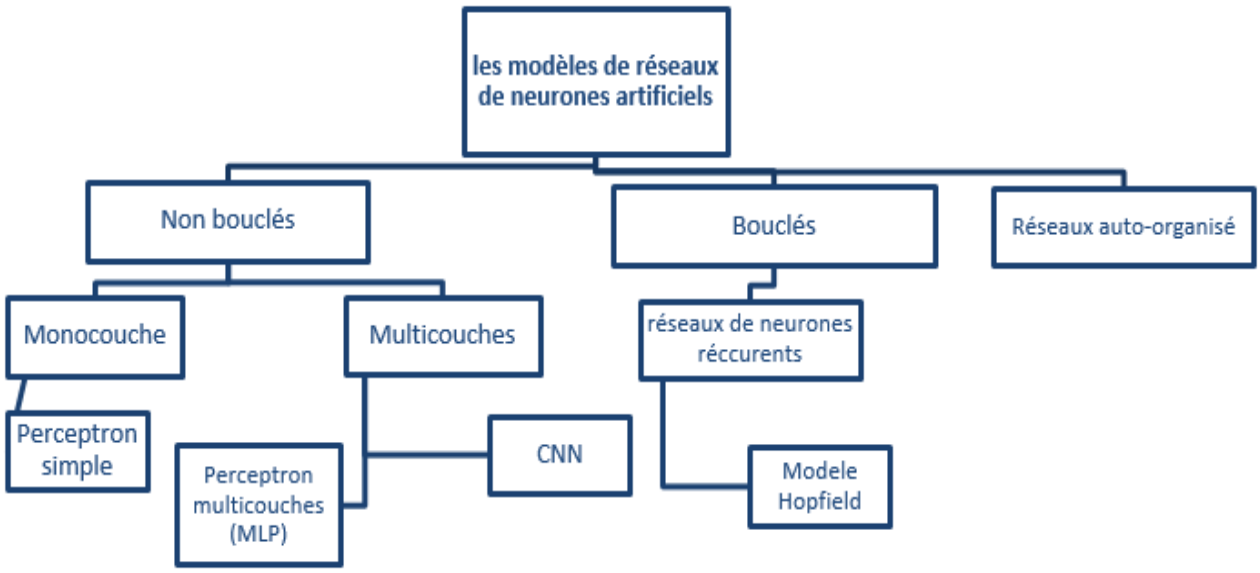

Figure 1.5: les types de réseaux de neurones artificiels.

Les réseaux de neurones non bouclés :

Un réseau de neurones non bouclé dit aussi Famille des réseaux à propagation avant ou feed-forward : est représenté graphiquement par un ensemble de neurones "connectés" entre eux,

l'information circulant des entrées vers les sorties sans "retour en arrière"; si l'on représente le réseau comme un graphe dont les nœuds sont les neurones et les arêtes les "connexions" entre ceux-ci, le graphe d'un réseau non bouclé est acyclique[9].

Réseaux de neurones monocouches

* **Perceptron simple :**
  Ce réseau est dit simple car il ne se compose que de deux couches : une couche d'entrée et une couche de sortie ce qui implique une seule matrice de poids. L'ensemble des unités de la couche d'entrée sont connectés à celles de la couche de sortie[12].

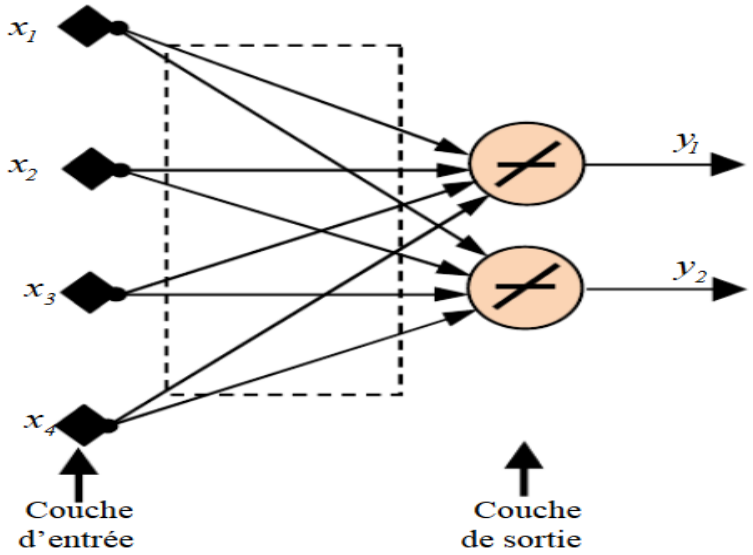

Figure 1.6: Schéma d'un réseau de neurones monocouche. [12].

Réseaux de neurones multicouches :

* **Le perceptron multicouche :**
  Désigné par le sigle MLP (pour Multi-layer Perceptron), le perceptron multicouche se compose d'une couche d'entrée, d'une couche de sortie et d'une ou plusieurs couches cachées. Si le réseau possède n couches, alors il possède n-1 matrice de poids (une entre chaque suite de couches)[12]

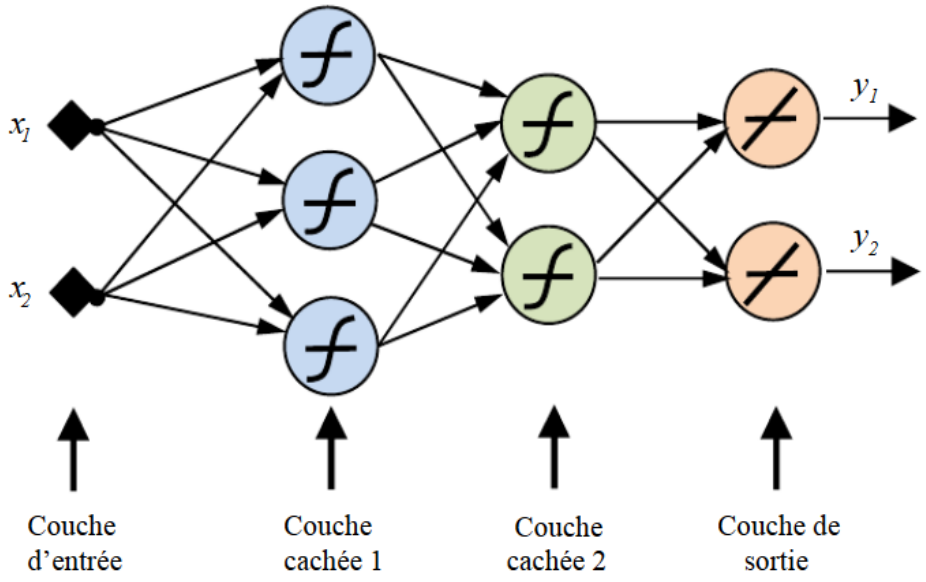

Figure 1.7: Schéma d'un réseau de neurones non bouclé (Perceptron multicouches). [12].

Le réseau de neurones à convolution : (CNN pour Convolutionnal Neural Network)sont aussi des réseaux feed-forward et qui correspondent, en simplifiant beaucoup, à un empilement de perceptron multicouche. Chacun traitant une portion de l'information globale. Ces réseaux sont surtout utilisés pour la reconnaissance d'image, de vidéos ou encore dans le traitement naturel du langage[9].

Les réseaux de neurones bouclés :

Contrairement aux réseaux de neurones non bouclés dont le graphe de connexions est acyclique, les réseaux de neurones bouclés peuvent avoir une topologie de connexions quelconque, comprenant notamment des boucles qui ramènent aux entrées la valeur d'une ou plusieurs sorties. Pour qu'un tel système soit causal, il faut évidemment qu'à toute boucle soit associé un retard : un réseau de neurones bouclé est donc un système dynamique, régi par des équations différentielles[9].

Il s'agit donc de réseaux de neurones avec retour en arrière (feedback network or recurrent network), (Figure 5).

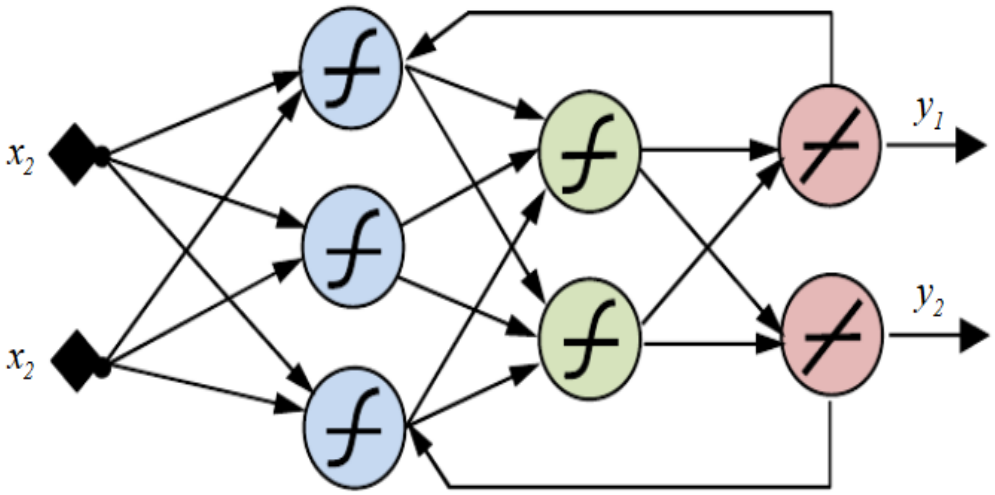

Figure 1.8: Schéma de réseau de neurones bouclé. [12].

* **Famille des réseaux de neurones récurrents (RNN recurrent neural network) :** Les RNNs sont des réseaux de neurones qui comportent des cycles dans leur graphe de connectivité. Ces cycles permettent au réseau d'entretenir une information en se l'envoyant à lui-même. Cela change la dynamique du réseau de neurones et l'amène à s'auto-entretenir. Ces modèles étaient souvent plébiscités notamment pour le traitement automatique de la parole, et plus généralement de séquences, car leurs caractéristiques leur permettent d'apprendre, de stocker et de prendre en compte l'information contextuelle passée lors de traitement de l'information à l'instant présent [12].
* **Le modèle de Hopfield :**
c'est un réseau qui se compose d'une seule couche où toutes les unités sont interconnectées. Il s'agit d'une mémoire auto-associative : constituée d'une seule couche qui représente à la fois l'entrée du réseau et sa sortie. Autrement dit, une mémoire auto-associative a autant d'entrées que de sorties. Ce réseau est assimilé à une mémoire adressable par son contenu : les connaissances mémorisées étant distribuées dans le réseau et non localisées à une adresse, il est possible de récupérer l'entièreté d'une donnée, juste en présentant une version dégradée (partielle ou bruitée) [Rougier, 2000][12].

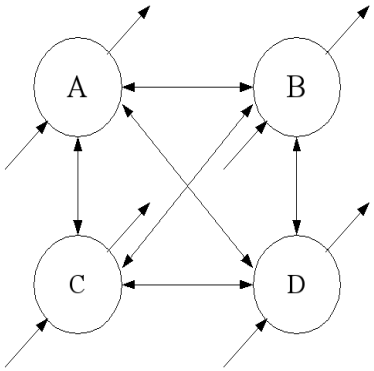

Figure 1.9: Réseau de neurone de Hopfield. [13].

Réseaux auto-organisé :

Inspiré de l'organisation du cortex, les réseaux auto-organisés se distinguent par une connectivité locale. Ils sont surtout adaptés pour le traitement d'informations spatiales. Le modèle le plus connu de ce type est la carte auto-organisatrices de Kohonen appelée aussi carte auto-adaptative ou SOM (Self Organizing Map)[6].

Ces réseaux utilisent des méthodes d'apprentissage non-supervisées. Ils peuvent être utilisé pour cartographier un espace réel ou encore étudier la répartition de données dans un espace de grandes dimensions comme dans le cas de problème de quantification vectorielle, de clusterisation ou de classification. [12]

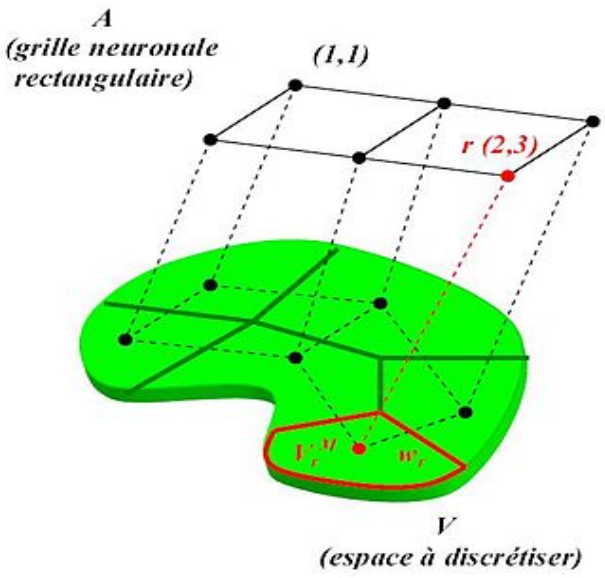

Figure 1.10: carte auto-adaptif[14].

1.4.3.3 Architecture des réseaux de neurones profond

Il existe un grand nombre d'architectures profondes. Nous allons détailler les réseaux de neurones convolutifs (CNNs), les réseaux de neurones récurrents(RNNs).

Les réseaux de neurones convolutifs (CNNs)

Ils sont un type de réseau de neurones spécialisés pour le traitement de données ayant une topologie semblable à une grille leurs principe c'est le filtrage. Les exemples comprennent des données de type série temporelle sont considérées comme une grille 1D en prenant des échantillons à des intervalles de temps réguliers, les données de type image se représentent en 2D ou en 3D, il y a 2 dimensions qui correspondent à la largeur et à la hauteur de l'image et une troisième dimension qui correspond à la composante couleur [15]

Le nom « réseau de neurones convolutif » indique que le réseau emploie une opération mathématique appelée convolution à la place de la multiplication matricielle dans au moins une de leurs couches. La convolution est une opération linéaire spéciale. [16]

Le rôle du ConvNet est de réduire les images sous une forme plus facile à traiter, sans perdre les fonctionnalités qui sont essentielles pour obtenir une bonne prédiction [17].

Il existe différentes architectures de CNN disponibles qui ont été essentielles dans la construction d'algorithmes qui alimentent et alimenteront l'IA dans son ensemble dans un avenir prévisible. Certains d'entre eux ont été énumérés ci-dessous[17] :

LeNet, AlexNet, VGGNet, GoogLeNet, ResNet, ZFNet.

* **Couche de convolution (CONV) :**
  C'est la composante clé des réseaux de neurones convolutifs, et constitue toujours au moins leur première couche. C'est la couche qui effectue le plus de calcul lourd. 3 paramètres pour dimensionner cette couche (la profondeur, le pas, la marge) :La profondeur(le nombre de neurones associés à un même champ de récepteur). Le pas(il contrôle le chevauchement des champs récepteurs).La marge(permet de contrôler la dimension spatiale du volume de la sortie).[4]
* **Couche de pooling (POOL) :**
  sa fonction est de réduire progressivement la taille spatiale de la représentation pour réduire le nombre de paramètre et de calcul dans le réseau, elle contrôle également le sur-apprentissage. Entre chaque deux couches de CONV il est préférable de mettre une couche pooling. Il y a plusieurs types de couches pooling : Max pooling (donne la valeur maximale des entrés), average pooling (donne la moyenne des entrés), L2-norm pooling [4].
* **Couche de correction (ReLU) :**
  Pour améliorer l'efficacité du traitement on intercale entre les couches de traitement une couche qui opère une fonction mathématique sur les signaux de sortie.

Elle renvoie 0 si l'entrée Z<0 et Z si Z >0[4].

* **Couche entièrement connectée(Fully Connected FC) :**
  Les neurones de cette couche ont des connexions complètes à toutes les activations de la couche précédentes [4].
* **Couche de perte (LOSS) :**
  Elle est normalement la dernière couche du réseau. Elle spécifie comment l'entrainement du réseau pénalise l'écart entre le signale prévue et réel. Les fonctions les plus utilisées softmax, la perte par entropie, la perte euclidienne[4].

Les réseaux de neurones récurrents(RNNs) :

Les couches de ce réseau sont des entités primitives qui permettent aux réseaux d'apprendre à partir d'une séquence d'entrées. Ils ont pour spécificité d'être composés de couches (couche d'entrée, couche(s) cachée(s) et couche de sortie) et d'apprendre des couples (d'entrées, sorties) comme les perceptrons. Ces réseaux reprennent à la fois l'idée de la propagation avant et de la rétro propagation de l'erreur (phénomène qui intervient lors de l'apprentissage), et l'idée de la récurrence de Hopfield (Hopfield, 1982). L'entrée globale du réseau se propage à la fois en avant dans le réseau de couche en couche, mais elle contient aussi une information sur le passé (Martinez, 2011) [12].

* **Long Short Term Memory (LSTM) :**
  LSTM sont un type spécial des RNNs, les RNNs oublient rapidement le passé (c'est ce que on appelle la disparition de gradient) la cellule LSTM est une adaptation de la couche récurrente qui permet aux signaux les plus anciens des couches profondes de se déplacer vers la cellule du présent.[12]
* **Gated Recurrent Unit (GRU) :**
  Introduits en 2014 [18]les GRU sont similaires à LSTM mais ils ont moins de paramètres. Ils ont aussi des unités fermées comme les LSTM qui contrôlent la flux d'informations à l'intérieur de l'unité mais sans des cellules de mémoire distinctes. Contrairement à LSTM, GRU n'a pas de porte de sortie, exposant ainsi son plein contenu[19].

1.4.3.4 Apprentissage des réseaux de neurones

L'apprentissage est une phase du développement d'un réseau de neurones durant laquelle le comportement du réseau est modifié jusqu'à l'obtention du comportement désiré. L'apprentissage neuronal fait appel à des exemples de comportement[6]. Deux modes d'apprentissage existent : l'apprentissage supervisé, et l'apprentissage non supervisé [20]

L'apprentissage supervisé :

Dans ce type d'apprentissage, les entrées et les sorties sont fournies au préalable. Ensuite, le

réseau traite les entrées et compare ses résultats aux sorties souhaitées. Les poids sont ensuite ajustés grâce aux erreurs propagées à travers le système. Ce processus se produit à plusieurs reprises tant que les poids sont continuellement améliorés. L'ensemble de données qui permet l'apprentissage est appelé l'ensemble d'apprentissage [20] [21].

L'apprentissage non supervisé :

Dans l'apprentissage non supervisé, le réseau est fourni avec des entrées mais pas avec les sorties souhaitées. Le système lui-même doit alors décider quelles fonctionnalités il utilisera pour regrouper les données d'entrée. C'est ce qu'on appelle souvent l'auto-organisation ou l'adaptation [20] [21].

1.5 Domaine d'application de l'apprentissage profond

L'apprentissage profond et dès son arrivé plusieurs chercheurs ont tenté de résoudre divers problèmes ou optimiser la solution de d'autre en l'utilisant comme approche , et cela dans beaucoup et différents domaines comme : l'intelligence artificielle, la sécurité, la médecine, le traitement automatique des données(données temporelles,données spatiales ,données spatio-temporelles) . . .

1.5.1 Dans le domaine du traitement automatique des langues naturelles(TALN)

Traduction automatique :

La Traduction Automatique(TA) est un processus par lequel des programmes informatiques sont utilisés pour traduire un texte d'une langue naturelle (langue source ou langue d'entrée) vers une autre langue naturelle (langue cible, langue de sortie). La sortie ne commence que lorsque nous avons vu l'entrée complète, car le premier mot de la phrase traduite nécessite des informations capturées à partir de la séquence d'entrée complète[16].

1.5.2 Dans le domaine de l'intelligence artificielle(IA)

Robots intelligents :

Les robots sont des machines programmables qui sont généralement capables d'effectuer une série d'actions de manière autonome ou semi-autonome. Les progrès de l'IA et DL feront évoluer les capacités des robots et les ont permis de détecter et de réagir à leur environnement[22].

Voiture autonome :

Les entreprises qui construisent de tels types de services d'aide à la conduite, ainsi que des

voitures autonomes telles que Google, doivent apprendre à un ordinateur à maîtriser certaines parties essentielles de la conduite à l'aide de systèmes de capteurs numériques au lieu de l'esprit humain. Pour ce faire, les entreprises commencent généralement par entraîner des algorithmes utilisant une grande quantité de données. Vous pouvez imaginer comment un enfant apprend grâce à des expériences constantes et à la réplication. Ces nouveaux services pourraient fournir des modèles commerciaux inattendus aux entreprises[23].

Chatbots (agents conversationnels) :

C'est un programme informatique capable de simuler une conversation en langage naturel avec un ou plusieurs humains par échange vocal ou textuel. Ils sont en quelque sorte des assistants virtuels dot´es d'une intelligence artificielle. Grâce au deep learning ils deviennent de plus en plus intelligents [24].

1.5.3 Dans le domaine de la médecine

Diagnostic médical :

L'intelligence artificielle et Le deep learning ont fait des progrès énormes dans l'interprétation de l'imagerie médicale, grâce à la disponibilité d'images de haute qualité et 'a la capacité des réseaux neuronaux convolutifs à classer les images l'IA et le DL sont désormais capable de réaliser un diagnostic médical avec autant, voire plus de précision qu'un humain[25].

1.5.4 Dans le domaine de la sécurité

Identification de pièces défectueuses :

La détection de défauts d'aspect et anomalies sur des objets en défilement à cadence élevée (10 pièces par seconde) demande la mise en place d'un système de contrôle robuste et rapide. Et on constate que les méthodes classiques de traitement d'images ne sont pas suffisamment efficaces. Alors L'utilisation du Deep Learning va aider 'à distinguer une pièce bonne d'une pièce défectueuse[26].

Détection de malwares ou de fraudes :

Le deep learning présente des résultats nettement plus performants en matière de détection de fraude lorsqu'ils sont aliment´es par un volume important de données . Des données sont collectées et fournies, Il peut s'agir de données comportementales, techniques ou d'informations issues de réseaux sociaux, leur permettant de « créer des profils clients », mis 'a jour régulièrement lesquels leur permet de prédire ou détecter une opération frauduleuse en se basant sur les comportements habituels, normaux du client afin de détecter un comportement anormal. Le malware est la contraction des termes anglais malicieux et software. Il désigne un

logiciel malveillant s'attaquant aux ordinateurs, terminaux mobiles et objets connectés. Les nouvelles techniques développées par le DL détectent rapidement les malwares[27].

1.6 Conclusion

Dans ce chapitre nous avons présentés ce que c'est le deep learning son principe ses différentes architectures et ses domaines d'application .

CHAPITRE 2

SPEECH EMOTION RECOGNITION (SER).

Sommaire

2.1 Introduction

Dans une interaction homme-homme la détection d'émotion est facile, elle est détectée soit par le visage et les expressions faciales, par les gestes du corps ou par la parole mais le défi c'est quand il s'agit d'une interaction homme-machine et que la machine soit capable de détecter l'émotion humaine. Afin d'améliorer cette interaction le terme SER apparait qui a comme objectif la reconnaissance d'émotion en utilisant uniquement l'intonation vocale.
La reconnaissance des émotions vocales (SER) est une tâche difficiles dans le domaine de l'analyse des signaux vocaux, c'est un problème de recherche qui tente de déduire l'émotion des signaux vocaux.

2.2 Les émotions et l'ère des sciences affectives

Dès l'antiquité, des philosophes tels qu'Aristote dans son ouvrage l'Ethique à Nicomaque se sont intéressés à l'étude des phénomènes affectifs en distinguant les affects (pathè) et le raisonnement logique (prohairèsis). Aristote note que l'émotion et la raison contribuent ensemble à la définition de l'être. Au XVIIème siècle d'autres auteurs tels que Descartes ou Spinoza s'intéressent également aux émotions. L'étude des états affectifs et plus particulièrement des émotions a connu un nouvel essor il y a 150 ans avec des auteurs comme (Darwin 1872) qui s'est intéressé aux formes d'expressions émotionnelles des hommes mais également des animaux. Darwin note que les émotions primaires sont universelles au niveau de l'expression faciale. Un siècle plus tard des chercheurs comme Scherer proposent un nouveau thème d'étude : les sciences affectives avec une série d'études portant sur l'expression émotionnelle vocale et/ou l'universalité d'expression et de perception des émotions chez des sujets de cultures différentes (Scherer et al. 1972 ; Ekman and Friesen 1975 ; Scherer 1981 ; Scherer 1984). Depuis le début des années 2000 et l'ouvrage « Affective Computing » (Picard 1997) le nombre de publications et conférences scientifiques dédiées aux états affectifs n'a cessé de croître. Le nombre de papiers acceptés dans les conférences autour du thème de la détection des émotions dans la voix a été multiplié par 100 par rapport à l'année 2000. L'édition de plusieurs ouvrages ont permis de synthétiser des avancées dans ce domaine (Sander and Scherer 2009 ; Scherer et al. 2010 ; Douglas-Cowie et al. 2011 ; Petta et al. 2011). La mise en place de challenges réunissant plusieurs laboratoires comme CEICES (Combining Efforts for Improving automatic. Classification of Emotional user States) mettant en jeu une collaboration entre plusieurs équipes du réseau HUMAINE (Batliner et al. 2006)a été l'occasion de constater les avancées réalisées au cours des dernières années. De nombreuses thèses sont soutenues chaque année sur ce thème (Kannan Venkataramanan 2019 ; Rory Beard 2018 ; Vladimir 2017 ; Mahdhaoui 2010), traduisant un enthousiasme grandissant de la part de la communauté scientifique. [28].

2.3 Définition de SER

La reconnaissance de l'émotion de la parole, abrégée en SER, est l'acte de tenter de reconnaître l'émotion humaine et les états affectifs à partir de la parole. Cela capitalise sur le fait que la voix reflète souvent l'émotion sous-jacente par le ton et la hauteur. C'est également le phénomène que les animaux comme les chiens et les chevaux utilisent pour comprendre l'émotion humaine. SER est difficile car les émotions sont subjectives et l'annotation audio est difficile. [29]

2.4 Les ambiguïtés du speech emotion recognition

L'étude des émotions en psychologie est caractérisée par un ensemble d'incertitudes qui sont reflétées par une absence de consensus entre théoriciens autour de certains éléments clés tels que[30] :

* *La définition de l'émotion* est à la fois importante et difficile car c'est un mot courant et un terme notoirement fluide. En fait, il existe différentes définitions de l'émotions dans la littérature scientifique. Dans le langage courant, l'émotion est toute expérience consciente relativement brève caractérisée par une activité mentale intense et un degré élevé de plaisir ou de déplaisir [30]. En psychologie, l'émotion est souvent définie comme un état de sensation complexe qui entraîne des changements physiques et psychologiques. Ces changements influencent la pensée et le comportement. C'est un terme qui est souvent confondu avec l'humeur (mood), l'attitudes interpersonnelles (interpersonal stances), attitudes et traits de personnalité affectifs.[30]
* *Différentes personnes classent différemment un discours émotionnel [31].*
* *Une autre difficulté qui concerne les propriétés temporelles des émotions* Souvent, la quasi-totalité de l'énoncé n'a pas d'émotion (le locuteur est dans un état neutre), mais l'émotivité n'est contenue que dans quelques mots ou phonèmes dans un énoncé [31].
* *le nombre et le nom des différentes catégories d'émotion* Ainsi, pour le nombre de catégories d'émotion, certaines écoles de pensée recensent 15 classes alors que d'autres proposent six classes d'émotions universelles (Big Six), et même pour ceux qui proposent six classes d'émotion, il n'existe pas d'entente sur la nature des six classes [32].

2.5 L'état de l'art

La reconnaissance des émotions a connu depuis le début des années 2000 de nombreux développements et de nombreux succès et pour atteindre les succès qu'il est aujourd'hui, ce domaine est passé par différentes périodes dont différentes approches ont vu le jour, au début

il y avait les approches classiques les HMM les GMM après un autre type d'approche apparait basé sur l'apprentissage dont le deep learning.

2.5.1 Les méthodes classiques utilisées pour des SER

Les méthodes statistiques(classiques) impliquent généralement l'utilisation de différents algorithmes d'apprentissage automatique supervisé dans lesquels un grand ensemble de données annotées est introduit dans les algorithmes pour que le système apprenne et prédise les types d'émotions appropriés[33]. L'un des défis à relever pour obtenir de bons résultats dans le processus de classification est la nécessité de disposer d'un ensemble d'entrainement suffisamment large[33]. Les algorithmes d'apprentissage automatique les plus couramment utilisés dans ce domaine il y'a : Support Vector Machines (SVM), Naive Bayes, Hidden Markov Model (HMM), Gaussian Mixture Model (GMM), Artificial Neuronal Network (ANN) HMM et GMM étaient le meilleur outil utilisé dans la reconnaissance des émotions vocales, sauf que les HMM présentaient plusieurs inconvénients ; telles que pour le configurer, un grand nombre de paramètres était nécessaire et une très grande quantité de données pour l'entraîner [34]. Pour augmenter les performances, ils ont mélangé ces deux techniques (GMM-HMM), et en effet ont donné de meilleures performances. Le seul inconvénient de ce modèle est l'indépendance ; en outre, GMM augmente la complexité du calcul[35]. Autre technique qui n'apparaît pas aussi efficace que le HMM mais donne de bons résultats qui est l'ANN (Artificial Neuronal Network), certains auteurs ont proposé de les associer au HMM mais le plus grand inconvénient qui apparaît est l'absence d'un schéma commun pour entrainer à la fois ANN et HMM[36].

les méthodes	**Accuracy**
SVM[37]	55.68 %
HMM[37]	64.77 %

Table 2.1: une comparaison entre les résultats d'un SVM et HMM [37].

2.5.2 Les limites des méthodes classiques

La parole est un signal complexe qui contient des informations sur : le message, le locuteur, la langue et les émotions. Il existe de nombreuses approches pour la reconnaissance automatique des émotions dans la parole : HMM, GMM, ANN, SVM, EM et bien d'autres.
La technique HMM a été largement utilisée par les chercheurs en raison de son efficacité même la combinaison HMM-GMM nous a offert plusieurs avantages mais il y a toujours des inconvénients, parmi eux ; la grande quantité de données à traiter, l'évaluation de l'expertise et un temps de calcul énorme.La limite des SVM est qu'il nécessite un temps énorme avec des données volumineuse et qu'il fait une classification binair[37].

Le plus grand inconvénient dans les méthodes classiques c'est qu'ils nécessitent un grand ensemble de données annotées est introduit pour que le système apprenne et prédise et pour obtenir de bons résultats. [33]

2.5.3 Les méthodes évoluées utilisées pour le SER (Deep Learning)

Avec l'avènement du deep learning, les résultats de la reconnaissance des émotions à partir de la parole se sont améliorés par rapport à ceux des méthodes classiques. Plusieurs études comparatives ont été menées entre les approches classiques et les approches plus évoluées parmi ces études il y'a [38] qui comparait entre le DNN (Deep Neuronal Network) et le SVM, dont le DNN était plus efficace. Un autre auteur a combiné le DNN-HMM et l'a comparé au GMM-HMM les résultats sont présentés dans le tableau suivant [4] et comme nous pouvons voir que le modèle DNN-HMM surpasse GMM-HMM [35].

les méthodes	Accuracy
GMM-HMM	42.22 %
DNN-HMM (DISCRIMINATIVE PRETRAINING, 6 HIDDEN LAYERS)	53.89 %

Table 2.2: Une comparaison entre GMM-HMM et DNN-HMM .

Une autre étude qui fait la comparaison entre GMM, GMM-DNN, GMM-ELM, DNN-ELMK et un humain les résultats de cette comparaison sont illustrés dans la figure ci-dessous :

Comme on peut bien le remarquer que l'humain a le plus la capacité de reconnaitre l'émotion par la parole mais aussi les combinaisons utilisées avec le DNN ont donné de bons résultats par rapport à celui du GMM.
La reconnaissance des émotions par la parole a largement évolué, après l'avènement du deep learning. Plusieurs auteurs l'ont testé avec différentes architectures, nous allons maintenant nous concentrer que sur ceux qui ont été testés sur le corpus RAVDESS. Parmi ces travaux, on retrouve le travail le plus récent [40], qui teste plusieurs architectures. Le tableau suivant [4], montre les résultats :

Caractéristiques	Architecture	Nb de Class	Validation Acc	Test Acc.
Log Mel Spectrogram	4 Layer 2D CNN	12	70.31%	65%
29Coefficients MFCC+Delta	1 Layer 2D CNN	12	56%	53%
Log Mel Spectrogram	3 Layer 3D CNN	12	66%	55%

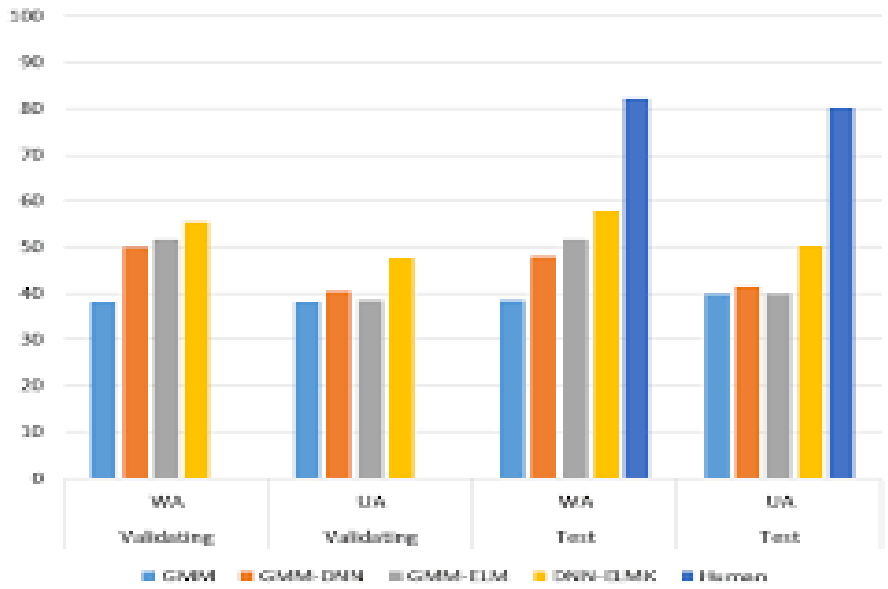

Figure 2.1: accuracy en% des méthodes proposées sur la reconnaissance des émotions de la parole.[39].

Table 2.3: les résultats de classification de l'émotion par la parole pour différentes architectures

Comme nous pouvons le voir, l'architecture qui donne la meilleure précision est celle du CNN 2D à 4 couches. Une autre recherche a été menée, qui a donné de bons résultats [41]. Le tableau 2.4 suivant montre ses résultats :

Features	**Architecture**	**Nb de Class**	**Accuracy**	**Epochs**
MFCC	CNN	8	42.22 %	500

Table 2.4: Précision de la classification des émotions vocales pour l'architecture CNN

L'auteur de l'article [42] a également travaillé avec l'architecture CNN. Le tableau 2.5 montre ses différents résultats

Features	**Architecture**	**Nb de Class**	**Accuracy**
COVAREP	LSTM	8	41.25 %

Table 2.5: Précision de la classification des émotions vocales pour les architectures LSTM

2.6 Conclusion

Dans ce chapitre nous avons parlé sur la reconnaissance des émotions par la parole nous avons vu comment ce domaine a fait son apparition, les différentes approches et techniques utilisé avant et maintenant et nous avons fait une comparaison entre ces techniques dont nous concluons que le deep learning nous apporte plus dans ce domaine.

CHAPITRE 3

CONTRIBUTION.

Sommaire

3.1 Introduction

Dans ce dernier chapitre nous allons présenter notre contribution dans le domaine de la reconnaissance des émotions à partir de la parole (SER), dans le chapitre précédent nous avons présenté les différentes techniques et approches qui ont été utilisé dans ce domaine dans notre recherche nous allons nous approfondir encore plus avec l'approche du deep learning dont nous allons tester plusieurs modèles qui n'ont pas été tester auparavant et nous allons présenter leurs résultats.

3.2 Présentation des outils de travail

3.2.1 Software

3.2.1.1 Jupyter Notebook

Un notebook, en programmation, permet de combiner des sections en langage naturel et des sections en langage de programmation dans un même document. Jupyter est une application web permettant de créer des notebooks. Jupyter permet de programmer en direct en langage python, langage Julia et le langage R. Le code s'y exécute. Le langage de balisage Markdown permet de commenter ce code en langage naturel [43].

3.2.1.2 Le Langage de Programmation Python

Python est un langage de programmation puissant et facile à apprendre. Il dispose de structures de données de haut niveau et permet une approche simple mais efficace de la programmation orientée objet. Parce que sa syntaxe est élégante, que son typage est dynamique et qu'il est interprété, Python est un langage idéal pour l'écriture de scripts et le développement rapide d'application dans de nombreux domaines et sur la plupart des plateformes[44].

3.2.1.3 TensorFlow

Créé par l'équipe Google Brain en 2011, sous la forme d'un système propriétaire dédié aux réseaux de neurones de Deep Learning, TensorFlow s'appelait à l'origine DistBelief. Par la suite, le code source de DistBelief a été modifié et cet outil est devenu une bibliothèque basée application. En2015, il a été renommé TensorFlow et Google l'a rendu open source. Depuis lors, il a subi plus de 21000 modification par la communication et est passé en version 1.0 en février 2017. Pour faire simple, TensorFlow est une bibliothèque de Machine Learning, il s'agit d'une boîte à outils permettant de résoudre des problèmes mathématiques extrêmement complexes avec aisance. Elle permet aux chercheurs de développer des architectures d'apprentissage expérimentales et de les transformer en logiciels[45].

3.2.1.4 Keras

Keras est un API de réseaux de neurones de haut niveau, écrite en Python et ineffaçable avec TensorFlow, CNTK et Theano. Elle a été développée pour permettre des expérimentations rapides. Les avantages de Keras :

* **Permet le prototypage rapide et facile** (de par sa convivialité, sa modularité et son extensibilité).
* **Supporte à la fois les réseaux convolutifs et les réseaux récurrents** ainsi que la combinaison des deux.
* **Fonctionne de façon transparente** sur CPU et GPU. [46]

3.2.2 Hardware

Le matériel utilisé est :

1.Un ordinateur personnel Lenovo I5 avec 8GB capacité mémoire(RAM), un processeur Intel(R) Core ™ i5-5300U 2.30 GHz 2.29 GHz, avec Windows 8, et un système de type 64 bit. Une carte graphique de type Intel(R) HD Graphics 5500.

2.Un ordinateur personnel ACER I5 avec 4GB capacité mémoire(RAM), un processeur Intel(R) Core ™ i5-4200U 1.60 GHz 2.30 GHz, avec Windows 8, et un système de type 64 bit. Une carte graphique de type AMD Radeon R5 M200 Series , Intel(R) HD Graphics Family.

3.3 Dataset

Pour démarrer un système de reconnaissance des émotions, trois éléments principaux doivent être pris en considération : le choix d'un ensemble de données approprié, la sélection des caractéristiques à partir des données audio et les classificateurs pour détecter les émotions[40].

Pour ce travail, nous choisissons de travailler avec les corpus RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song). Il contient 7356 fichiers (taille totale : 24,8 Go) et 24 acteurs professionnels (12 femmes, 12 hommes), vocalisant deux déclarations lexicalement appariées dans un accent nord-américain neutre. Chaque fichier audio dure 3 secondes et contient de la parole classée comme une émotion spécifique. Le format des fichiers audio est au format WAVE 16 bits, 48 kHz (.wav) [47]. Le discours comprend des expressions : calme, heureux, triste, en colère, craintives, surpris et de dégoût, et la chanson contient des émotions calmes, heureuses, tristes, en colère et craintives. Chaque expression est produite à deux niveaux d'intensité émotionnelle (normal, fort), avec une expression neutre supplémentaire. Toutes les conditions sont disponibles dans trois formats de modalité : audio uniquement (16 bits, 48 kHz .wav), audio-vidéo (720p H.264, AAC 48 kHz, .mp4) et vidéo uniquement (pas

de son). Remarque, il n'y a aucun fichier de morceau pour Actor_18 [48]. Chacun des fichiers 7356 RAVDESS a un nom de fichier unique. Le nom de fichier se compose d'un identifiant numérique en 7 parties (par exemple, 02-01-06-01-02-01-12.mp4). Ces identifiants définissent les caractéristiques du stimulus :

Modalité	01 = full-AV, 02 = vidéo uniquement, 03 = audio uniquement
Canal vocal	01 = discours, 02 = chanson
Émotion	01 = neutre, 02 = calme, 03 = heureux, 04 = triste, 05 = en colère, 06 = peureux, 07 = dégoût, 08 = surpris
Intensité émotionnelle	01 = normal, 02 = fort. REMARQUE : Il n'y a pas d'intensité forte pour l'émotion «neutre».
Déclaration	01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door"
Répétition	01 = 1ère répétition, 02 = 2ème répétition
Acteur	De 01 à 24. Les acteurs impairs sont des hommes, et les pairs sont des femmes

Table 3.1: Identificateurs de nom de fichier.[49]

Exemple de nom de fichier [49].

02-01-06-01-02-01-12.mp4 :
Vidéo uniquement (02).
Discours (01).
Peur (06).
Intensité normale (01).
Déclaration «chiens» (02).
1ère répétition (01).
12em acteur (12).
Femme, car le numéro d'identification de l'acteur est pair.

3.4 Extraction de caractéristiques (Feature Extraction)

Le défi avec les réseaux de neurones est de savoir comment gérer la taille des données de l'entrée d'origine, qui est souvent très importante en termes de mémoire. Par exemple, les fichiers image et audio ont souvent la taille de plusieurs Mo. Cela rend le processus de formation très coûteux, en termes d'allocation de mémoire et de nombre d'opérations de calcul nécessaires[47]. Avant de prendre les données brutes telles quelles, nous extrayons des caractéristiques spécifiques afin qu'elles soient des entrées pour le réseau neuronal, de sorte que

la taille sera réduite et donc le nombre d'opérations nécessaires pour entraîner le réseau au fur et à mesure pour augmenter les performances du modèle [50]. Dans ce mémoire, nous utilisons la méthode d'extraction de caractéristiques MFCC (The Mel-Frequency Cepstral Coefficients), qui est une approche de pointe pour l'extraction de caractéristiques vocales[51].

3.5 Implémentation et mise en œuvre

3.5.1 Prétraitement des fichiers audio

Notre corpus contient des fichiers audio dont nous n'allons pas les utilisés tels qu'ils sont ; le signal de son va devenir une image de pixels lesquelles nous allons traitées Pour résoudre les problèmes liés à la complexité de la parole, il est possible de calculer des coefficients représentatifs du signal traité. En simplifiant les choses, le signal de parole est transformé en une image de pixels. Ces pixels doivent représenter au mieux le signal, et extraire le maximum d'informations utiles pour la reconnaissance [52].

Les étapes de la transformation sont les suivant :[53]

* *Une forme d'onde acoustique est parlée dans le microphone.*
* *Une forme d'onde acoustique est parlée dans le microphone.*
* *Séparé par des énoncés de silence, le signal sonore est divisé en petits morceaux pour un traitement plus précis.*
* *Une conversion du morceau d'audio (du domaine temporel) en domaine fréquentiel qui va nous retourner une matrice des fréquences d'échantillonnage qui sera ensuite transformée en une séquence de N images carrées.*

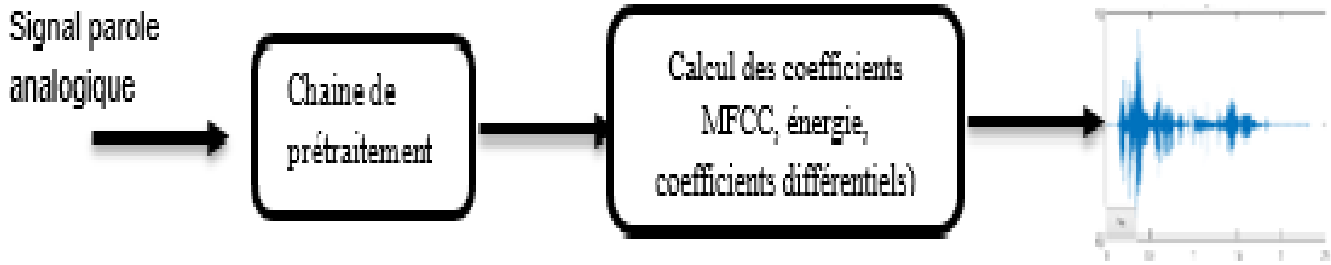

Figure 3.1: Phase de paramétrisation acoustique [52].

3.5.2 Les architectures

3.5.2.1 Modèle n°1

Pour la première architecture, nous avons proposé d'utiliser trois couches de convolutions conv1D, suivies d'une couche de correction (Relu) ; qui est utilisé pour améliorer l'efficacité du traitement ; c'est une couche qui opère une fonction mathématique sur les signaux de sortie, elle est insérée entre les couches de traitement. Suivi d'une couche de Batch normalisation qui est utilisée pour améliorer la vitesse, la performance et la stabilité des réseaux de neurones artificiels, la même chose pour la deuxième et la troisième couches conv1D, après cella nous avons mis deux couches entièrement connectées. La dernière couche est la couche de perte qui spécifie comment la formation réseau pénalise la différence entre le signal attendu et réel dans ce cas nous avons utilisé l'activation Softmax l'architecture est illustrée dans la figure ci-dessous :

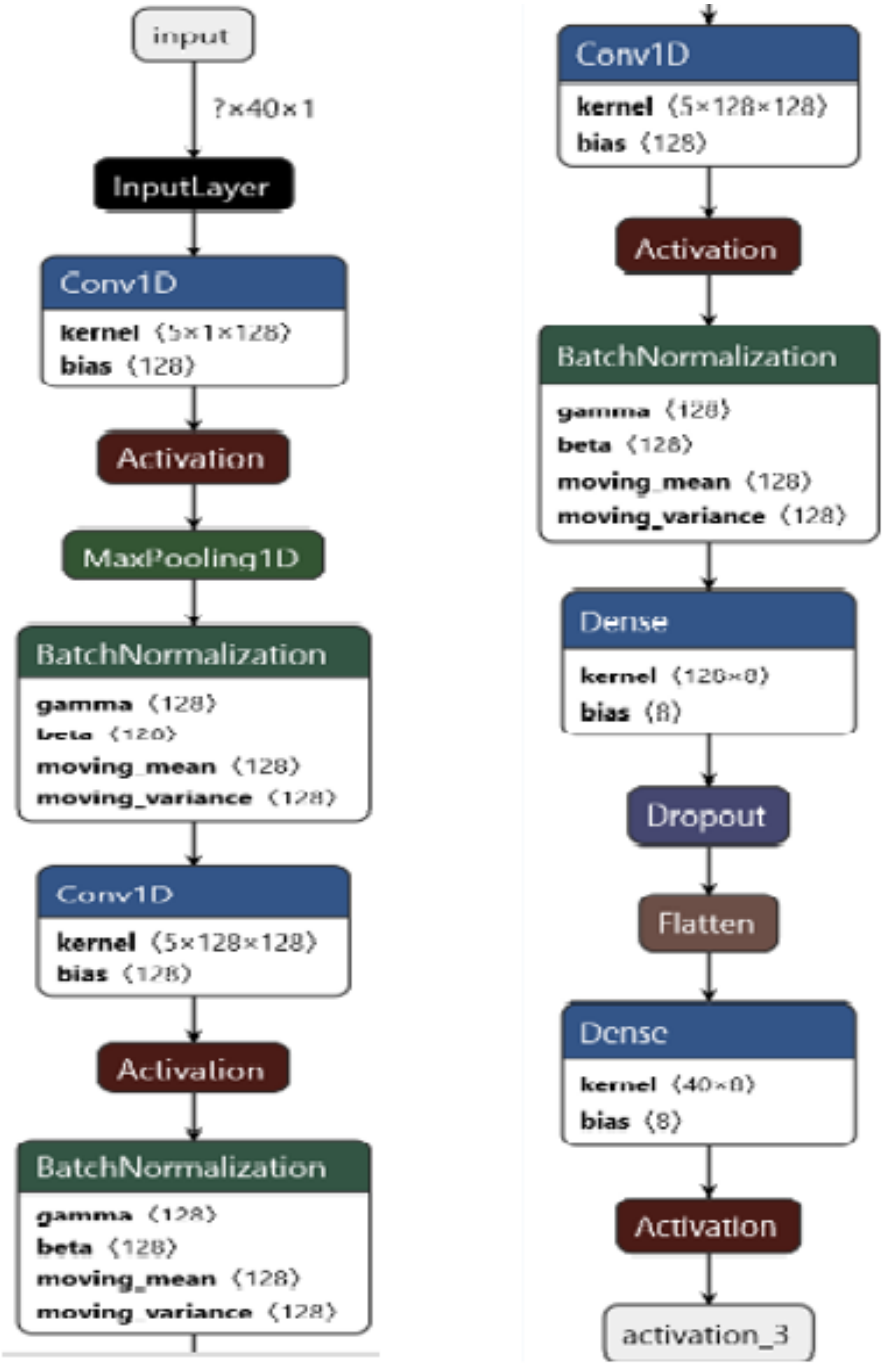

Figure 3.2: Architecture du model n°1.

Layer (type)	Output Shape	Param #
conv1d_2 (Conv1D)	(None, 40, 128)	768
activation_3 (Activation)	(None, 40, 128)	0
max_pooling1d_1 (MaxPooling1	(None, 5, 128)	0
batch_normalization_2 (Batch	(None, 5, 128)	512
conv1d_3 (Conv1D)	(None, 5, 128)	82048
activation_4 (Activation)	(None, 5, 128)	0
batch_normalization_3 (Batch	(None, 5, 128)	512
conv1d_4 (Conv1D)	(None, 5, 128)	82048
activation_5 (Activation)	(None, 5, 128)	0
batch_normalization_4 (Batch	(None, 5, 128)	512
dense_2 (Dense)	(None, 5, 8)	1032
dropout_1 (Dropout)	(None, 5, 8)	0
flatten_1 (Flatten)	(None, 40)	0
dense_3 (Dense)	(None, 8)	328
activation_6 (Activation)	(None, 8)	0

Total params: 167,760
Trainable params: 166,992
Non-trainable params: 768

Figure 3.3: le résume du modèle 1.

3.5.2.2 Modèle n°2

Dans ce deuxième modèle nous avons proposé d'utiliser 4 couches convolutives, suivi de couches de Batch normalisation qui sont utilisées pour améliorer la vitesse, la performance et la stabilité des réseaux de neurones artificiels, suivies de 2 couches entièrement connectées l'architecture est illustrée dans la figure ci-dessous :

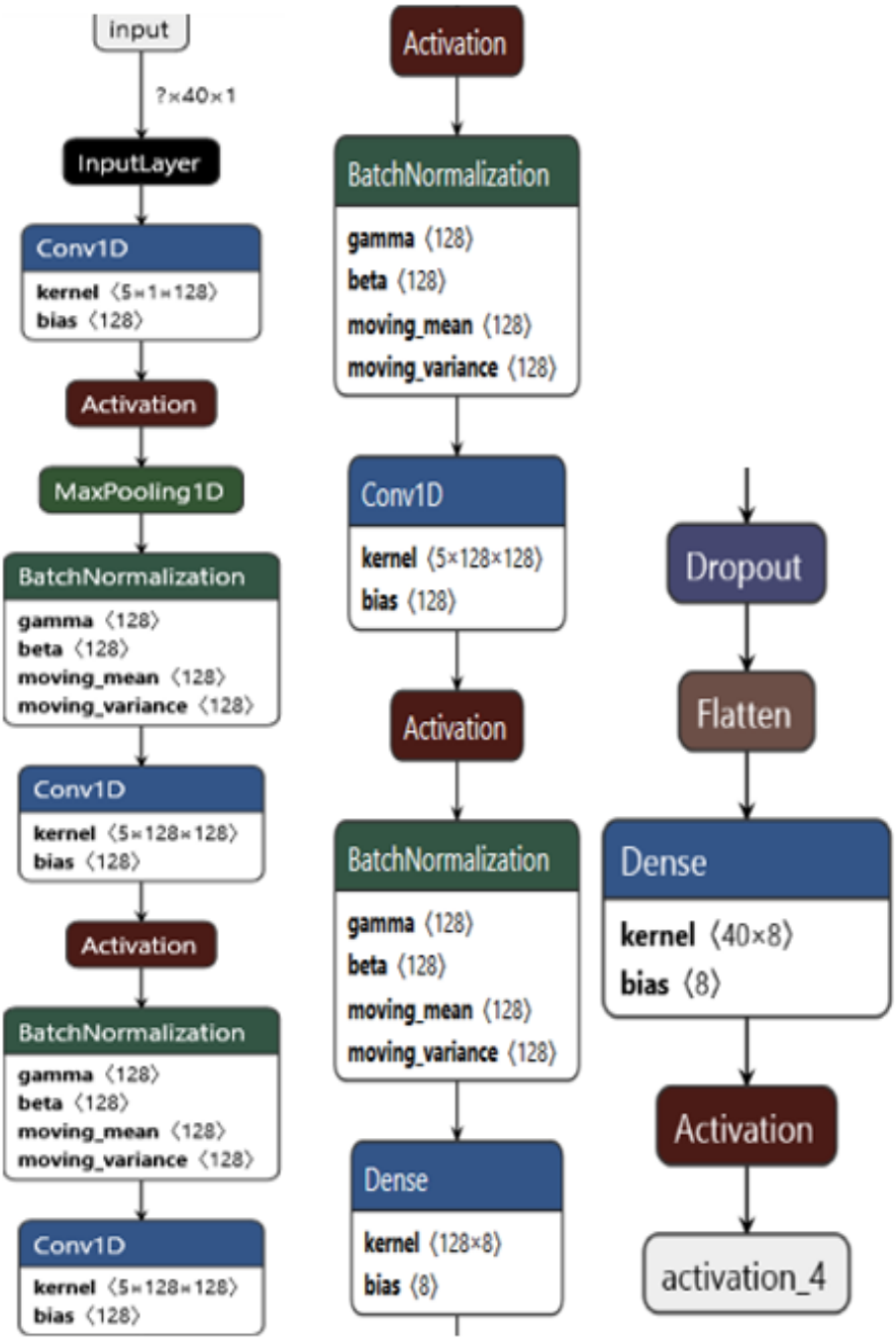

Figure 3.4: Architecture du model n°2.

Layer (type)	Output Shape	Param #
conv1d_30 (Conv1D)	(None, 40, 128)	768
activation_40 (Activation)	(None, 40, 128)	0
max_pooling1d_10 (MaxPooling	(None, 5, 128)	0
batch_normalization_30 (Batc	(None, 5, 128)	512
conv1d_31 (Conv1D)	(None, 5, 128)	82048
activation_41 (Activation)	(None, 5, 128)	0
batch_normalization_31 (Batc	(None, 5, 128)	512
conv1d_32 (Conv1D)	(None, 5, 128)	82048
activation_42 (Activation)	(None, 5, 128)	0
batch_normalization_32 (Batc	(None, 5, 128)	512
conv1d_33 (Conv1D)	(None, 5, 128)	82048
activation_43 (Activation)	(None, 5, 128)	0
batch_normalization_33 (Batc	(None, 5, 128)	512
dense_20 (Dense)	(None, 5, 8)	1032
dropout_10 (Dropout)	(None, 5, 8)	0
flatten_10 (Flatten)	(None, 40)	0
dense_21 (Dense)	(None, 8)	328
activation_44 (Activation)	(None, 8)	0

Total params: 250,320
Trainable params: 249,296
Non-trainable params: 1,024

Figure 3.5: le résumé du modèle 2.

3.5.2.3 Modèle n°3

Dans ce troisième modèle nous avons proposé d'utiliser 6 couches convolutives, suivi de couches de Batch normalisation qui sont utilisées pour améliorer la vitesse, la performance et la stabilité des réseaux de neurones artificiels, suivies de 2 couches entièrement connectées l'architecture est illustrée dans la figure ci-dessous :

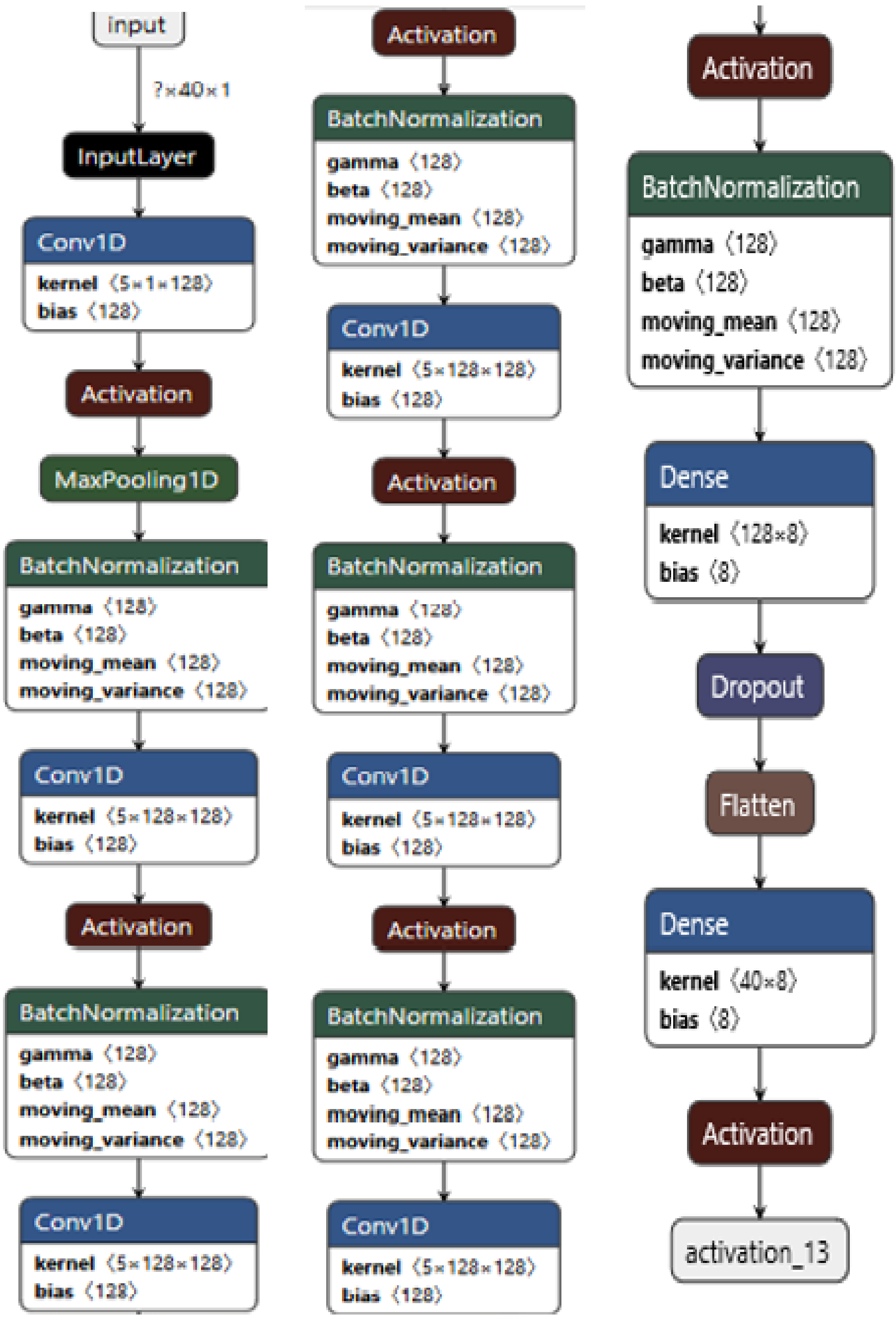

Figure 3.6: Architecture du model n° 3.

Layer (type)	Output Shape	Param #
conv1d_6 (Conv1D)	(None, 40, 128)	768
activation_7 (Activation)	(None, 40, 128)	0
max_pooling1d_1 (MaxPooling1	(None, 5, 128)	0
batch_normalization_6 (Batch	(None, 5, 128)	512
conv1d_7 (Conv1D)	(None, 5, 128)	82048
activation_8 (Activation)	(None, 5, 128)	0
batch_normalization_7 (Batch	(None, 5, 128)	512
conv1d_8 (Conv1D)	(None, 5, 128)	82048
activation_9 (Activation)	(None, 5, 128)	0
batch_normalization_8 (Batch	(None, 5, 128)	512
conv1d_9 (Conv1D)	(None, 5, 128)	82048
activation_10 (Activation)	(None, 5, 128)	0
batch_normalization_9 (Batch	(None, 5, 128)	512
conv1d_10 (Conv1D)	(None, 5, 128)	82048
activation_11 (Activation)	(None, 5, 128)	0
batch_normalization_10 (Batc	(None, 5, 128)	512
conv1d_11 (Conv1D)	(None, 5, 128)	82048
activation_12 (Activation)	(None, 5, 128)	0
batch_normalization_11 (Batc	(None, 5, 128)	512
dense_2 (Dense)	(None, 5, 8)	1032
dropout_1 (Dropout)	(None, 5, 8)	0
flatten_1 (Flatten)	(None, 40)	0
dense_3 (Dense)	(None, 8)	328
activation_13 (Activation)	(None, 8)	0

Total params: 415,440
Trainable params: 413,904
Non-trainable params: 1,536

Figure 3.7: le résumé du modèle 3.

3.5.2.4 Modèle n°4

Dans ce quatrième modèle nous avons proposé d'utiliser le réseau de neurone récurrent (RNN) de type GRU dont nous avons testé 4 couches GRU, suivi chacune d'entre elles de couches de Batch normalisation qui sont utilisées pour améliorer la vitesse, la performance et la stabilité des réseaux de neurones artificiels, suivies de 2 couches entièrement connectées l'architecture est illustrée dans la figure ci-dessous :

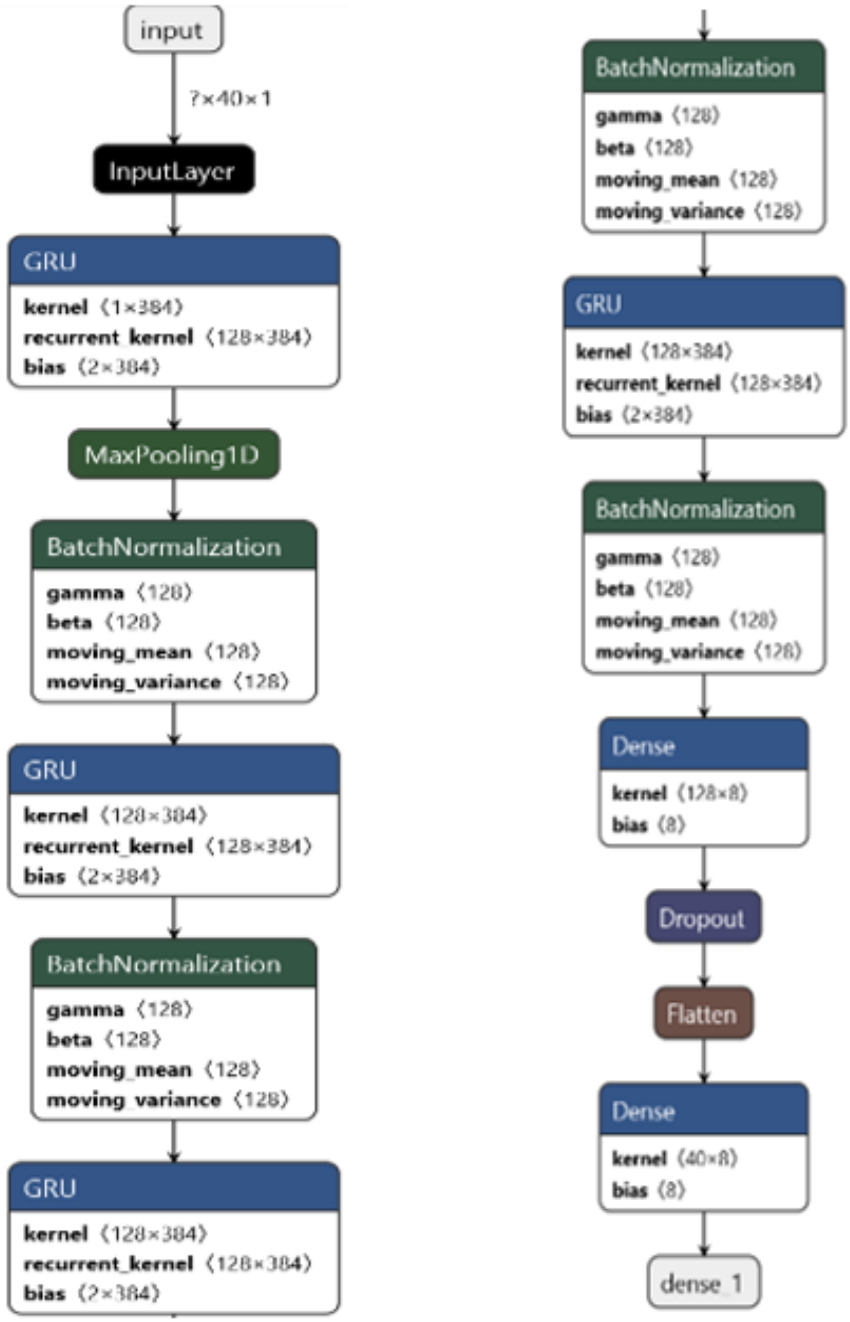

Figure 3.8: Architecture du model n°4.

Layer (type)	Output Shape	Param #
gru_12 (GRU)	(None, 40, 128)	50304
max_pooling1d_3 (MaxPooling1	(None, 5, 128)	0
batch_normalization_15 (Batc	(None, 5, 128)	512
gru_13 (GRU)	(None, 5, 128)	99072
batch_normalization_16 (Batc	(None, 5, 128)	512
gru_14 (GRU)	(None, 5, 128)	99072
batch_normalization_17 (Batc	(None, 5, 128)	512
gru_15 (GRU)	(None, 5, 128)	99072
batch_normalization_18 (Batc	(None, 5, 128)	512
dense_6 (Dense)	(None, 5, 8)	1032
dropout_3 (Dropout)	(None, 5, 8)	0
flatten_3 (Flatten)	(None, 40)	0
dense_7 (Dense)	(None, 8)	328

Total params: 350,928
Trainable params: 349,904
Non-trainable params: 1,024

Figure 3.9: le résumé du modèle 4.

3.5.2.5 Modèle n°5

La même chose pour ce cinquième modèle dont nous avons proposé d'utiliser le réseau de neurone récurrent (RNN) de type GRU dont nous avons testé 6 couches GRU, suivi chacune d'entre elles de couches de Batch normalisation qui sont utilisées pour améliorer la vitesse, la performance et la stabilité des réseaux de neurones artificiels, suivies de 2 couches entièrement connectées l'architecture est illustrée dans la figure ci-dessous :

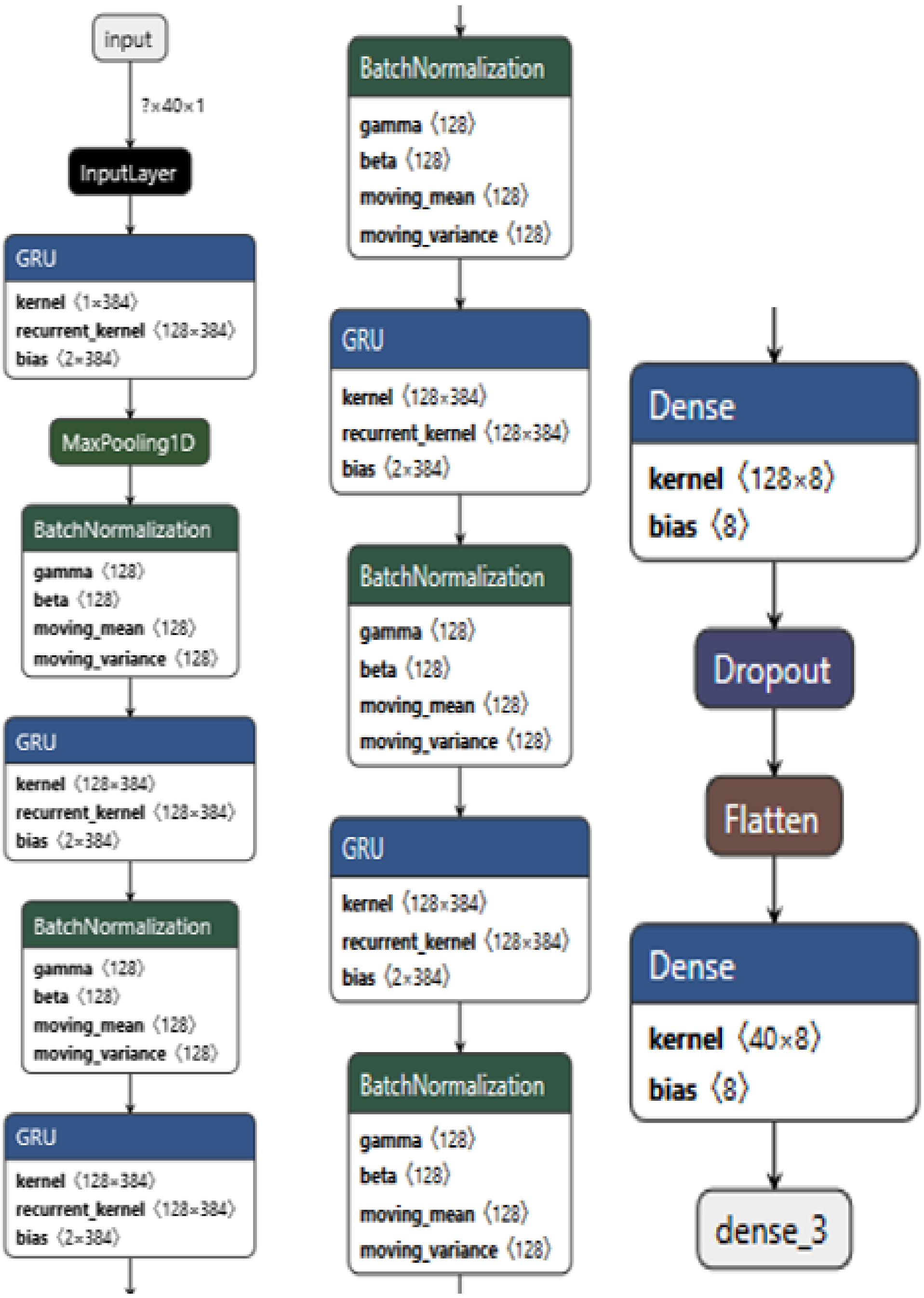

Figure 3.10: Architecture du model n° 5.

Layer (type)	Output Shape	Param #
gru_4 (GRU)	(None, 40, 128)	50304
max_pooling1d_1 (MaxPooling1	(None, 5, 128)	0
batch_normalization_4 (Batch	(None, 5, 128)	512
gru_5 (GRU)	(None, 5, 128)	99072
batch_normalization_5 (Batch	(None, 5, 128)	512
gru_6 (GRU)	(None, 5, 128)	99072
batch_normalization_6 (Batch	(None, 5, 128)	512
gru_7 (GRU)	(None, 5, 128)	99072
batch_normalization_7 (Batch	(None, 5, 128)	512
gru_8 (GRU)	(None, 5, 128)	99072
batch_normalization_8 (Batch	(None, 5, 128)	512
dense_2 (Dense)	(None, 5, 8)	1032
dropout_1 (Dropout)	(None, 5, 8)	0
flatten_1 (Flatten)	(None, 40)	0
dense_3 (Dense)	(None, 8)	328

Total params: 450,512
Trainable params: 449,232
Non-trainable params: 1,280

Figure 3.11: le résumé du modèle 5.

3.5.2.6 Modèle n°6

La même chose pour ce sixième modèle dont nous avons proposé d'utiliser le réseau de neurone récurrent (RNN) de type GRU dont nous avons testé 6 couches GRU, suivi chacune d'entre elles de couches de Batch normalisation qui sont utilisées pour améliorer la vitesse, la performance et la stabilité des réseaux de neurones artificiels, suivies de 2 couches entièrement connectées l'architecture est illustrée dans la figure ci-dessous :

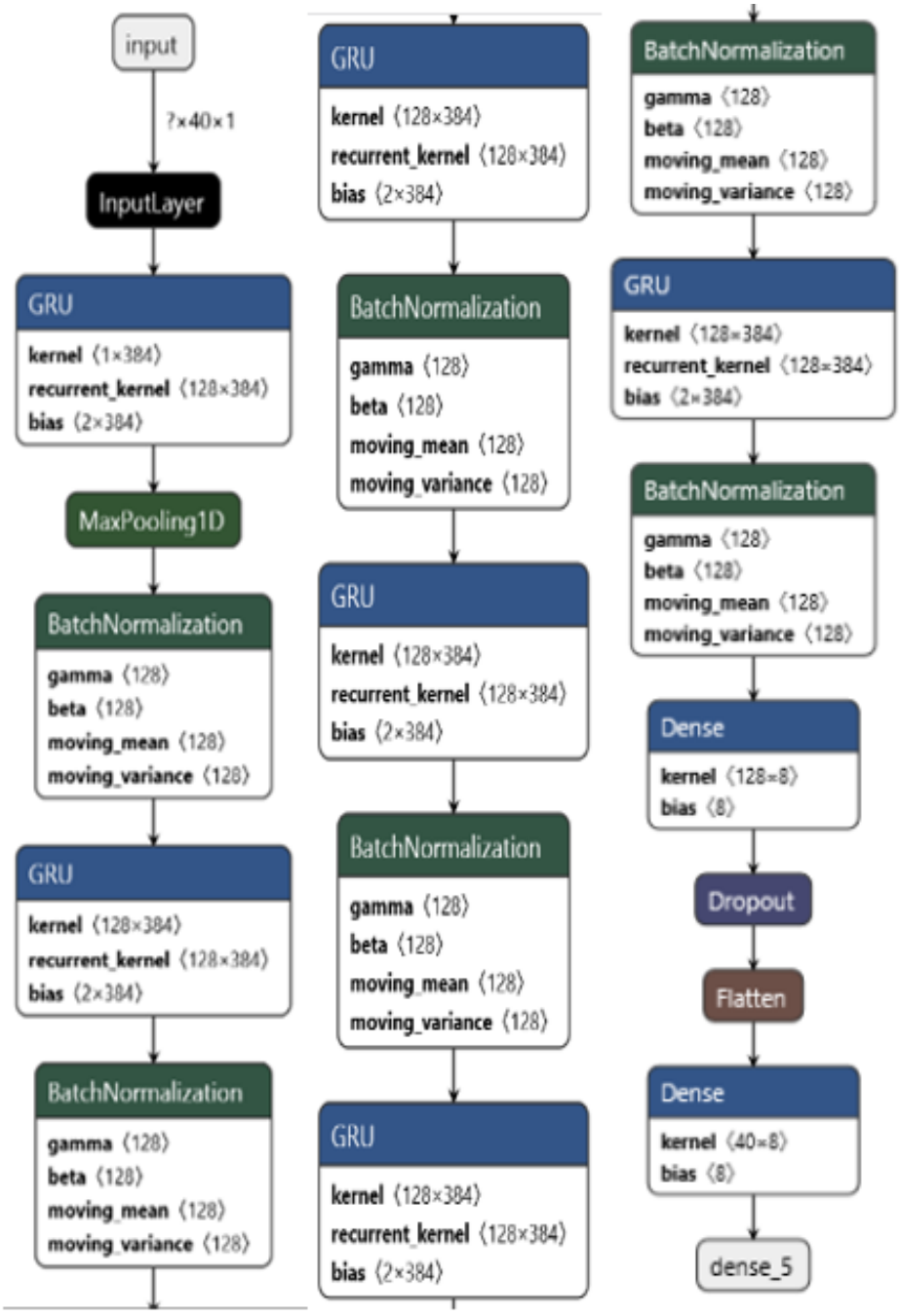

Figure 3.12: Architecture du model n° 6.

Layer (type)	Output Shape	Param #
gru (GRU)	(None, 40, 128)	50304
max_pooling1d (MaxPooling1D)	(None, 5, 128)	0
batch_normalization (BatchNo	(None, 5, 128)	512
gru_1 (GRU)	(None, 5, 128)	99072
batch_normalization_1 (Batch	(None, 5, 128)	512
gru_2 (GRU)	(None, 5, 128)	99072
batch_normalization_2 (Batch	(None, 5, 128)	512
gru_3 (GRU)	(None, 5, 128)	99072
batch_normalization_3 (Batch	(None, 5, 128)	512
gru_4 (GRU)	(None, 5, 128)	99072
batch_normalization_4 (Batch	(None, 5, 128)	512
gru_5 (GRU)	(None, 5, 128)	99072
batch_normalization_5 (Batch	(None, 5, 128)	512
dense (Dense)	(None, 5, 8)	1032
dropout (Dropout)	(None, 5, 8)	0
flatten (Flatten)	(None, 40)	0
dense_1 (Dense)	(None, 8)	328

Total params: 550,096
Trainable params: 548,560
Non-trainable params: 1,536

Figure 3.13: le résumé du modèle 6.

3.5.3 Résultat obtenus et discutions

3.5.3.1 Les métriques d'évaluation utilisés

L'accuracy : est la métrique la plus utilisé pour la mesure de la performance des modèles, mais n'est pas suffisante pour réellement évaluer un modèle[54].

$$Accuracy = \frac{Predictions.correctes}{Total.des.prdictions} \tag{3.1}$$

Matrice de confusion : Confusion Matrix ou tableau de contingence : est un résumé des résultats de prédictions sur un problème de classification. Les prédictions correctes et

incorrectes sont mises en lumière et réparties par classe. Les résultats sont ainsi comparés avec les valeurs réelles. Elle permet de comprendre de quelle façon le modèle de classification est confus lorsqu'il effectue des prédictions[54].

* **1. Modèle 1 :**

La valeur du loss et du val-loss diminue de plus en plus avec l'entrainement du modèle et l'augmentation du nombre d'epochs dont la valeur du loss arrive jusqu'à 0.0164 et val-loss à 0.3545 à l'epochs 1000. Par contre la valeur de l'accuracy augmente ce qui est normale et arrive jusqu'à 0.9934 et la valeur de val-acc arrive à 0.9430 à l'epochs 1000.

Les résultats sont bien illustrés dans les figures suivantes :

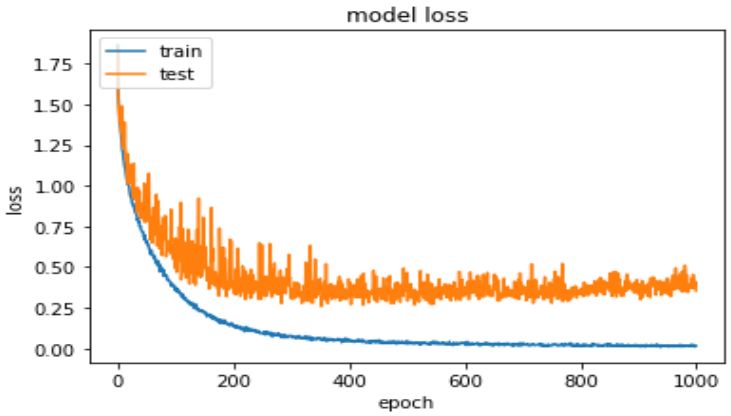

Figure 3.14: valeurs du Loss et Val-Loss du modèle1.

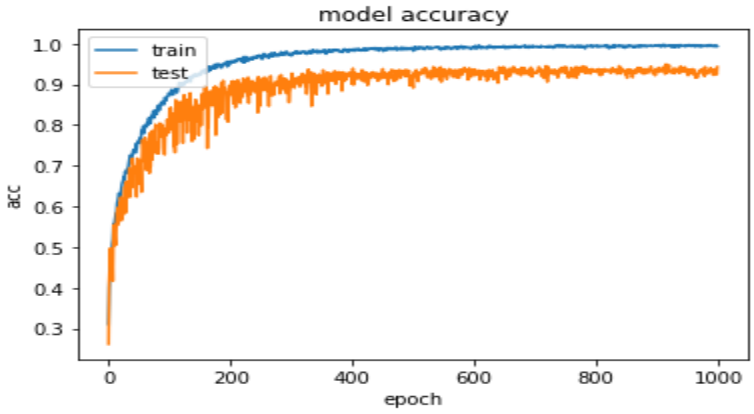

Figure 3.15: valeurs accuracy et Val-acc du modèle1.

On peut évaluer la performance d'un modèle par sa matrice de confusion aussi, la matrice de confusion du premier modèle est representée dans le tableu suivant :

	Neutral=0	Calm=1	Happy=2	Sad=3	Angry=4	Fearful=5	Disgust=6	Surprised=7
neutral=0	134	0	0	0	0	0	0	0
Calm=1	0	243	3	4	1	0	0	0
Happy=2	0	4	218	5	2	6	0	7
Sad=3	0	4	2	246	2	9	4	4
angry=4	0	2	6	0	243	0	0	2
fearful=5	0	0	2	4	0	233	0	0
Disgust=6	0	0	2	0	6	0	117	2
Surprised=7	0	2	6	0	0	0	2	106

Table 3.2: Matrice de confusion du modèle 1 .

* **2. Modèle 2 :**

Dans ce 2em modèle la valeur du loss et du val-loss diminue également avec l'entrainement du modèle dont la valeur du loss arrive jusqu'à 0.0046 et val-loss à 0.5667 à l'epochs 1000.

Par contre la valeur de l'accuracy augmente et arrive jusqu'à 0.9982 et la valeur de val-acc arrive à 0.9326 à l'epochs 1000. On remarque bien que l'accuracy de ce modèle a diminué par rapport au premier modèle (3CNN) ce qui signifie que le 1er modèle est meilleur que le 2em modèle.

Les résultats sont bien illustrés dans les figures suivantes :

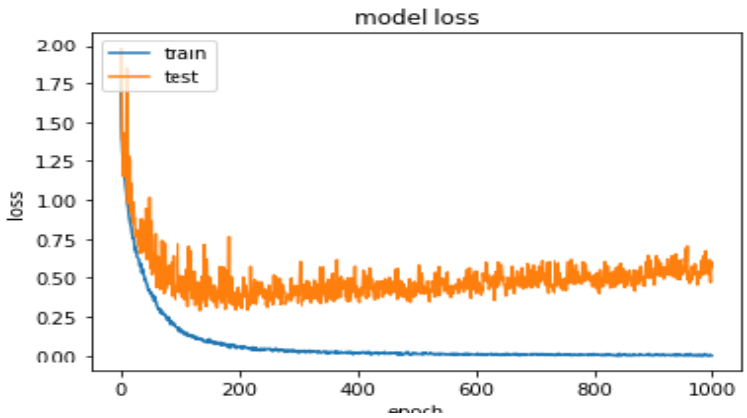

Figure 3.16: valeurs du Loss et Val-Loss du modèle2.

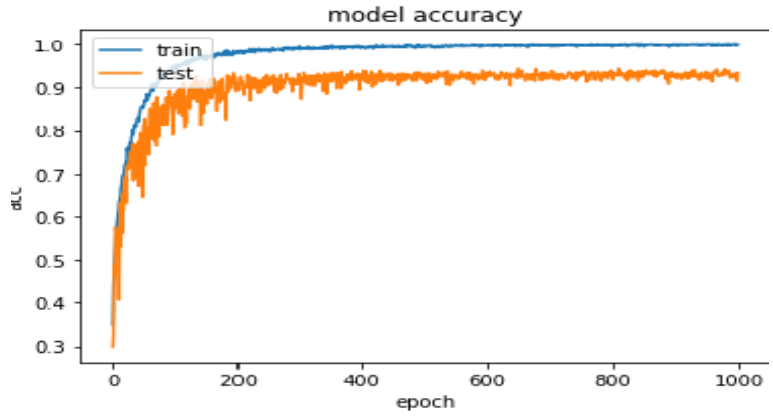

Figure 3.17: valeurs accuracy et Val-acc du modèle2.

la matrice de confusion du deuxieme modèle est representée dans le tableu suivant :

	Neutral=0	Calm=1	Happy=2	Sad=3	Angry=4	Fearful=5	Disgust=6	Surprised=7
neutral=0	129	0	1	2	0	2	0	0
Calm= 1	4	239	0	2	0	2	2	2
Happy=2	0	0	220	4	2	6	4	6
Sad=3	2	0	2	240	1	8	8	10
angry=4	2	2	2	0	243	2	0	2
fearful=5	0	0	0	10	0	229	0	0
Disgust=6	0	0	0	0	6	2	115	4
Surprised=7	0	0	2	0	0	4	2	108

Table 3.3: Matrice de confusion du modèle 2 .

* **3. Modèle 3 :**

Dans ce 3em modèle la valeur du loss et du val-loss diminue également avec l'entrainement du modèle dont la valeur du loss arrive jusqu'à 0.0060 et val-loss à 0.7400 à l'epochs 1000. Par contre la valeur de l'accuracy augmente et arrive jusqu'à 0.9985 et la valeur de val-acc arrive à 0.9228 à l'epochs 1000. Ce modèle est moins performant que le 1er modèle(3CNN) mais plus performant que le 2em modèle car il aboutit à de meilleurs résultats à une meilleure performance également.

Les résultats sont bien illustrés dans les figures suivantes :

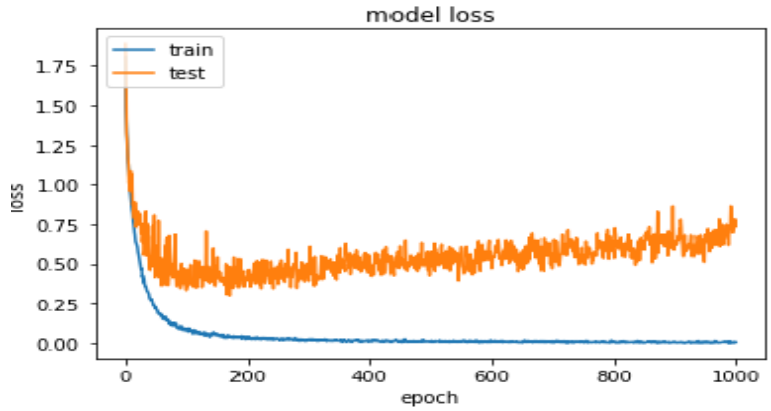

Figure 3.18: valeurs du Loss et Val-Loss du modèle3.

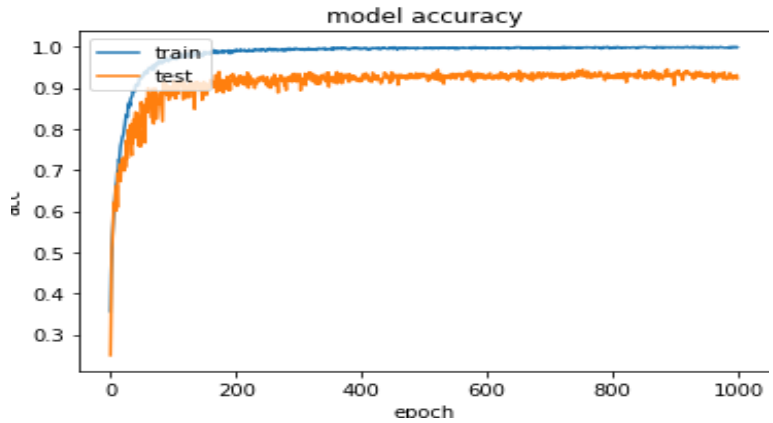

Figure 3.19: valeurs accuracy et Val-acc du modèle3.

la matrice de confusion du troisième modèle est representée dans le tableu suivant :

	Neutral=0	**Calm=1**	**Happy=2**	**Sad=3**	**Angry=4**	**Fearful=5**	**Disgust=6**	**Surprised=7**
neutral=0	129	4	0	1	0	0	0	0
Calm= 1	0	239	6	4	0	2	0	0
Happy=2	0	0	226	4	2	6	0	4
Sad=3	0	6	5	237	2	5	7	9
angry=4	0	0	4	0	245	0	2	2
fearful=5	0	0	4	17	2	216	0	0
Disgust=6	0	0	0	1	6	3	115	2
Surprised=7	0	2	4	0	0	6	4	100

Table 3.4: Matrice de confusion du modèle 3 .

* **4. Modèle 4 :**

Dans ce 4em modèle ou on utilise le RNN la valeur du loss et du val-loss diminue également avec l'entrainement du modèle dont la valeur du loss arrive jusqu'à 0.016 à et val-loss à 0.6665 à l'epochs 1000. Par contre la valeur de l'accuracy augmente et arrive jusqu'à 0.9955 et la valeur de val-acc arrive à 0.9290 à l'epochs 1000. Ce modèle est moins performant que le 1er (3CNN) et le 2em(4CNN) modèle mais plus performant que le 3em (6CNN) modèle.

Les résultats sont bien illustrés dans les figures suivantes :

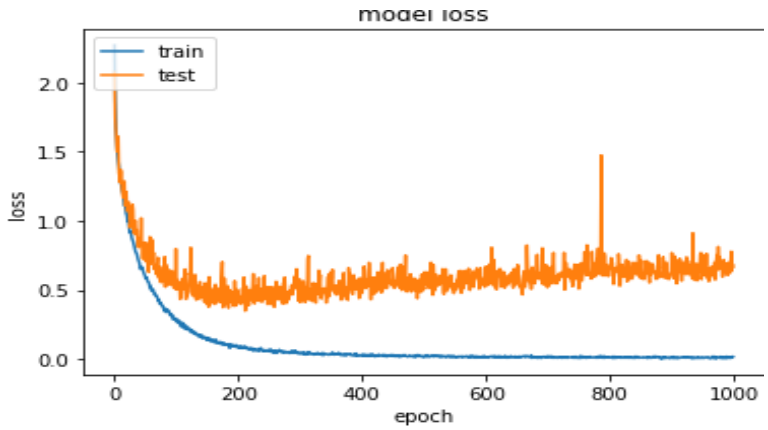

Figure 3.20: valeurs du Loss et Val-Loss du modèle4.

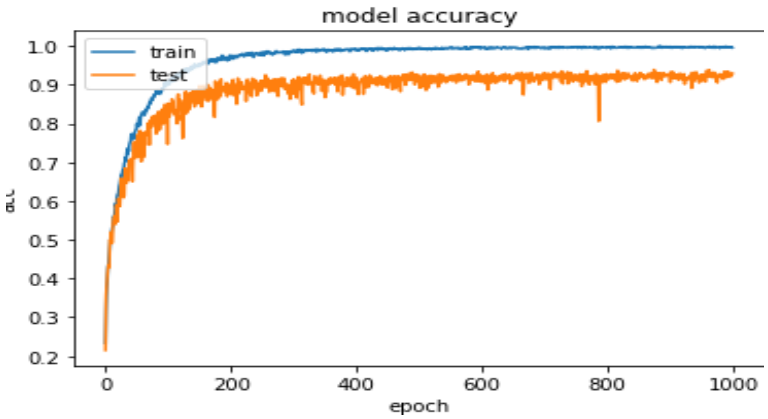

Figure 3.21: valeurs accuracy et Val-acc du modèle4.

la matrice de confusion du quatrième modèle est representée dans le tableu suivant :

	Neutral=0	Calm=1	Happy=2	Sad=3	Angry=4	Fearful=5	Disgust=6	Surprised=7
neutral=0	130	2	0	2	0	0	0	0
Calm= 1	0	243	3	1	0	2	2	0
Happy=2	0	1	216	0	4	14	0	7
Sad=3	6	3	0	232	1	11	4	14
angry=4	0	2	0	0	245	0	4	2
fearful=5	0	1	2	7	0	229	0	0
Disgust=6	0	0	5	4	3	2	112	1
Surprised=7	0	0	1	0	1	4	0	110

Table 3.5: Matrice de confusion du modèle 4.

* **5. Modèle 5 :**

Dans ce 5em modèle ou on a utilisé le RNN également et la valeur du loss et du val-loss diminue également avec l'entrainement du modèle dont la valeur du loss arrive jusqu'à 0.0174 et val-loss à 0.736 à l'epochs 1000. Par contre la valeur de l'accuracy augmente et arrive jusqu'à 0.9952 et la valeur de val-acc arrive à 0.9271 à l'epochs 1000. Ce modèle est moins performant que le 1er (3CNN) et le 2em(4CNN) et la 4em (4GRU) modèle mais plus performant que le 3em modèle(6CNN).

Les résultats sont bien illustrés dans les figures suivantes :

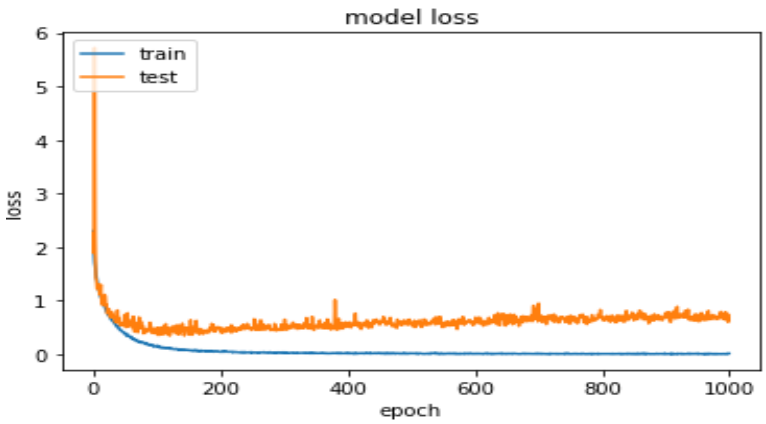

Figure 3.22: valeurs du Loss et Val-Loss du modèle5.

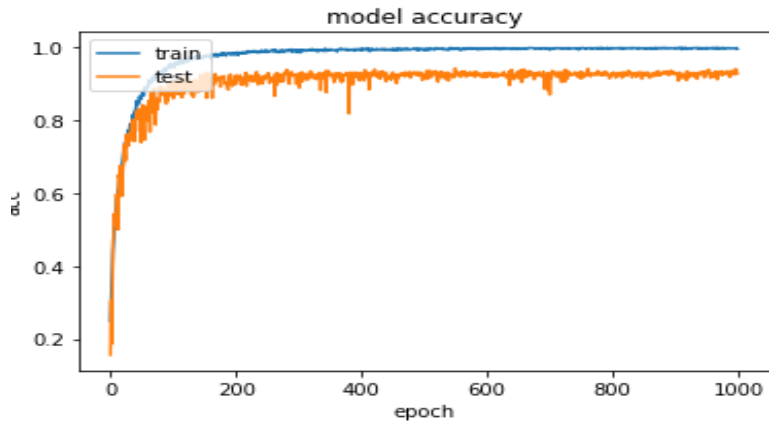

Figure 3.23: valeurs accuracy et Val-acc du modèle5.

la matrice de confusion du cinquième modèle est representée dans le tableu suivant :

	Neutral=0	Calm=1	Happy=2	Sad=3	Angry=4	Fearful=5	Disgust=6	Surprised=7
neutral=0	122	8	0	2	0	2	0	0
Calm= 1	0	245	0	6	0	0	0	0
Happy=2	4	3	213	2	2	9	4	5
Sad=3	2	6	0	246	4	4	4	5
angry=4	0	0	0	0	241	3	7	2
fearful=5	0	0	4	10	0	225	0	0
Disgust=6	0	0	0	0	6	4	115	2
Surprised=7	0	2	0	2	2	0	3	107

Table 3.6: Matrice de confusion du modèle 5 .

* **6. Modèle 6 :**

Dans ce 3em modèle la valeur du loss et du val-loss diminue également avec l'entrainement du modèle dont la valeur du loss arrive jusqu'à 0.0113 et val-loss à 0.695 à l'epochs 1000. Par contre la valeur de l'accuracy augmente et arrive jusqu'à 0.995 et la valeur de val-acc arrive à 0.9241 à l'epochs 1000. Ce modèle est plus performant que le 3em (6CNN) modèle mais toujours moins performant que le 1er(3CNN), le 2em(4CNN), le 3em(4GRU) et le 5em(5GRU) modèle.

Les résultats sont bien illustrés dans les figures suivantes :

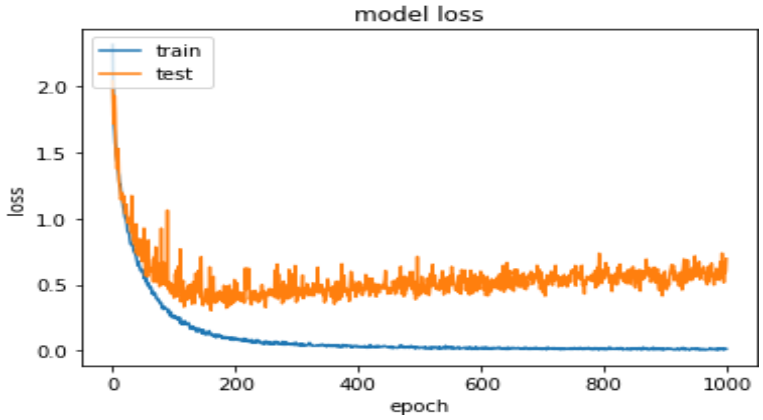

Figure 3.24: valeurs du Loss et Val-Loss du modèle6.

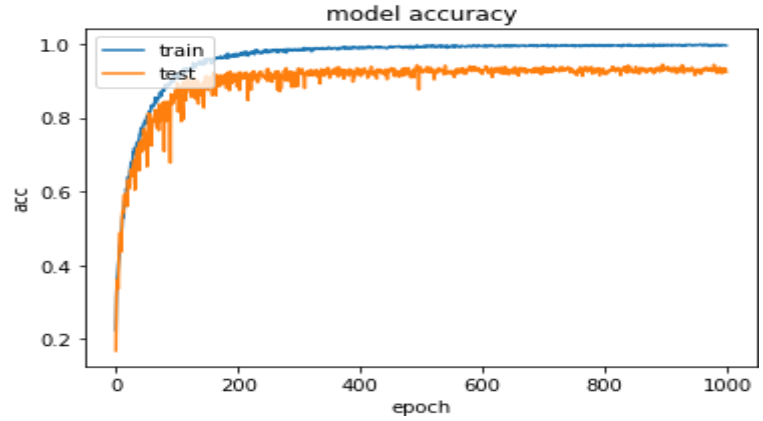

Figure 3.25: valeurs accuracy et Val-acc du modèle6.

la matrice de confusion du dernier modèle est representée dans le tableu suivant :

	Neutral=0	**Calm=1**	**Happy=2**	**Sad=3**	**Angry=4**	**Fearful=5**	**Disgust=6**	**Surprised=7**
neutral=0	128	2	0	2	0	2	0	0
Calm= 1	5	239	2	4	0	0	0	1
Happy=2	4	6	217	4	3	5	0	3
Sad=3	6	5	2	237	6	8	3	4
angry=4	0	2	4	0	245	0	2	0
fearful=5	0	0	6	10	0	223	0	0
Disgust=6	0	0	3	0	7	5	112	0
Surprised=7	2	2	0	0	2	0	2	108

Table 3.7: Matrice de confusion du modèle 6 .

Après avoir vu les résultats obtenus du loss et val-loss des 6 modèles on peut bien remarquer que à la fin de l'entrainement des modèles 1,2,3,4 et 6 ,ces modèle commencent à faire un sur-apprentissage (overfitting),si l'entrainement avait continué on aurait pu tomber dans un sur-apprentissage dont nos modèles auront pu être inefficaces et moins performants.

3.5.4 La comparaison de notre travail et d'autre travaux similaires

Le tableau suivant montre les performances de nos modèles, comparés à des travaux connexes récents qui utilisaient également le même corpus RAVDESS :

	Architecture	epochs	Training Accuracy	Test accuracy	Temps d'exécution
Travaux connexes	4 Layer 2D CNN [40].	100	70.31%	65%	-
	2 2D CNN+ MaxPooling 2D +Dropout+2 2D CNN+ MaxPooling 2D + Dropout+ flatten+ dense + Dropout+ dense [41].	500	90%	85%	-
	7 CNN + flatten +2 FC +softmax[42].	50	81%	79.5%	14 min
	LSTM [55].	-	-	41.25%	-
	BLSTM-CNN[56].	-	-	51.3%	-
Notre travail	3 1D CNN+ 2 FC	1000	99,34%	94.30%	30min
	4 1D CNN+ 2FC	1000	99,82%	93,26%	39min
	6 1D CNN+ 2FC	1000	99,85%	92,28%	40min10s
	4 GRU+ 2FC	1000	99,55%	92,90%	3h03s
	5 GRU+ 2FC	1000	99,52%	92,71%	3h10min
	6GRU + 2FC	1000	99,58%	92,41%	4h07min29s

Table 3.8: Une comparaison entre les précisions des travaux connexes et nos travaux

Pour bien évaluer nos modèles, nous les avons comparés avec des travaux similaires mais qui sont très récemment faites proposant le même corpus RAVDESS. Dans ces travaux les auteurs ont utilisé la même métrique d'évaluation que nous « l'accuracy ». Grâce à notre

expérience, et comme nous pouvons le voire dans le tableau ci-dessus nous avons pu augmenter la performance, cela est vrai que le nombre d'epochs de nos modèles est plus grand mais n'empêche que les résultats obtenus sont beaucoup meilleurs. Nous pouvons dire que nos modèles ont pu obtenir des résultats satisfaisants par rapport aux modèles des travaux connexes ; en question de temps d'exécution et de performance.

3.6 Quelques captures des résultats d'entrainement .

* **1. Modèle 1 :**

```
- val_accuracy: 0.9363
Epoch 995/1000
208/208 [==============================] - 2s 8ms/step - loss: 0.0141 - accuracy: 0.9946 - val_loss: 0.4543
- val_accuracy: 0.9228
Epoch 996/1000
208/208 [==============================] - 2s 7ms/step - loss: 0.0229 - accuracy: 0.9928 - val_loss: 0.4244
- val_accuracy: 0.9271
Epoch 997/1000
208/208 [==============================] - 2s 8ms/step - loss: 0.0163 - accuracy: 0.9931 - val_loss: 0.3689
- val_accuracy: 0.9375
Epoch 998/1000
208/208 [==============================] - 2s 8ms/step - loss: 0.0099 - accuracy: 0.9964 - val_loss: 0.3901
- val_accuracy: 0.9388
Epoch 999/1000
208/208 [==============================] - 1s 7ms/step - loss: 0.0150 - accuracy: 0.9946 - val_loss: 0.4114
- val_accuracy: 0.9241
Epoch 1000/1000
208/208 [==============================] - 2s 7ms/step - loss: 0.0164 - accuracy: 0.9934 - val_loss: 0.3545
- val_accuracy: 0.9430
```

Figure 3.26: Capture du résultat d'entrainement du Modèle1.

* **2. Modèle 2 :**

```
- val_accuracy: 0.9277
Epoch 995/1000
208/208 [==============================] - 3s 13ms/step - loss: 0.0032 - accuracy: 0.9991 - val_loss: 0.5480
- val_accuracy: 0.9314
Epoch 996/1000
208/208 [==============================] - 3s 15ms/step - loss: 0.0024 - accuracy: 0.9991 - val_loss: 0.6124
- val_accuracy: 0.9216
Epoch 997/1000
208/208 [==============================] - 3s 17ms/step - loss: 0.0111 - accuracy: 0.9961 - val_loss: 0.5241
- val_accuracy: 0.9326
Epoch 998/1000
208/208 [==============================] - 3s 15ms/step - loss: 0.0058 - accuracy: 0.9982 - val_loss: 0.4690
- val_accuracy: 0.9326
Epoch 999/1000
208/208 [==============================] - 3s 14ms/step - loss: 0.0056 - accuracy: 0.9988 - val_loss: 0.6012
- val_accuracy: 0.9149
Epoch 1000/1000
208/208 [==============================] - 3s 14ms/step - loss: 0.0046 - accuracy: 0.9982 - val_loss: 0.5667
- val_accuracy: 0.9326
```

Figure 3.27: Capture du résultat d'entrainement du Modèle2.

∗ **3. Modèle 3 :**

```
- val_accuracy: 0.9241
Epoch 995/1000
208/208 [==============================] - 3s 14ms/step - loss: 0.0048 - accuracy: 0.9979 - val_loss: 0.6985
- val_accuracy: 0.9302
Epoch 996/1000
208/208 [==============================] - 3s 13ms/step - loss: 0.0018 - accuracy: 0.9997 - val_loss: 0.7205
- val_accuracy: 0.9296
Epoch 997/1000
208/208 [==============================] - 3s 13ms/step - loss: 0.0055 - accuracy: 0.9988 - val_loss: 0.7665
- val_accuracy: 0.9290
Epoch 998/1000
208/208 [==============================] - 3s 13ms/step - loss: 0.0066 - accuracy: 0.9985 - val_loss: 0.7335
- val_accuracy: 0.9222
Epoch 999/1000
208/208 [==============================] - 3s 13ms/step - loss: 0.0025 - accuracy: 0.9997 - val_loss: 0.7843
- val_accuracy: 0.9314
Epoch 1000/1000
208/208 [==============================] - 3s 14ms/step - loss: 0.0060 - accuracy: 0.9985 - val_loss: 0.7400
- val_accuracy: 0.9228
```

Figure 3.28: Capture du résultat d'entrainement du Modèle3.

∗ **4. Modèle 4 :**

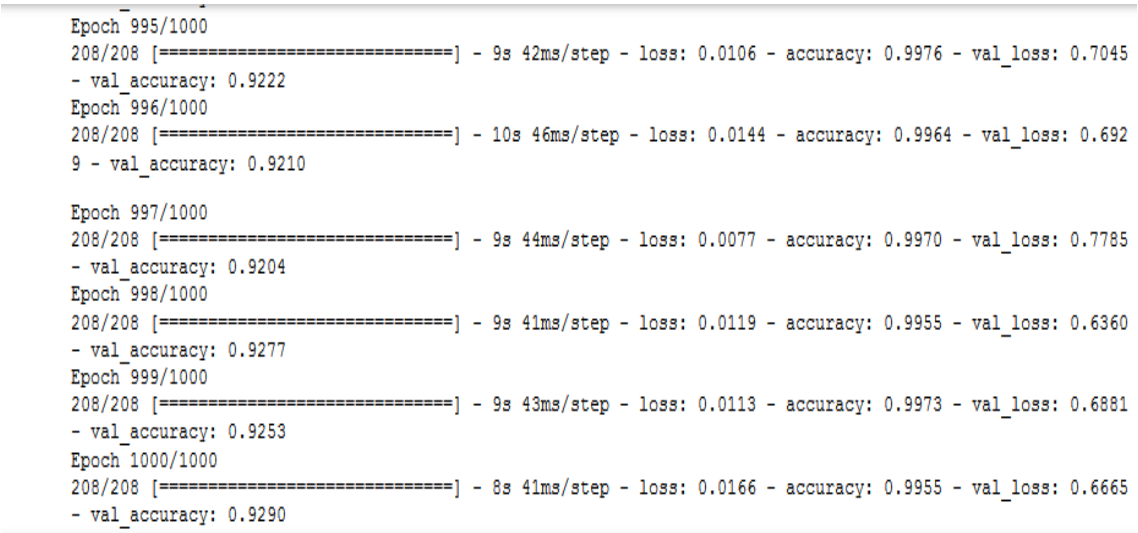

Figure 3.29: Capture du résultat d'entrainement du Modèle4.

* **5. Modèle 5 :**

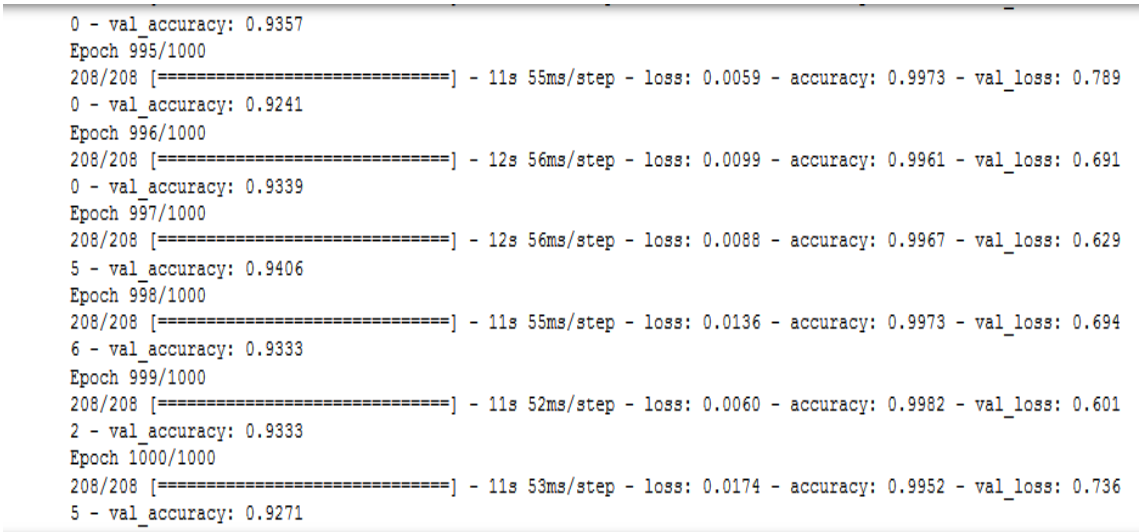

Figure 3.30: Capture du résultat d'entrainement du Modèle5.

* **6. Modèle 6 :**

```
5 - val_accuracy: 0.9265
Epoch 995/1000
208/208 [==============================] - 10s 50ms/step - loss: 0.0058 - accuracy: 0.9976 - val_loss: 0.664
4 - val_accuracy: 0.9241
Epoch 996/1000
208/208 [==============================] - 10s 50ms/step - loss: 0.0129 - accuracy: 0.9970 - val_loss: 0.514
9 - val_accuracy: 0.9345
Epoch 997/1000
208/208 [==============================] - 10s 48ms/step - loss: 0.0058 - accuracy: 0.9973 - val_loss: 0.605
5 - val_accuracy: 0.9345
Epoch 998/1000
208/208 [==============================] - 11s 52ms/step - loss: 0.0143 - accuracy: 0.9970 - val_loss: 0.586
8 - val_accuracy: 0.9290
Epoch 999/1000
208/208 [==============================] - 11s 54ms/step - loss: 0.0105 - accuracy: 0.9958 - val_loss: 0.597
6 - val_accuracy: 0.9320
Epoch 1000/1000
208/208 [==============================] - 10s 50ms/step - loss: 0.0113 - accuracy: 0.9958 - val_loss: 0.695
9 - val_accuracy: 0.9241
```

Figure 3.31: Capture du résultat d'entrainement du Modèle6.

3.7 Conclusion

Dans ce chapitre nous avons donné les outils nécessaires pour la réalisation de ce travail. Nous avons aussi présenté l'environnement de développement, à la fin nous avons montré les résultats obtenus ainsi que quelques captures d'écran qui explique le déroulement de l'entrainement de nos modèles.

Après avoir tester les 6 modèle on peut déduire que :

les modèles avec CNN sont les meilleurs modèles en particulier le premier modèle (3CNN) du coté performance et temps d'exécution pour la reconnaissance des émotions à partir de la parole.

CONCLUSION GENERALE

Ce mémoire avait pour ambition d'exploré un domaine qui est un champ de recherche qui passionne beaucoup de chercheurs, et c'est la reconnaissance des émotions à partir de la parole.

Pour réaliser ce travail on a utilisé le deep learning comme technique, ce choix de technique est justifié par la simplicité et l'efficacité de ses méthodes.

Il a fallu dans un premier temps définir la notion même d'apprentissage profond et savoir comment l'appliquer à notre thème de recherche. Bien que la performance des activités réalisées dans le domaine de la reconnaissance d'émotions à partir de la parole soient nombreuses, aucune méthode n'est jugée performante à 100%, mais au fur et à mesure les nouveaux travaux essayent d'améliorer la performance pour de meilleurs résultats.

Les résultats que nous avons obtenus confirment l'efficacité de notre approche. Ce travail reste ouvert pour de nouvelle améliorations dont on peut citer comme perspectives dans ce domaine pour l'avenir :

- ☐ Tester nos modèles sur d'autres ensembles de donnés (FAU AIBO Emotion ,Emo-DB ,TESS...).
- ☐ Tester notre thème avec des architectures encore plus profondes.
- ☐ Tester nos modèles sur de grandes quantité de données.
- ☐ Tester l'hybridation de différentes architectures comme (CNN-GRU) (CNN-LSTM)

BIBLIOGRAPHIE

[1] R, R. L., "Différence entre intelligence artificielle Machine learning et Deep learning," Décembre 2018.

[2] Chartrand, G., Cheng, P. P. M., MD, Vorontsov, M. E., Drozdzal, B. E. S. M., Turcotte, P. S., Pal, M. C. J., Kadoury, P. S., Tang, P. A., and MSc, "Deep Learning : A Primer for Radiologists," November 2017.

[3] "Intelligence artificielle machine learning deep learning kézako," https :// www.ledigitalab.com/ fabrique/intelligence -artificielle-machine-learning-deep-learning-kezako/ Consulté le 25/03/2020 à 02 :06.

[4] Boughaba Mohammed, B. B., *L'apprentissage profond (Deep Learning) pour la classification et la recherche d'images par le contenu*, Master, UNIVERSITE KASDI MERBAH OUARGLA, 2017.

[5] Bial, R., "Data Scientist comprendre le machine learning et le deep learning," https ://www.bial-r.com/2019/05/22/comprendre-le-machine-learning-et-le-deep-learning/ consulté le 25/03/2020 à 00 :57.

[6] Touzet, C., *LES RESEAUX DE NEURONES ARTIFICIELS INTRODUCTION AU CONNEXIONNISME*, Master's thesis, 27 Jun 2016.

[7] Chtita, S., *Modélisation de molécules organiques hétérocycliques biologiquement actives par des méthodes QSAR/QSPR Recherche de nouveaux médicaments*, Theses, casablanca, July 2017.

[8] "Bastien : Introduction aux réseaux de neurones – 2/3 : Neurone biologique et neurone formel. Nov 15." consulté sur le site https ://blog.clevy.io/nlp-et-ia/introduction-aux-reseaux-de-neurones-2-3-neurone-biologique-et-neurone-formel/.

[9] Djeriri, "Les Réseaux de Neurones Artificiels," Septembre 2017.

[10] Ahmed, H. M., *Commande de la machine asynchrone à double alimentation – apport des techniques de l'intelligence artificielle*, Theses, University of Sidi-Bel-Abbès, Jul 2017.

[11] ALI, L., *SÉLECTION DES MOTS CLÉS BASÉE SUR LA CLASSIFICATION ET L'EXTRACTION DES RÈGLES D'ASSOCIATION*, Master, MÉMOIRE PRÉSENTÉ À L'UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES, JUIN 2017.

[12] Kaadoud, I. C., "Architecture des réseaux de neurones : Réseaux de neurones artificiels classiques(2/3)," Novembre 2018.

[13] Boes, J., *Apprentissage du contrôle de systèmes complexes par l'auto-organisation coopérative d'un système multiagent Application à la calibration de moteurs à combustion*, Theses, Universite de Toulouse III – Paul Sabatier.

[14] Mohamed, K. B., *Réseau de Kohonen les Carte Auto-Organisatrices*, Theses, Université des Sciences et de la Technologie d'Oran USTOM B.

[15] Dejasmin, J., "Les réseaux de neurones convolutifs . Avril 17, 2018," Récupérer sur le site https ://www.natural-solutions.eu/blog/la-reconnaissance-dimage-avec-les-rseaux-de-neurones-convolutifs .

[16] Youcef, M. D., *Deep Learning pour la classification des images*, Master, Université Abou Bakr Belkaid Tlemcen, 2017.

[17] Saha, S., "A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way," Dec 15 2018.

[18] Bengio, J. C. C. G. K. H. Y., "Empirical Evaluation of Gated Recurrent Neural Networkson Sequence Modeling," .

[19] Shewalkar, A., Nyavanandi, D., and Ludwig, S. A., "PERFORMANCE EVALUATION OF DEEP NEURAL NETWORKS APPLIED TO SPEECH RECOGNITION RNN, LSTM AND GRU," March 2019.

[20] k, G., *an introduction to neural networks*, 1997.

[21] l v, F., *fundamentals of neural networks architectures algorithms and applications*, 1993.

[22] "Mais c'est quoi un robot ?." Consulté sur le site : https ://robogenie.fr/avantages-apprentissage-robotique/robot-definition/.

[23] Abdelkader, M., *L'ANALYSE DU SENTIMENT UTILISANT LE DEEP LEARNING*, Master, Université Dr TAHAR MOULAY SAIDA, 2019.

[24] "Futura Tech : Chatbot." Récupéré sur le site : https ://www.futurasciences.com/tech/definitions/internet-chatbot-15778/.

[25] Cimino, V., " : L'IA peut réaliser un diagnostic médical avec plus de précision qu'un humain . Publié le 25 septembre 2019 à 14h44 ." Récupéré sur le

site : https ://siecledigital.fr/2019/09/25/lia-peut-realiser-un-diagnostic-medical-avec-plus-de-precision-quun-humain/.

[26] S. Bouillant, J. Mitéran, M. P., "DÉTECTION DE DÉFAUTS TEMPS RÉEL SUR DES OBJETS A GÉOMÉTRIE COMPLEXE : ÉTUDE PAR SVM, BOOSTING ET HYPER-RECTANGLES," .

[27] KOSSI, K. K., "L'intelligence artificielle face à la fraude bancaire. 28 mars 2019epuis," Récupéré sur le site : http ://blog.economie-numerique.net/2019/03/28/lintelligence-artificielle face-a-la-fraude-bancaire/.

[28] Vaudable, C., *Analyse et reconnaissance des émotions lors de conversations de centres d'appels*, Theses, Université Paris Sud - Paris XI, 2012.

[29] DataFlair, T., January 7 2020, https ://data-flair.training/blogs/python-mini-project-speech-emotion-recognition/ consulté le 12/5/2020 à 12 :19.

[30] Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., Mahjoub, M. A., and Cleder, C., "Automatic Speech Emotion Recognition Using Machine Learning," March 2019.

[31] Chernykh, V., Sterling, G., and Prihodko, P., "Emotion Recognition from Speech with Recurrent Neural Networks," Janvier 2017.

[32] ATTABI, Y., *RECONNAISSANCE AUTOMATIQUE DES ÉMOTIONS SPONTANÉES À PARTIR DU SIGNAL DE PAROLE*, theses, ÉCOLE DE TECHNOLOGIE SUPÉRIEURE UNIVERSITÉ DU QUÉBEC, NOVEMBRE 2015.

[33] Cambria, E., "Affective Computing and Sentiment Analysis IEEE Intelligent Systems," March 2016.

[34] Chandralika Chakraborty, P. H. T., "Issues and Limitations of HMM in Speech Processing : A Survey. International Journal of Computer Applications," Vol. 141, May 2016.

[35] Padmanabhana, J. and Premkumar, M. J. J., "Machine Learning in Automatic Speech Recognition," .

[36] Benalioua, G., BOUDIA, M. A., HAMOU, R. M., and AMINE, A., *Deep learning for speech recognition and vocal emotion detection now and next*, Master, Department of Computer Science, Tahar Moulay University of Saïda, Algeria, 2019.

[37] Min, L. C., Yildirim, S., Bulut, M., Busso, A. K. C., Deng, Z., and Narayanan, S., "Emotion Recognition based on Phoneme Classes Emotion Research Group Speech Analysis and Interpretation Lab and Integrated Media Systems," October 2004.

[38] Stuhlsatz, A., Meyer, C., Eyben, F., Zielke1, T., Meier, G., and Schuller, B., "DEEP NEURAL NETWORKS FOR ACOUSTIC EMOTION RECOGNITION : RAISING THE BENCHMARKS," 2011.

[39] Tashev, I. J., Wang, Z.-Q., and Godin, K., "Speech Emotion Recognition based on Gaussian Mixture Models and Deep Neural Networks," .

[40] Venkataramanan, K. and Rajamohan, H. R., "Emotion Recognition from Speech," Décembre 2019.

[41] Huang, A. and Bao, M. P., "Human Vocal Sentiment Analysis," May 2019.

[42] Mustaqeem and Kwon :, S., "A CNN Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition," 28 December 2019.

[43] Dumont, F., Mai 2017, http ://python.lecoinduprogrammeur.org/2017/05/07/jupyternotebook-ecrivez-executer-documentez-et-publiez-votre-code-python/ , consulté le 12/5/2020 à 15 : 17.

[44] "B34 Python, D(s d)," https ://docs.python.org/fr/3/tutorial/index.html consulté le 11/05/2020 à 03 : 06.

[45] "L, B TensorFlow : tout savoir sur la bibliothèque Machine Learning," https ://www.lebigdata.fr/tensorflow-definition-tout-savoir .consulté le 11/05/2020 à 03 : 20.

[46] "KERAS, D(sd)," https ://www.actuia.com/keras/ .consulté le 11/05/2020 à 03 : 50.

[47] Holmstrom, H. and Zars, V., "Effect of Feature Extraction when Classifying Emotions in Speech," 2018.

[48] Russo, S. R. and A, F., "the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)," https ://zenodo.org/record/1188976.XmdobPTjLIU. Consulté le 10/03/2020 à 12 :25.

[49] Najbauer, J., "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) :A dynamic, multimodal set of facial and vocal expressions in North American English," May 16 2018.

[50] Cunningham, P., "Dimension Reduction. In Machine Learning Techniques for Multimedia," .

[51] Hossan, M. A., Memon, S., and Gregory, M. A., "A novel approach for MFCC feature extraction," https ://ieeexplore.ieee.org/abstract/document/5709752 consulté le 11/03/2020 à 13 :31.

[52] Djemil, M., "SYSTÈME D'AIDE AUX NON-VOYANTS PAR COMMANDE VOCALE," 2011-2013.

[53] Fakhry, C., "La reconnaissance des émotions de la parole." 2 septembre 2019 récupéré sur le site https ://meritis.fr/ia/la-reconnaissance-des-emotions-de-la-parole/ consulté le 20/06/2020.

[54] Tato, A., "Apprentissage Machine : Techniques et applications(Partie 2)." 6 décembre 2018.

[55] Beard, R., Das, R., Ng, R. W. M., Gopalakrishnan, P. G. K., Swietojanski, L. E. P., and Miksik, O., "Multi-modal Sequence Fusion via Recursive Attention for Emotion Recognition," November 2018.

[56] Jalal, M. A., Loweimi, E., Moore, K. R., and Hain, T., "Learning Temporal Clusters Using Capsule Routing for Speech Emotion Recognition," September 15–19 2019.