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

**Université Abou Bakr Belkaid - Tlemcen**
**Faculté des Sciences**
**Département d'Informatique**


**Mémoire de fin d'études**

Pour l'obtention du diplôme de Master en Informatique

**Option : Intelligence Artificielle (I.A)**

# Détection d'émotions vocales et alerte en temps réel pour la protection des enfants à l'aide de l'intelligence artificielle.

***Réalisé par :***
Saidane Lamya
Benali Sid ahmed

*Soutenu le 01/07/2025, Devant le jury composé de :*

| | |
|---|---|
| **Président :** | M.BENAMAR Abdelkrim |
| **Encadrant :** | M.ABDERRAHIM Mohammed Alaeddine |
| **Examinatrice :** | Mme.MEHIAOUI Asma |

Promotion : 2024/2025

# *Remerciements*


ٱلْحَمْدُ للهِ وَٱلشُّكْرُ للهِ

En premier lieu, nous remercions Allah, le Très-Haut, qui nous a accordé la force, le courage ainsi que la volonté nécessaires pour mener à bien ce modeste travail.

Nous saisissons cette occasion pour exprimer nos sincères remerciements ainsi que notre profonde gratitude à toutes les personnes qui, de près ou de loin, ont contribué à la réalisation de ce mémoire.

Nos remerciements les plus chaleureux vont à **M.Abderrahim Mohammed Alaeddine**, notre encadrant, pour ses précieux conseils, sa disponibilité constante ainsi que son accompagnement tout au long de ce travail de recherche.

Nous adressons également nos remerciements aux membres du jury pour avoir accepté d'évaluer notre travail avec bienveillance et intérêt.

Nos remerciements s'adressent aussi à **Mlle Dib Samira**, notre déléguée, ainsi qu'à nos camarades et amis pour leur soutien, leur entraide ainsi que pour tous les moments de partage vécus tout au long de ce parcours universitaire.

Enfin, nous exprimons notre profonde reconnaissance à l'ensemble des enseignants et responsables de l'Université Abou Bekr Belkaïd  Tlemcen, pour la qualité de leur encadrement, la richesse de leur savoir ainsi que pour leur contribution précieuse à notre formation académique.

## إهداء

ها أنا أصل،
ليس كما حلمتُ أول مرة، ولا في الوقت الذي ظننته لي، لكنني وصلت... بعد كل سقوط، بعد كل سؤال مؤلم، بعد كل ليلة بكيتُ فيها بصمت.
تخرّجي ليس مجرد إنجاز، بل هو اعتذار صغير لطفلةٍ بداخلي، كانت تحلم أن تصبح طبيبة، لكن الأبواب أوصدت في وجهها، فأعادت ترتيب الألم على هيئة حلمٍ جديد، حتى وصلت... لطالما آمنتُ أن في القلب بذورًا لا تُثمر إلا حين تُسقى بالدمع والصبر، وها أنا اليوم أكتب هذه الكلمات بعد أن عبرتُ فصولًا من الشدائد، ووقفتُ في وجه العاصفة بعزيمةٍ وإيمان، وبأن الحلم، مهما تأخّر، لن يضلّ طريقه لمن أصرّ عليه.
تعثّرت، تألّمت، وبكيت وحدي... لكنني نهضت، لأن النور ليس من حولنا، بل منا، نحن الذين آمنا دومًا بأن في الصبر جمالًا، وفي الضعف قوة، وفي الألم ميلادًا جديدًا، وإن طال... ولو بعد حين.
إلى أولئك الذين رسموا أحلامهم على جدران الجامعات، وحملوا دفاترهم على طريقهم إلى الشهادة، إلى طلاب غزة الجامعيين الذين رحلوا قبل أن تكتمل الحكاية، قبل أن يُنادى بأسمائهم يوم التخرج... فصاروا نورًا لا ينطفئ في سماء العلم والكرامة، سلامًا على أرواحهم الطاهرة، وموعدنا عند ربٍّ لا ينسى.
إلى أبي، يا من طيّب الله خاطري بك، كنت أول من بارك نجاحي، وأقرب من آمن به، كرّست وقتك، ومالك، وصحتك، لترانا نكبر ونعلو، وها أنا اليوم أهديك تخرّجي.
إلى أمي، اليد التي امتدّت إليّ في العتمة، وأضاءت الطريق، الأنيسة التي احتضنتني دائمًا، وكنتِ السند حين وهن الجسد، والصوت حين خفت الأمل، فلولاكِ، بعد الله، ما أدركت أن بداخلي امرأةً لا تُهزم، وطموحًا لا يعرف المستحيل.
إلى زميلي في هذا المشوار، شريك التعب والنجاح، جمعنا حلمٌ واحد، فتقاسمنا السهر، والمسؤولية، والتحدّي. وها نحن اليوم نقطف ثمار اجتهادنا معًا. كان العمل معك نعمة، ونجاحي لا يكتمل دونك، فأنت جزء لا يتجزأ من هذه الحكاية، ومن هذا الإنجاز.
إلى جدّتاي، منبع الخير والحنان، اللتان تعلماننا بصمت معنى المحبة الصادقة، كم مشيتُ على وقع خطاهما ودعواتهما، وها أنا أصل إلى هذه القمة بفضل محبتهما، فكل إنجاز لي هو امتداد لعطائهما.
إلى جنود الخفاء... أخوالي، من دعموني بصمت، وآمنوا بي حين نسيت نفسي.
إلى إخوتي، النور حين تعتم الدروب،
وإلى صديقتي التي كانت أختًا من أم ثانية، صباح،
وإلى رفيقات الدرب، مَن جمعتنا الجامعة غرباء، وافترقنا إخوة... لضحكات خفّفت كل هم، ولتفاصيل لن أنساها.
هذا النجاح لكم كما هو لي،
أنا خريجة تعبٍ وانتظارٍ وأمل،
أنا حكايةٌ تأخّرت، لكنها جاءت كما أحبّها الله لي،
أنا مَن قاومت الألم، لا لأفوز بلقب، بل لأفوز بنفسي.

**لمياء**

# *Dédicace*

*Je dédie ce modeste travail à :*

*À mon père, pilier de ma vie, exemple de droiture et de travail acharné ;*
*À ma mère, source inépuisable d'amour, de soutien et de prières, elle qui m'a accompagné à chaque étape de ce parcours ;*
*À mes frères et surs, pour leur soutien constant, leur affection sincère et leur confiance en moi ;*
*À toute ma famille, qui m'a entouré de bienveillance tout au long de ce chemin ;*
*À mes amis, compagnons de route et de cur, pour tous les moments partagés, pour vos encouragements, vos rires et votre amitié sincère ;*
*À mon binôme, pour son engagement, sa patience et l'esprit d'équipe tout au long de ce projet.*

*Ce succès est le vôtre autant que le mien, car chacun de vos gestes, chacun de vos mots, a été une pierre à l'édifice de ma réussite.*

*Je rends grâce à Allah pour m'avoir permis d'arriver jusqu'ici, et je Lui demande de continuer à guider mes pas vers ce qui est utile, juste et bienveillant.*

***Sid Ahmed***

# Table des matières

# Table des figures

# Liste des tableaux

# Abréviations

| Abréviation | Signification |
|---|---|
| IA | Intelligence Artificielle |
| SER | Speech Emotion Recognition |
| MFCC | Mel Frequency Cepstral Coefficients |
| ASR | Automatic Speech Recognition |
| API | Application Programming Interface |
| GUI | Graphical User Interface |
| ONNX | Open Neural Network Exchange |
| dB | Décibel |
| Hz | Hertz |
| MESD | Mexican Emotional Speech Database |
| BESD-c | Bilingual Emotion Speech Dataset-C |

Tab. 1: Liste des abréviations utilisées dans le cadre du projet

# Chapitre 1

Introduction générale

## 1.1 Contexte

La voix joue un rôle central dans le développement humain et dans nos échanges quotidiens, en particulier pour exprimer les émotions et tisser des relations sociales. Chez les enfants, elle reflète souvent leur état émotionnel, surtout lorsqu'ils sont stressés, effrayés ou en détresse. Or, comme leur aptitude à exprimer verbalement ce qu'ils ressentent est encore en cours de maturation, il n'est pas toujours simple pour les parents de repérer les signes avant-coureurs d'un malaise ou d'une situation à risque (Alemu et al.,2023) [4]. Dans le même temps, les progrès récents en intelligence artificielle, en analyse audio et en apprentissage profond ont rendu possible la création d'outils capables d'identifier automatiquement les émotions à partir de la voix. Ces technologies ouvrent des perspectives prometteuses pour mieux protéger les personnes vulnérables, notamment les enfants.

## 1.2 Problématique

Avec la hausse des cas de maltraitance, d'accidents domestiques et de troubles psychologiques chez les enfants, il devient impératif de concevoir des solutions technologiques capables de repérer des signes de détresse en temps réel. Les méthodes classiques de surveillance parentale présentant certaines limites, il est légitime de se demander comment les progrès réalisés dans le domaine de la reconnaissance des émotions vocales peuvent permettre d'identifier rapidement et de manière automatique des états émotionnels critiques chez les enfants, même lorsque ceux-ci ne sont pas en mesure de s'exprimer verbalement.

## 1.3 Objectifs

Ce projet a pour but de développer un système intelligent capable de détecter les émotions vocales des enfants, afin de suivre leur état émotionnel à travers leur voix. L'objectif principal est de déclencher une alerte pour avertir les parents ou les tuteurs lorsqu'un signe de peur ou de détresse est détecté. Pour atteindre cet objectif, les étapes suivantes sont prévues :

- Utiliser un modèle d'intelligence artificielle déjà formé pour la reconnaissance des émotions dans la voix .
- Intégrer ce modèle dans un dispositif mobile ou embarqué qui fonctionne en temps réel .
- Assurer la précision et la fiabilité du système dans des environnements variés et réalistes.

## 1.4 Méthodologie

Dans le cadre de ce projet, nous utiliserons le modèle pré-entraîné ehcalabres/wav2vec2-lg-xlsr-en-speech-emotion-recognition, reposant sur l'architecture Wav2Vec2.0, spécialement adapté à l'analyse des émotions à partir de la voix. La méthodologie adoptée se décompose en plusieurs étapes clés :

- Traitement des données audio : les enregistrements seront nettoyés, découpés et normalisés afin d'optimiser la qualité des signaux analysés ;
- Utilisation du modèle Wav2Vec2 : ce modèle permettra d'extraire les informations sonores pertinentes pour identifier des émotions telles que la colère, la peur, la joie ou la tristesse ;
- Développement d'une application mobile : le système sera intégré dans une application capable de surveiller l'état émotionnel en continu et de transmettre des alertes en cas de détection de signaux préoccupants ;
- Phase de test et validation : des essais seront menés dans des conditions contrôlées puis réelles afin d'évaluer l'efficacité et la robustesse du dispositif.

## 1.5 Structure du mémoire

Ce mémoire est structuré en deux chapitres :

- **Première chapitre : État de l'art**
  Ce chapitre présente les notions théoriques liées aux émotions humaines ainsi que leur classification, notamment du point de vue vocal. Il détaille le fonctionnement général d'un système de reconnaissance des émotions vocales (SER), les méthodes de traitement du signal ainsi que les modèles d'intelligence artificielle les plus performants, tels que les architectures de type Wav2Vec2.

- **Deuxième chapitre : Expérimentation et Implémentation**
  Ce chapitre détaille l'ensemble des étapes techniques du projet, de la préparation du jeu de données à l'entraînement, à l'évaluation des modèles ainsi qu'à l'analyse des résultats obtenus.

# Chapitre 2

État de lart

## 2.1 Introduction

Les émotions sont au cur du comportement humain. Elles influencent la perception, la mémoire, la prise de décision et les interactions sociales. Chez les enfants, les émotions jouent un rôle fondamental dans le développement affectif et cognitif. Toutefois, leur expression est souvent limitée ou mal interprétée, en particulier à un jeune âge où le langage verbal reste en cours d'acquisition.

Parmi les différents moyens d'exprimer les émotions, la voix occupe une place privilégiée. Elle reflète des indices acoustiques subtils qui trahissent l'état émotionnel d'une personne, tels que le ton, le rythme ou l'intensité. C'est dans ce contexte que s'inscrit la reconnaissance des émotions vocales (Voice Emotion Recognition  VER), un domaine émergent de l'intelligence artificielle qui vise à détecter automatiquement les émotions à partir de la parole.

Grâce aux progrès en apprentissage profond et en traitement du signal audio, il est désormais possible d'entraîner des modèles capables d'identifier avec précision des émotions telles que la peur, la tristesse, la joie ou la colère, simplement à partir d'un enregistrement vocal. Chez les enfants, cette technologie offre un potentiel considérable pour détecter des signes de mal-être, améliorer la communication et renforcer leur sécurité.

## 2.2 Les émotions

Les émotions sont des états psychophysiologiques complexes, issus de l'interaction entre des processus cognitifs, corporels et sociaux [5]. Elles se traduisent par une expérience subjective, des réactions physiologiques (comme l'activation cardiaque ou hormonale), et des comportements expressifs observables (expressions faciales, posture, ton vocal). Les émotions remplissent des fonctions adaptatives fondamentales en orientant la prise de décision, en modulant les comportements et en facilitant les relations sociales [6].

Elles peuvent être classées selon leur valence en émotions positives (joie, fierté, gratitude) et négatives (colère, peur, tristesse), bien que cette dichotomie puisse entraîner des biais socioculturels liés, par exemple, au genre ou à l'ethnicité [7]. Contrairement à de simples réactions passagères, les émotions peuvent influencer durablement les attitudes et les traits de personnalité [8].

Par ailleurs, leur perception et leur expression sont profondément façonnées par les contextes culturels et linguistiques, ce qui en fait des phénomènes contextualisés, à la fois universels et culturellement variables [9]. Comprendre les émotions requiert ainsi une approche interdisciplinaire intégrant psychologie, neurosciences, sociologie et linguistique [10].

### 2.2.1 Types des émotions

On peut classer les émotions humaines en diverses catégories selon leur valence, ou leur nature plaisante ou déplaisante. Cette classification sépare généralement les émotions dites positives de celles négatives, même si cette séparation peut parfois alimenter des stéréotypes associés à des facteurs individuels tels que le sexe ou l'origine ethnique [7].

#### Émotions positives

Les émotions positives sont associées à des expériences plaisantes et contribuent au bien-être psychologique. Elles favorisent l'ouverture cognitive, les comportements prosociaux et l'adaptation à l'environnement. Parmi les principales émotions positives, on retrouve la joie, l'amour, la gratitude, l'enthousiasme, la fierté et la sérénité.

#### Émotions négatives

Les émotions négatives, quant à elles, sont liées à des expériences perçues comme désagréables ou menaçantes. Bien qu'elles soient souvent perçues comme indésirables, elles jouent un rôle crucial dans la survie, en signalant un danger ou un déséquilibre à corriger. Les émotions négatives incluent la peur, la colère, la tristesse, la honte, la culpabilité et le dégoût.

### 2.2.2 Classification des émotions

La classification des émotions est une démarche visant à organiser et catégoriser les émotions humaines pour mieux comprendre leur nature, leur origine et leur fonction. Plusieurs modèles théoriques majeurs existent, chacun apportant une perspective différente.

#### Modèle discret

La théorie des émotions discrètes repose sur l'hypothèse selon laquelle il existe un ensemble limité d'émotions primaires, considérées comme fondamentales, universelles et biologiquement déterminées. Ces émotions sont dites ń discrètes ż car elles forment des catégories distinctes, chacune ayant ses propres caractéristiques expressives, physiologiques et comportementales.

**Émotions primaires** Les émotions primaires sont émotions de base universelles et innées qui sont présentes chez le bébé [11]. Non seulement elles sont innées mais en plus de ça elles sont automatiques, elles sont inconscientes et elles ont un déclenchement rapide, telles des réflexes [2].

**Émotions secondaires** Les émotions secondaires, ou complexes, naissent du mélange d'émotions primaires (par exemple, la honte combine peur et colère). Elles apparaissent plus tardivement (entre 1 et 4 ans), ne sont pas innées et ne se déclenchent pas automatiquement : leur émergence est plus progressive et leur durée plus longue. Parmi elles : la jalousie, la culpabilité, l'embarras ou l'envie [2].

TAB. 2.1: Émotions basiques identifiées par différents auteurs[2]

| **Acteurs** | **Émotions basiques** |
|---|---|
| Ekman et al. (1982) | Colère, dégoût, joie, tristesse, peur, surprise, neutralité |
| Izard (1971) | Colère, mépris, dégoût, détresse, peur, culpabilité, intérêt, joie, honte, surprise |
| Gray (1982) | Rage, terreur, anxiété, joie |
| Mower (1960) | Douleur, plaisir |
| James (1884) | Peur, chagrin, amour, rage, colère, dégoût |
| McDougall (1926) | Exaltation, peur, soumission, tendresse, émerveillement |
| Plutchik (1980) | Acceptation, colère, anticipation, dégoût, peur, joie, tristesse, surprise |
| Frijda (1986) | Intérêt, joie, désir, chagrin, émerveillement |
| Tomkins (1984) | Colère, intérêt, mépris, dégoût, détresse, peur, joie, honte, surprise |
| Panksepp (1982) | Attente, peur, rage, panique |
| Arnold (1960) | Courage, colère, aversion, désir, désespoir, tristesse, amour, espoir, abattement, haine, peur |

### Modèle tridimensionnel

le modèle tridimensionnel [12], appelé aussi modèle PAD (Pleasure, Arousal, Dominance), est uncadre en psychologie environnementale qui conceptualise les réponses émotionnelles humaines face à un environnement selon trois dimensions fondamentales selon trois dimensions fondamentales

- **Plaisir (Valence)** : Cette dimension évalue dans quelle mesure une personne ressent du bien-être, de la satisfaction ou du bonheur dans un environnement donné. Elle s'étend d'un pôle négatif (douleur, mécontentement) à un pôle positif (plaisir, joie).
- **Activation (Arousal)** : Elle mesure le niveau d'éveil mental et physique d'un individu, allant de la somnolence ou l'ennui à des états d'excitation intense ou de stimulation accrue.
- **Domination (Dominance)** : Cette dimension traduit le degré de contrôle ou d'influence perçu par l'individu dans une situation. Elle oscille entre des sentiments de soumission ou d'impuissance et ceux de maîtrise, de liberté ou d'autorité.

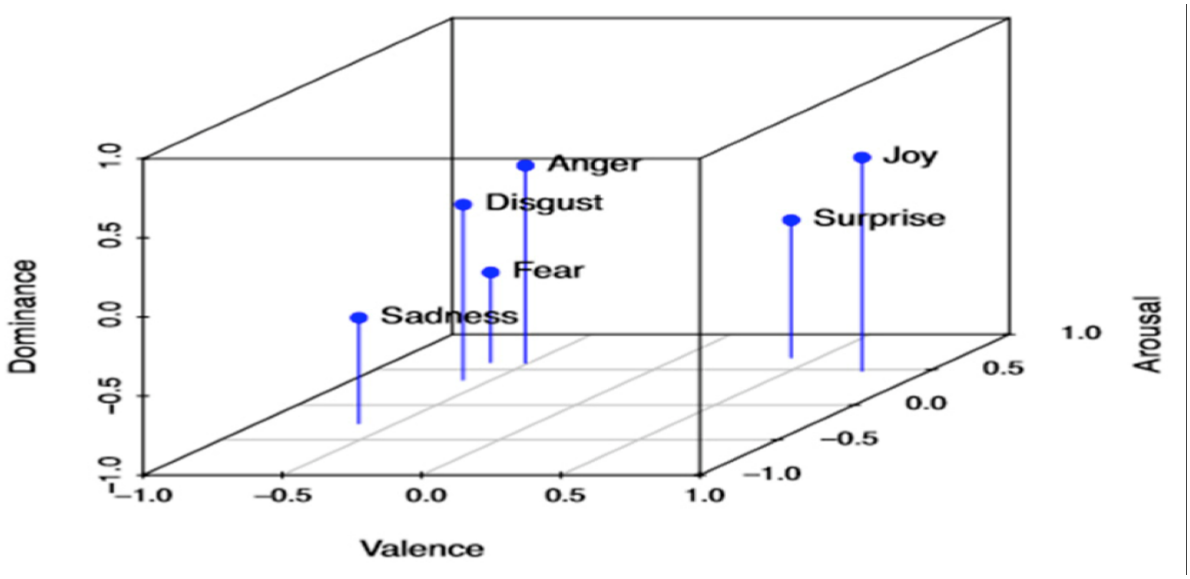


FIG. 2.1: The VAD (Valence-Arousal-Dominance) model spanned across the six basic emotion[1]

### 2.2.3 Caractéristiques des émotions vocales :

Le tableau ci-dessous illustre certaines caractéristiques vocales associées aux émotions primaires :

| **Émotions** | **Caractéristique de la voix** |
|---|---|
| Joie | Énergie vocale élevée.<br>Volume élevé<br>Rythme rapide<br>Intensité élevée<br>Variation mélodique marquée<br>Durée des pauses plus courte |
| Tristesse | Faible énergie<br>Voix basse<br>Rythme lent<br>Intensité réduite<br>Durée des pauses plus longue |
| Colère | Énergie vocale élevée<br>Volume élevé<br>Rythme rapide et irrégulier<br>Intensité élevée<br>Variation mélodique abrupte |
| Dégoût | Énergie vocale variable<br>Volume variable<br>Rythme variable<br>Variation mélodique instable<br>Durée des pauses variable |
| Peur | Énergie vocale variable<br>Volume élevé<br>Variation mélodique instable<br>Rythme rapide et saccadé<br>Durée des pauses imprévisible |
| Neutre | Énergie vocale stable<br>Volume moyen<br>Rythme régulier<br>Intensité modérée<br>Variation mélodique faible<br>Pauses régulières |

TAB. 2.2: Caractéristiques de la voix selon les émotions

### 2.2.4 La protection émotionnelle des enfants

Les enfants vivent chaque jour de nombreuses émotions, parfois intenses, qu'ils ne savent pas toujours comment gérer. Apprendre à réguler ces émotions  c'est-à-dire à les comprendre, les exprimer calmement, et ne pas se laisser submerger  est un élément essentiel de leur développement. C'est ce qu'on appelle la protection émotionnelle : elle

permet à l'enfant de se sentir en sécurité avec ce qu'il ressent, tout en apprenant à vivre avec ses émotions [13].

Dès le plus jeune âge, le cerveau de l'enfant commence à traiter les émotions grâce à des régions spécifiques. Une de ces régions est l'amygdale, qui réagit fortement aux émotions comme la peur ou la colère [14]. C'est comme une alarme qui s'active quand l'enfant ressent quelque chose de fort. Chez les jeunes enfants, cette alarme est très active, ce qui explique pourquoi ils peuvent réagir de manière impulsive (en criant, pleurant ou fuyant, par exemple).

Heureusement, une autre partie du cerveau, appelée cortex préfrontal, va progressivement apprendre à calmer cette alarme. Cette zone permet à l'enfant de réfléchir, de se concentrer et de contrôler ses réactions. Mais elle se développe plus lentement et continue à mûrir jusqu'à l'adolescence [15]. C'est pour cela qu'il est normal qu'un enfant ait parfois du mal à se contrôler : son cerveau est encore en construction.

Il existe aussi des connexions importantes entre ces deux parties du cerveau. Quand elles fonctionnent bien, l'enfant peut mieux comprendre ce qu'il ressent et y répondre de manière adaptée. Ce travail en équipe entre les différentes zones du cerveau est essentiel pour développer ce qu'on appelle la conscience émotionnelle : la capacité à reconnaître ses émotions, à leur donner un nom, et à en parler [16].

Les enfants qui ont du mal à réguler leurs émotions peuvent, en grandissant, rencontrer des difficultés comme l'anxiété, la tristesse excessive ou des comportements agressifs [17]. C'est pourquoi il est si important de les accompagner tôt. Un environnement bienveillant, avec des adultes à l'écoute, aide énormément. Il existe aussi des approches éducatives et thérapeutiques qui peuvent soutenir les enfants dans ce développement [18].

### 2.2.5 La reconnaissance automatique des émotions vocales :

La reconnaissance vocale des émotions SER (Speech Emotion Recognition) est une branche de l'intelligence artificielle et du traitement du signal qui vise à détecter et interpréter les états affectifs d'un individu à partir de sa voix. En analysant des caractéristiques acoustiques telles que la hauteur (pitch), l'intensité, le rythme, le timbre et les variations prosodiques, il est possible d'identifier les émotions de base comme la joie, la colère, la peur, la tristesse, le dégoût ou la surprise. Cette approche est particulièrement utile dans des contextes où l'analyse visuelle n'est pas disponible, comme les conversations téléphoniques ou les systèmes vocaux interactifs. Dans le cadre de la protection des enfants, elle permettrait de repérer des signaux de détresse émotionnelle, même en l'absence de mots explicites, et d'alerter les parents ou les professionnels en temps réel. Couplée à des algorithmes de deep learning, cette technologie offre des perspectives prometteuses pour créer des systèmes sensibles aux besoins affectifs des enfants.[19]

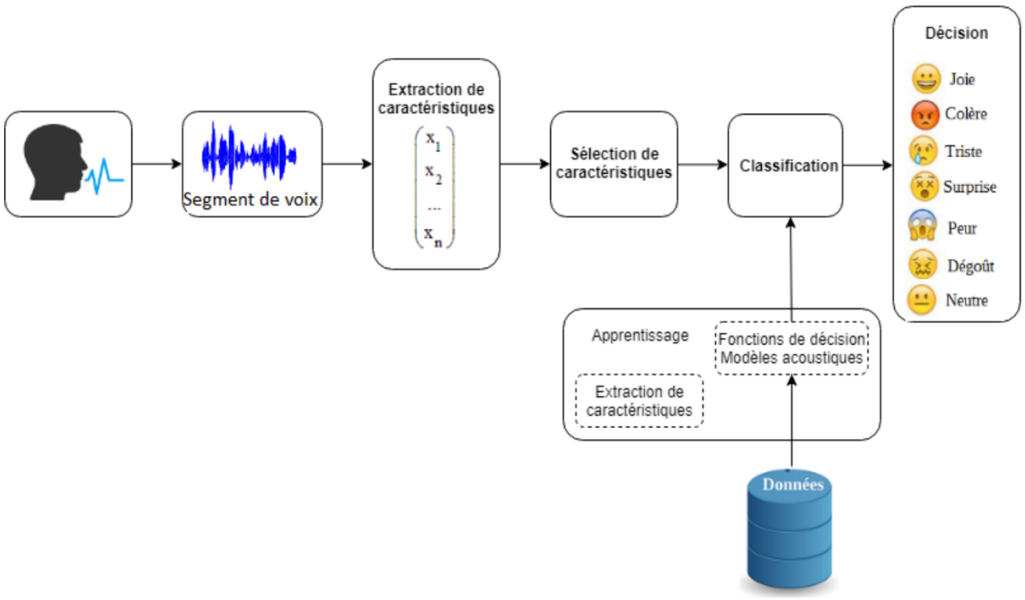


FIG. 2.2: Architecture d'un système de reconnaissance des émotions[2]

### 2.2.6 Fonctionnement général d'un système SER

Le fonctionnement d'un système de reconnaissance des émotions vocales (SER) repose sur plusieurs étapes clés permettant de traiter un signal vocal brut pour en extraire les émotions exprimées. Ce processus est fondé sur une chaîne de traitement du signal audio couplée à des techniques d'intelligence artificielle. Les principales étapes sont les suivantes :

#### Acquisition du signal vocal

La première étape consiste à recueillir le signal vocal à analyser. Cela peut être réalisé à l'aide de microphones ou d'enregistreurs numériques, dans un environnement contrôlé ou non. La qualité de l'enregistrement est cruciale, car elle influence directement la précision de la détection émotionnelle. Les bases de données publiques ou les enregistrements réalisés sur le terrain constituent les sources principales pour l'acquisition [20].

#### Prétraitement du signal audio

Le prétraitement du signal audio représente une étape fondamentale dans le traitement de la parole, particulièrement en reconnaissance des émotions vocales. Il consiste à appliquer un ensemble de techniques visant à améliorer la qualité du signal vocal brut, afin d'en faciliter l'analyse et d'en garantir la fiabilité pour les étapes ultérieures telles que l'extraction des caractéristiques acoustiques et la classification des émotions [21, 22].

En effet, les signaux vocaux enregistrés sont souvent affectés par des bruits de fond, des variations d'intensité, ou encore des distorsions dues à l'environnement ou au matériel utilisé. Le rôle du prétraitement est alors de nettoyer, normaliser et structurer le signal pour le rendre plus robuste et plus représentatif de l'émotion exprimée.

Dans le cadre de notre système de reconnaissance des émotions, le prétraitement repose principalement sur trois opérations successives (voir Figure 2.3) :

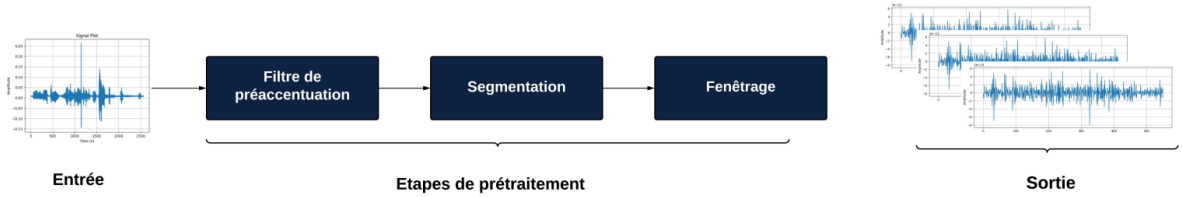


FIG. 2.3: Étapes du prétraitement de signal de la parole [3]

- **Filtre de préaccentuation :** La préaccentuation est une technique couramment employée dans le traitement du signal audio, consistant à appliquer un filtre passe-haut au signal vocal brut. Ce filtre a pour objectif d'amplifier les composantes haute fréquence, souvent atténuées lors de l'acquisition du signal par les microphones.

  Cette opération améliore la clarté des transitions vocales et facilite une analyse fréquentielle plus fine et plus fiable. Elle est particulièrement utile dans les contextes où les détails acoustiques jouent un rôle crucial, comme l'analyse de la prosodie ou du timbre émotionnel.

  Sur le plan mathématique, la préaccentuation est généralement formulée ainsi :

$$y(t) = x(t) - \alpha \cdot x(t-1) \tag{2.1}$$

  où :

  - $x(t)$ est le signal original,
  - $y(t)$ est le signal pré-accentué,
  - $\alpha$ est le coefficient de préaccentuation, souvent choisi entre 0,9 et 0,97.

  Cette opération permet ainsi de renforcer les transitions rapides du signal vocal, en accentuant les variations de haute fréquence qui sont souvent porteuses d'informations émotionnelles importantes [21, 22].

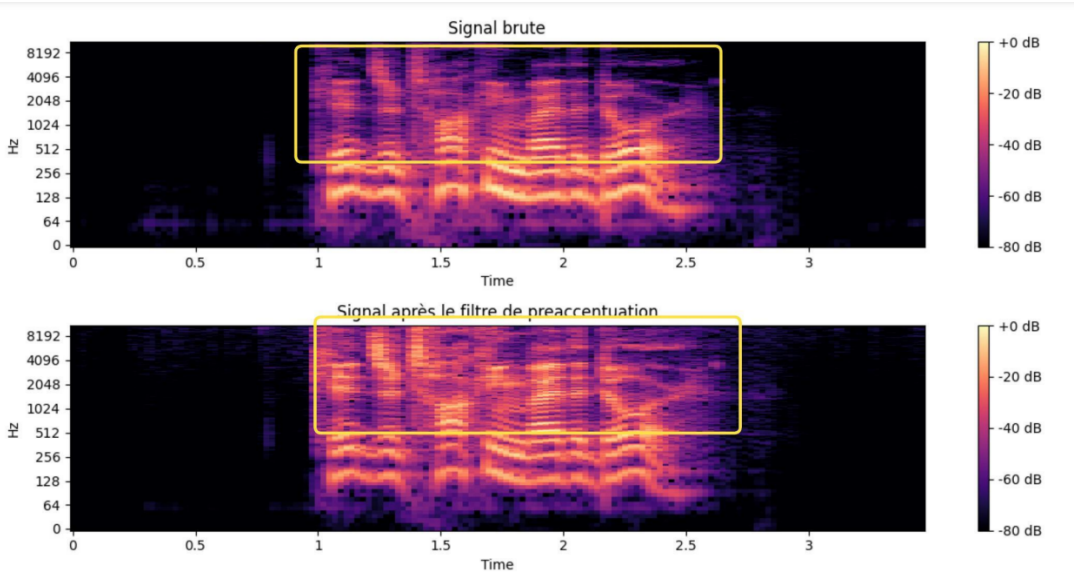


FIG. 2.4: Effet du filtre de préaccentuation sur le spectrogramme d'un signal audio

- **Segmentation :** Suite à la préaccentuation, le signal est segmenté en trames temporelles de courte durée, typiquement comprises entre 20 et 30 millisecondes. Cette opération, dite de segmentation, vise à isoler les variations locales du signal vocal, intrinsèquement non stationnaire. L'hypothèse de quasi-stationnarité est ainsi valable au sein de chaque trame, condition indispensable pour l'application d'analyses spectro-temporelles telles que la transformée de Fourier discrète ou l'extraction des coefficients cepstraux [20, 23].

  Afin d'assurer la continuité temporelle et limiter les pertes d'information acoustique, un recouvrement (overlap) de l'ordre de 50 % est couramment appliqué entre trames successives.

- **Fenêtrage :** Chaque segment est ensuite multiplié par une fonction fenêtre, telle que la fenêtre de Hamming ou de Hanning. Cette étape, appelée fenêtrage, a pour but d'atténuer les discontinuités aux extrémités des trames, qui pourraient engendrer des artefacts dans le domaine fréquentiel. Elle permet notamment de réduire les effets de repliement spectral (spectral leakage) lors de l'application de la transformée de Fourier [24, 25].

  Par exemple, la fenêtre de Hamming est définie par la formule suivante :

$$w(n) = 0{,}54 - 0{,}46 \cdot \cos\left(\frac{2\pi n}{N-1}\right)$$

  Le fenêtrage contribue ainsi à préserver la structure harmonique du signal et à assurer une représentation spectrale précise, condition indispensable pour une détection d'émotions fiable.

- **Normalisation :** La normalisation de l'amplitude consiste à ajuster le volume sonore du signal vocal pour le ramener dans une plage cohérente (souvent entre $-1$ et 1). Cette étape permet d'uniformiser l'intensité sonore des enregistrements, souvent affectée par des facteurs externes (distance au micro, environnement, etc.).

Elle joue un rôle crucial dans la reconnaissance émotionnelle, en assurant que les variations d'amplitude utilisées par les modèles reflètent bien les émotions réelles (comme la colère ou la tristesse), et non des artefacts techniques. La méthode la plus courante consiste à diviser le signal par son amplitude maximale :

$$x_{\text{norm}}(t) = \frac{x(t)}{\max(|x(t)|)} \tag{2.2}$$

Cela garantit un traitement plus stable et robuste lors de l'extraction des caractéristiques [21].

## 2.3 Extraction des caractéristiques vocales

L'extraction des caractéristiques vocales constitue une étape fondamentale dans le processus de reconnaissance des émotions à partir de la voix. Elle permet de convertir le signal acoustique brut en une représentation numérique compacte et pertinente, exploitable par les algorithmes de classification pour déduire l'état émotionnel du locuteur. Cette étape repose sur des fondements psycho-acoustiques et linguistiques selon lesquels certaines propriétés du signal vocal sont corrélées à des états émotionnels spécifiques [26, 19].

Les caractéristiques extraites peuvent être regroupées en plusieurs catégories. Les caractéristiques prosodiques décrivent les variations globales du signal vocal, notamment l'intonation, la durée et l'intensité. La hauteur fondamentale (pitch) (voir Figure 2.5), correspondant à la fréquence de vibration des cordes vocales, tend à augmenter dans des émotions telles que la peur ou la joie, tandis qu'elle diminue dans la tristesse [27]. L'énergie du signal (voir Figure 2.6) représente l'intensité vocale et peut refléter le degré d'engagement émotionnel : une émotion comme la colère est généralement associée à une énergie plus élevée que la tristesse ou la neutralité [28].

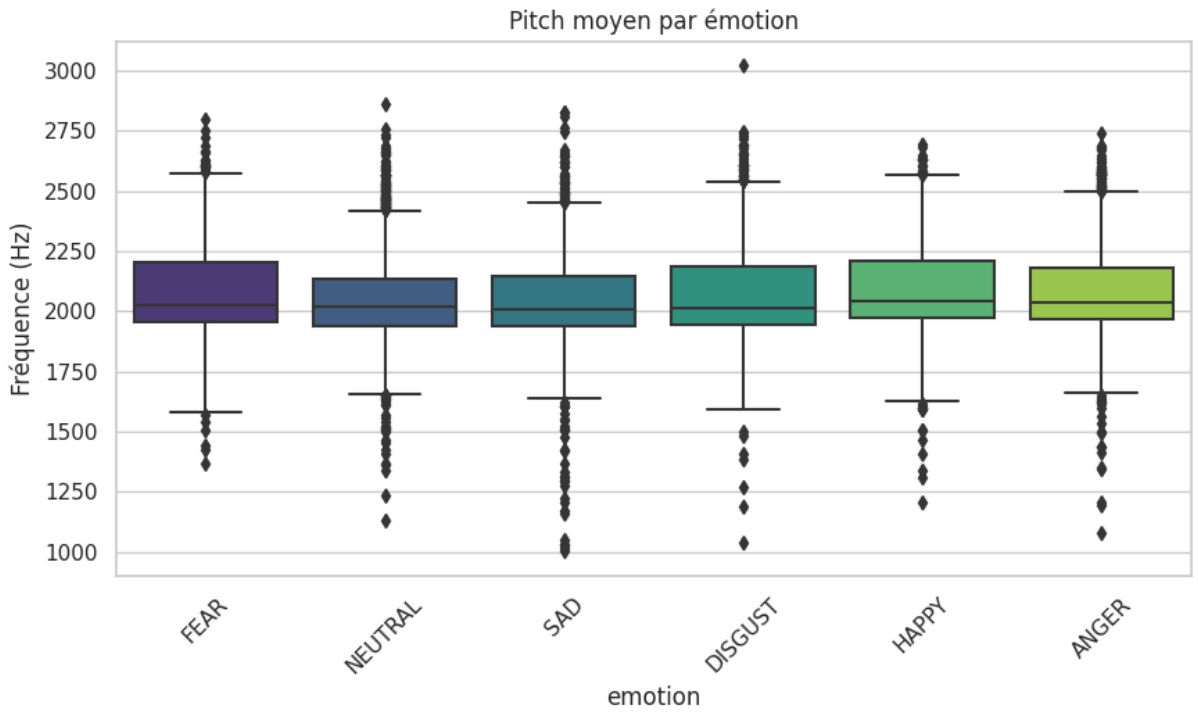


Fig. 2.5: Pitch moyen par émotion

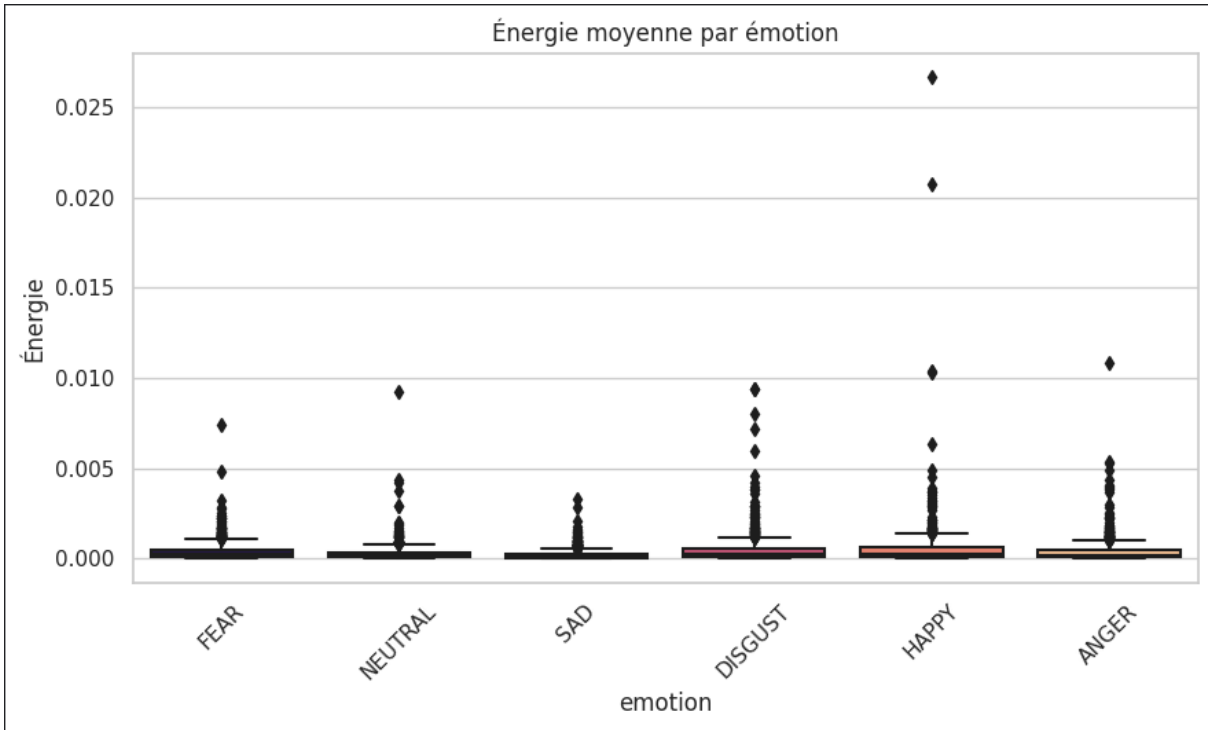


Fig. 2.6: Énergie moyenne par émotion

Les **caractéristiques spectrales** permettent quant à elles de décrire la structure fréquentielle du signal. Les MFCC (Mel-Frequency Cepstral Coefficients), largement utilisés en reconnaissance vocale, modélisent la réponse fréquentielle du système auditif humain.

Leur extraction passe par plusieurs étapes : segmentation du signal en trames, transformation de Fourier rapide (FFT), application d'une banque de filtres sur l'échelle de Mel, compression logarithmique, puis transformée en cosinus discrète (DCT) pour obtenir une représentation compacte et informative [29]. D'autres représentations, telles que le spectrogrammee et le Mel spectrogramme, fournissent une visualisation temps-fréquence du contenu spectral du signal, avec une échelle de fréquence non linéaire dans le cas du Mel spectrogramme. Ces représentations sont particulièrement adaptées comme entrées pour les modèles d'apprentissage profond tels que les CNN et les transformers [30].

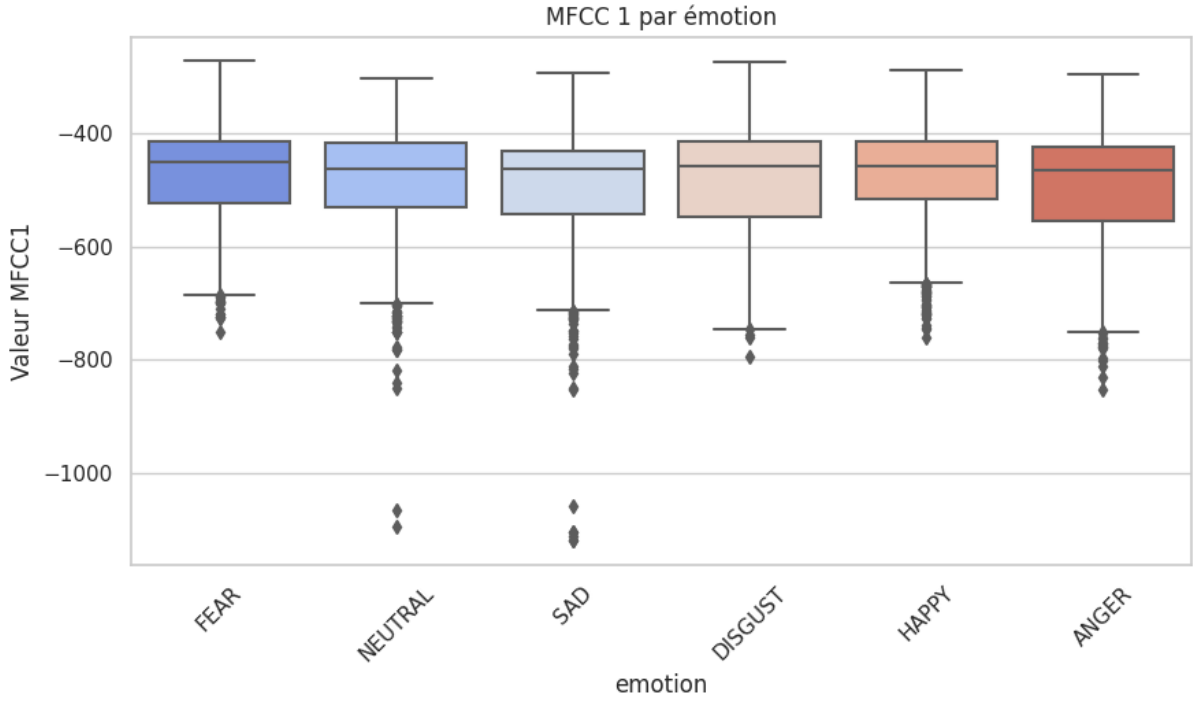


Fig. 2.7: MFCC 1er coefficient par émotion

## 2.4 Méthodes de classification des émotion :

La classification des émotions est une étape cruciale dans la reconnaissance des états affectifs à partir de données vocales. Elle consiste à attribuer une catégorie émotionnelle précise à partir des caractéristiques extraites du signal audio. Pour cela, plusieurs types d'algorithmes peuvent être utilisés, allant des méthodes traditionnelles d'apprentissage automatique aux approches récentes basées sur l'apprentissage profond et les modèles pré-entraînés. Chaque méthode présente ses avantages et ses limites selon la nature des données et les objectifs visés.

**Algorithme :**

Les algorithmes classiques regroupent des techniques comme les Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), les Hidden Markov Models (HMM) et les forêts aléatoires (Random Forests). Ces méthodes nécessitent souvent une étape préalable de sélection ou réduction des caractéristiques pour optimiser leur performance. Elles sont

appréciées pour leur simplicité et leur rapidité, mais leur efficacité dépend fortement de la qualité des données et des caractéristiques extraites.

**Apprentissage profond :**

L'apprentissage profond utilise des réseaux de neurones profonds capables d'apprendre automatiquement des représentations complexes à partir des données brutes. Les architectures comme les réseaux convolutifs (CNN), les réseaux récurrents (LSTM) et les transformers permettent de modéliser efficacement les caractéristiques spatiales et temporelles des signaux vocaux, offrant ainsi des performances supérieures dans la classification des émotions.

**Les modèles pré-entraînés :**

Les modèles pré-entraînés, tels que Wav2Vec 2.0, HuBERT ou Whisper, sont formés sur de vastes corpus audio en mode auto-supervisé. Ils fournissent des représentations riches du signal vocal, qui peuvent être adaptées (fine-tuned) à la classification émotionnelle avec relativement peu de données spécifiques. Cette approche permet de gagner en précision tout en réduisant les besoins en données annotées.

## 2.5 Modèles de détection d'émotions à partir de la voix :

**Wav2Vec2 XLSR-53 (Facebook) :**

Le modèle facebook/wav2vec2-large-xlsr-53 , développé par Facebook AI (Meta), est une version multilingue avancée du modèle wav2vec 2.0. Il a été introduit par Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed et Michael Auli dans le but de proposer une approche unifiée pour la reconnaissance vocale dans plusieurs langues, y compris celles à ressources limitées.

Ce modèle repose sur une stratégie d'apprentissage auto-supervisé, où il est pré-entraîné sur des signaux audio bruts à une fréquence de 16 kHz. L'apprentissage se fait par la prédiction de représentations masquées dans l'audio à l'aide d'un objectif *contrastif*, tout en apprenant une quantification partagée entre les langues. Cette approche permet au modèle de capturer des représentations communes à différentes langues dans un espace acoustique cohérent.

**Objectif principal** Former un seul modèle capable de comprendre et de traiter efficacement des discours issus de 53 langues, en favorisant le transfert de connaissances interlingue, notamment au profit des langues faiblement dotées.

**Résultats expérimentaux**

- Une réduction de 72 % du taux d'erreur phonémique (PER) sur le benchmark *CommonVoice*.
- Une amélioration de 16 % du taux d'erreur de mots (WER) sur le corpus *BABEL*, comparé à un système monolingue équivalent.

- Des représentations latentes partagées entre les langues, avec un taux de partage plus élevé pour les langues apparentées.

**Wav2Vec2-Robust Emotion :**

Le modèle "audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim" est un système de reconnaissance émotionnelle basé exclusivement sur le signal audio. Développé par l'entreprise audEERING GmbH, spécialisée dans l'analyse émotionnelle automatique, ce modèle n'effectue aucune transcription textuelle : il se concentre uniquement sur les caractéristiques acoustiques de la voix pour estimer les émotions humaines dans un espace continu à trois dimensions [31].

L'architecture utilisée comme base est Wav2Vec2-Large-Robust, un modèle de traitement de la parole développé par Facebook AI Research (FAIR). Ce dernier repose sur un encodeur convolutionnel, combiné à des couches Transformer, dans une version initiale comportant 24 couches, conçue pour apprendre des représentations riches à partir d'audio brut [32]. Afin de réduire la complexité computationnelle, la moitié des couches a été supprimée, aboutissant à une version allégée à 12 couches, appelée wav2vec2-large-robust-12 [33].

Le modèle ainsi simplifié a ensuite été fine-tune (affiné) par audEERING sur la tâche spécifique de régression émotionnelle, à l'aide du corpus MSP-Podcast, version 1.7. Ce corpus contient plusieurs milliers d'enregistrements vocaux issus de podcasts en anglais, annotés selon des dimensions émotionnelles continues, à l'aide du crowdsourcing. Il constitue aujourd'hui une référence majeure pour la recherche en reconnaissance des émotions [34].

En sortie, ce modèle fournit pour chaque segment audio un triplet de scores continus (valence, arousal, dominance), normalisés dans une plage allant approximativement de 0 à 1 [31]. Concrètement, il évalue [35] :

- valence, représentant la positivité ou la négativité de l'émotion ;
- l'arousal, indiquant le niveau d'éveil ou d'énergie émotionnelle ;
- la dominance, qui mesure le degré de contrôle ou de puissance émotionnelle [34].

Ces scores résultent d'un apprentissage supervisé par régression, et non d'une classification discrète des émotions, ce qui permet une interprétation plus nuancée et continue des états émotionnels exprimés vocalement.

**Wav2Vec2-LG XLSR Emotion :**

Ce modèle, nommé "ehcalabres/wav2vec2-lg-xlsr-en-speech-emotion-recognition", est un système de reconnaissance des émotions vocales basé exclusivement sur le signal audio. Il s'agit d'une version fine-tunée du modèle pré-entraîné jonatasgrosman/wav2vec2-large-xlsr-53-english, spécifiquement adaptée à la tâche de reconnaissance des émotions exprimées dans la parole[36].

Le fine-tuning a été réalisé sur le corpus RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song), qui contient 1440 enregistrements audio d'acteurs exprimant huit émotions distinctes en anglais : colère (angry), calme (calm), dégoût (disgust), peur (fearful), joie (happy), neutre (neutral), tristesse (sad) et surprise (surprised)[36].

Ce dataset est une référence reconnue dans la recherche en reconnaissance des émotions vocales.

L'architecture de base repose sur Wav2Vec 2.0 Large XLSR, un modèle de traitement de la parole multilingue développé pour apprendre des représentations audio riches à partir de données brutes. Ce modèle combine un encodeur convolutionnel avec des couches Transformer profondes, permettant de capter des caractéristiques acoustiques complexes[37].

En sortie, le modèle effectue une classification multi-classes sur les huit émotions mentionnées, atteignant une précision (accuracy) de 82,23 % et une perte (loss) de 0,5023 sur le jeu d'évaluation RAVDESS.[37] Ces performances en font un outil efficace pour des applications variées telles que la surveillance émotionnelle, l'analyse de la satisfaction client, ou encore la santé mentale.

### 2.5.1 Travaux connexes :

De nombreuses études ont été menées afin d'extraire des informations émotionnelles à partir du signal vocal, notamment dans des contextes d'assistance à des personnes vulnérables ou de prévention de situations critiques. La reconnaissance des émotions à travers la voix est en effet un domaine essentiel dans le développement d'interfaces intelligentes capables d'interagir de manière plus humaine et plus empathique. Plusieurs recherches se sont concentrées sur la détection des émotions chez les enfants, bien que cette population reste moins étudiée que les adultes. D'autres approches ont également proposé des systèmes d'alerte automatique dans des contextes de protection, mais l'intégration conjointe des deux dimensions  détection vocale des émotions infantiles et alerte parentale reste encore marginale.

L'étude de (Yan et al). Propose un système dédié à la détection de situations potentiellement abusives chez les enfants, basé sur l'analyse des signaux audio. Le système repose sur la transformée de Fourier à court terme (STFT) pour l'extraction des caractéristiques spectrales, couplée à des modèles d'apprentissage supervisé afin de classifier différents types de sons, tels que les pleurs, les cris ou les rires. Le traitement est déployé sur une plateforme Nvidia Jetson (Edge GPU), optimisée pour des applications embarquées en temps réel. L'objectif principal est de garantir une réaction immédiate du personnel encadrant ou des tuteurs lorsque des comportements inhabituels sont détectés, en particulier dans des environnements tels que les centres d'accueil ou les foyers spécialisés [38].

Une autre contribution significative est celle de Yao et al., qui ont exploré la reconnaissance automatique des pleurs de nourrissons dans des environnements domestiques sujets à des interférences sonores. Leur méthodologie combine des représentations spectrales profondes issues de réseaux de neurones convolutifs (CNN) avec des descripteurs acoustiques traditionnels (MFCC, spectrogrammes). L'évaluation expérimentale montre que l'association de ces deux types de caractéristiques améliore sensiblement la robustesse du système, même en présence de bruits ambiants. Ce travail illustre la capacité de l'intelligence artificielle à s'adapter à des conditions d'enregistrement variées et à fournir des résultats fiables pour une application pratique à domicile [39].

En ce qui concerne la détection des émotions proprement dites chez l'enfant, le projet Emotirob s'est orienté vers une approche multimodale appliquée à la robotique sociale. Il vise à améliorer l'interaction entre des enfants fragilisés et un robot compagnon. Le système prend en compte non seulement le contenu linguistique des phrases prononcées,

mais aussi leur prosodie (intonation, rythme, intensité), afin d'estimer la valence émotionnelle des énoncés. Cette double approche permet au robot d'adapter ses réponses émotionnelles, et contribue ainsi à créer un environnement plus rassurant et engageant pour les enfants concernés [40].

Ces recherches montrent l'évolution du domaine vers des solutions intelligentes capables d'interpréter la voix humaine dans un objectif de protection ou de soutien. Cependant, elles ne proposent pas de mécanisme intégré qui relie directement la détection vocale d'émotions infantiles à un système d'alerte parentale en temps réel. Le projet que nous proposons s'inscrit donc dans une démarche novatrice, en réunissant les capacités de classification émotionnelle, de traitement embarqué, et de notification immédiate à destination des parents.

### 2.5.2 Conclution :

Ce chapitre a dressé un panorama complet des fondements liés à la reconnaissance des émotions, en abordant à la fois les dimensions psychologiques et technologiques. Dans une première partie, nous avons clarifié la notion d'émotion, en présentant ses différentes définitions, ses typologies ainsi que les principales classifications utilisées dans le domaine scientifique. L'accent a également été mis sur l'importance de la protection émotionnelle des enfants, soulignant la nécessité de détecter leurs états émotionnels pour mieux les accompagner.

La deuxième partie du chapitre s'est centrée sur la reconnaissance des émotions à partir de la voix. Nous y avons décrit le fonctionnement global d'un système SER (Speech Emotion Recognition), en détaillant les étapes clés telles que le prétraitement, l'extraction des caractéristiques et la classification. Plusieurs approches ont été explorées, allant des méthodes classiques aux modèles récents fondés sur l'intelligence artificielle. Enfin, les travaux connexes ont permis de situer notre problématique dans le contexte de la recherche actuelle.

Ainsi, ce chapitre pose les bases théoriques et techniques nécessaires à la compréhension des enjeux et des solutions liées à la détection des émotions vocales, en particulier chez les enfants.

# Chapitre 3

Expérimentation et Implémentation

## 3.1 Introduction

La présente section est consacrée à la description méthodique de la démarche expérimentale entreprise dans le cadre de ce projet. Elle vise à exposer, de manière structurée, l'ensemble des étapes techniques mises en uvre pour concevoir, entraîner, évaluer et adapter un modèle de reconnaissance automatique des émotions dans la voix des enfants.

Notre approche repose principalement sur le *fine-tuning* de modèles pré-entraînés de type `Wav2Vec2`, associés à des bases de données spécialisées contenant des enregistrements vocaux d'enfants annotés selon différentes émotions.

Ce processus a nécessité plusieurs étapes clés, allant du prétraitement des données audio à l'évaluation des performances du modèle à l'aide de métriques standards (précision, rappel, F1-score, matrice de confusion), en passant par l'extraction des caractéristiques et le choix du modèle le plus adapté.

L'objectif est de développer un modèle performant et suffisamment robuste pour être intégré dans une application mobile à destination des parents et des professionnels de la petite enfance.

## 3.2 Données utilisées :

En raison du manque de bases de données vocales émotionnelles spécialisées pour les enfants, et dans le but d'accroître la quantité et la diversité des données disponibles pour l'apprentissage automatique, nous avons choisi de combiner deux bases principales obtenues via la plateforme Kaggle : MESD (Mexican Emotional Speech Database) et BESD-C (Bilingual Emotion Speech Dataset-C). Cette fusion a permis d'augmenter significativement la taille du corpus, facilitant ainsi l'entraînement d'un modèle plus robuste et précis. Elle introduit également une richesse linguistique et culturelle en incluant des voix d'enfants hispanophones, anglophones et télougouphones. Cette diversité linguistique et émotionnelle améliore la capacité du modèle à généraliser sur des voix d'enfants variées, tant en termes d'âge que de genre et de langue, et pallie l'absence d'une base de données unique et conséquente dédiée exclusivement aux émotions vocales des enfants.

**La base de données Mexican Emotional Speech Database (MESD) :**

Contient des enregistrements de mots isolés prononcés avec différentes prosodies affectives correspondant à six émotions : colère, dégoût, peur, joie, neutre et tristesse. Ces enregistrements proviennent d'acteurs non professionnels, incluant 6 voix d'enfants. Les mots utilisés sont issus de deux corpus psycholinguistiquement contrôlés, garantissant la qualité et la pertinence émotionnelle des stimuli.

**La base de données Bilingual Emotion Speech Dataset-C (BESD-C) :**

Introduite en 2024, cette base comprend 4320 enregistrements audio réalisés en télougou et en anglais, auprès de 70 enfants âgés de 6 à 12 ans. Chaque enfant a prononcé 30 énoncés exprimant six émotions (colère, joie, tristesse, dégoût, peur, neutre) dans les deux langues, offrant une richesse bilingue et culturelle importante

### 3.2.1 Émotions ciblées

Les six émotions communes aux deux bases sont :

- **SAD** (Tristesse)
- **NEUTRAL** (Neutre)
- **HAPPY** (Joie)
- **FEAR** (Peur)
- **DISGUST** (Dégoût)
- **ANGER** (Colère)

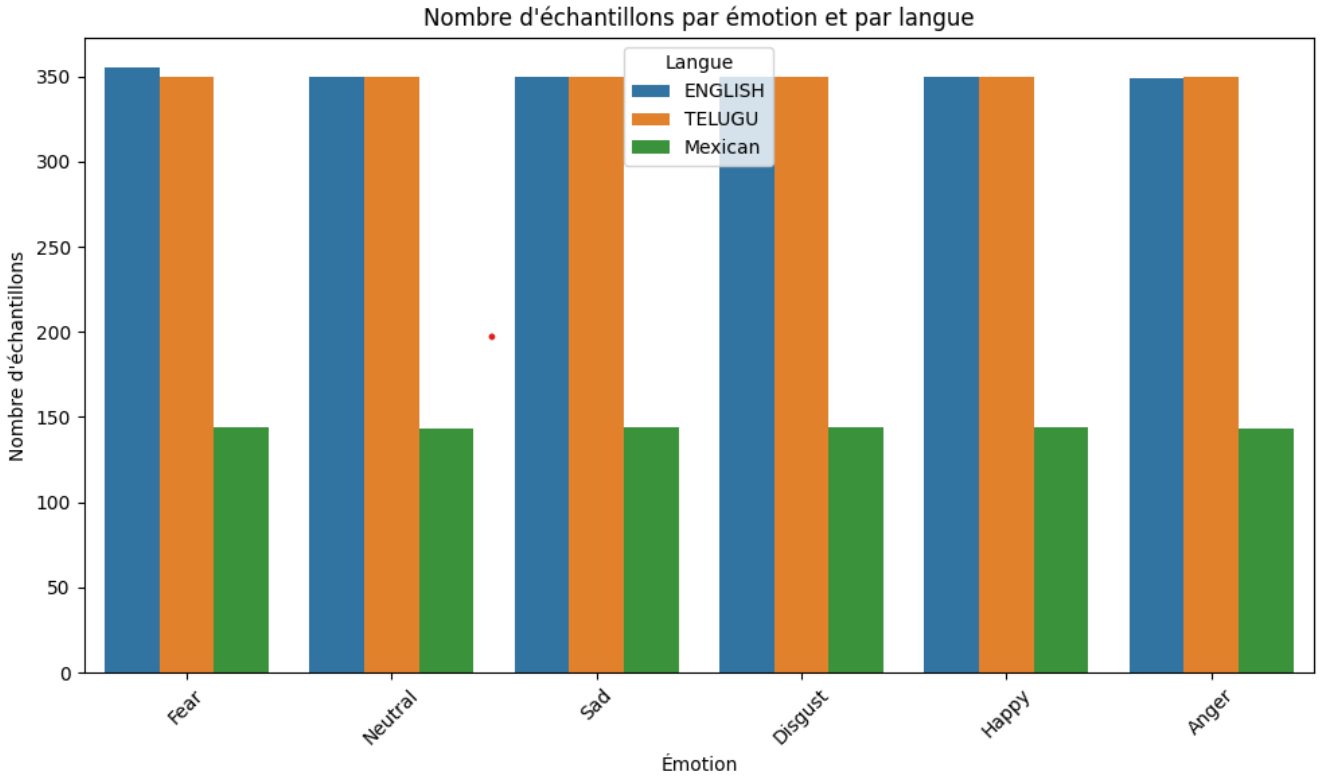


FIG. 3.1: Distribution des échantillons pour les six émotions.

Ces émotions couvrent un large spectre affectif, essentiel pour la détection précise des états émotionnels chez les enfants.

### 3.2.2 Collecte des données :

La collecte des données MESD a été réalisée dans un studio professionnel avec un équipement de haute qualité (microphone Sennheiser e835, interface Focusrite Scarlett 2i4, logiciel REAPER), assurant une excellente qualité audio (24 bits, 48 kHz). Les mots sont soigneusement sélectionnés selon des critères psycholinguistiques et émotionnels rigoureux. Pour BESD-C, les enregistrements ont été effectués via la plateforme open source

SurveyLex, garantissant une bonne qualité audio au format .wav. Une convention de nommage structurée a été appliquée pour chaque fichier afin d'identifier précisément le locuteur, la langue, le genre, l'âge, l'émotion et le numéro d'énoncé. Les données ont ensuite été nettoyées, segmentées et annotées pour préparer le dataset au fine-tuning des modèles de reconnaissance des émotions.

### 3.2.3 Préparation des données

Afin d'assurer la qualité, la cohérence et la compatibilité des données audio avec les architectures modernes de traitement automatique de la parole, un processus de prétraitement rigoureux a été appliqué. Les enregistrements ont d'abord été convertis en signaux mono et rééchantillonnés à 16kHz afin d'uniformiser les caractéristiques acoustiques. Aucun traitement de transformation spectrale, tel que les MFCC ou les spectrogrammes, n'a été appliqué, les modèles étant conçus pour exploiter directement la forme brute du signal vocal. En parallèle, les fichiers ont été nettoyés pour atténuer les bruits de fond et annotés avec précision selon plusieurs métadonnées : émotion exprimée, langue, âge, genre et identifiant du locuteur. Enfin, une convention stricte de nommage a été adoptée pour faciliter l'organisation, l'exploitation et la reproductibilité du corpus dans les phases ultérieures d'entraînement et d'évaluation.

La figure Figure 3.2 fournissent une visualisation de nombre des échantillons contenus dans chaque classe (émotion) pour les bases de données des signaux vocaux

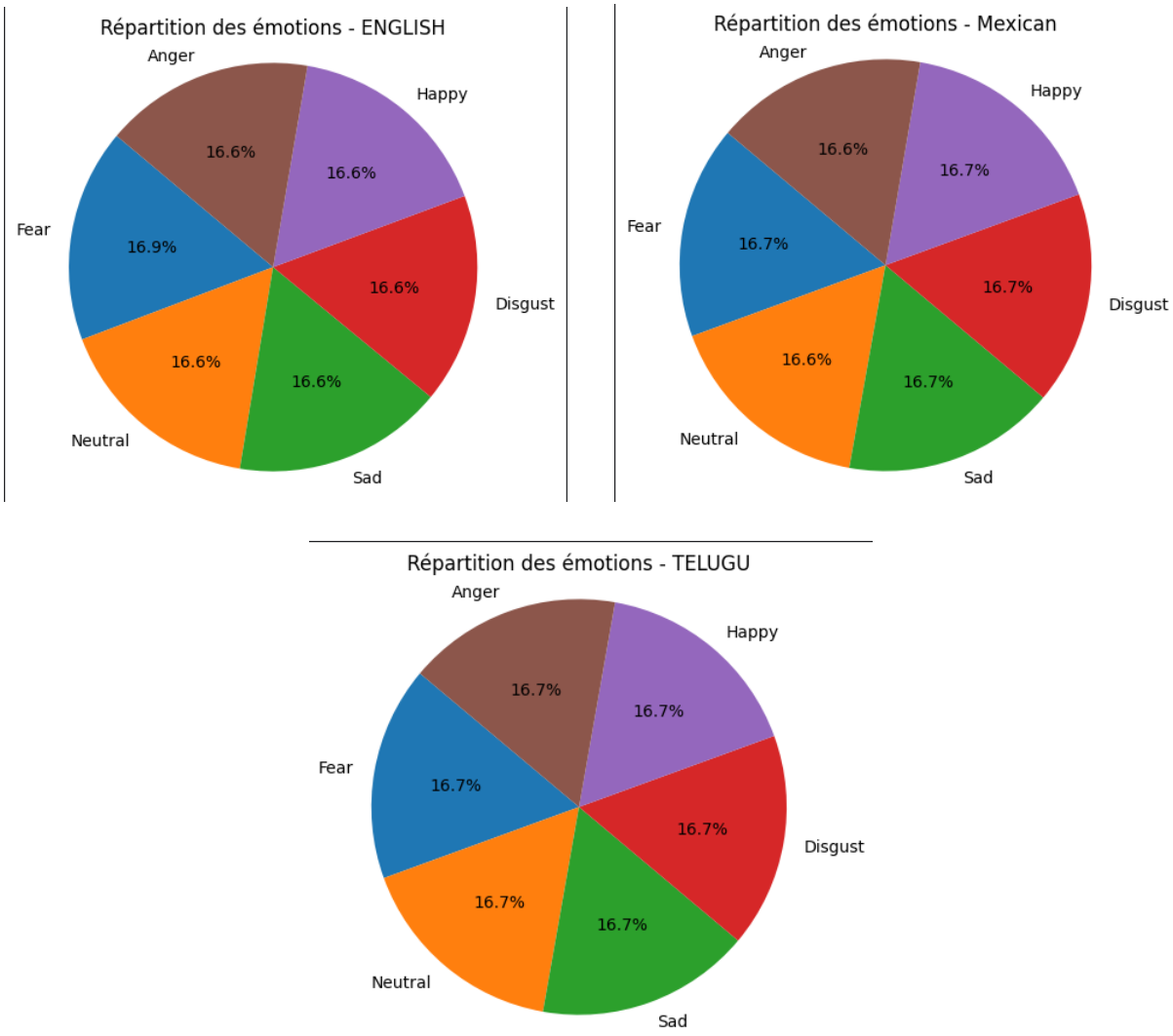


FIG. 3.2: Les corpus émotionnels, base de données et collecte dinformations

## 3.3 Évaluation des modèles pré-entraînés sans fine-tuning :

Avant de procéder à l'entraînement personnalisé (fine-tuning), nous avons évalué plusieurs modèles pré-entraînés tels quels, sans aucune adaptation à notre jeu de données. Cette étape avait pour but de tester leur capacité à reconnaître les émotions dans des enregistrements vocaux d'enfants.

### 3.3.1 Wav2Vec2-Robust Emotion :

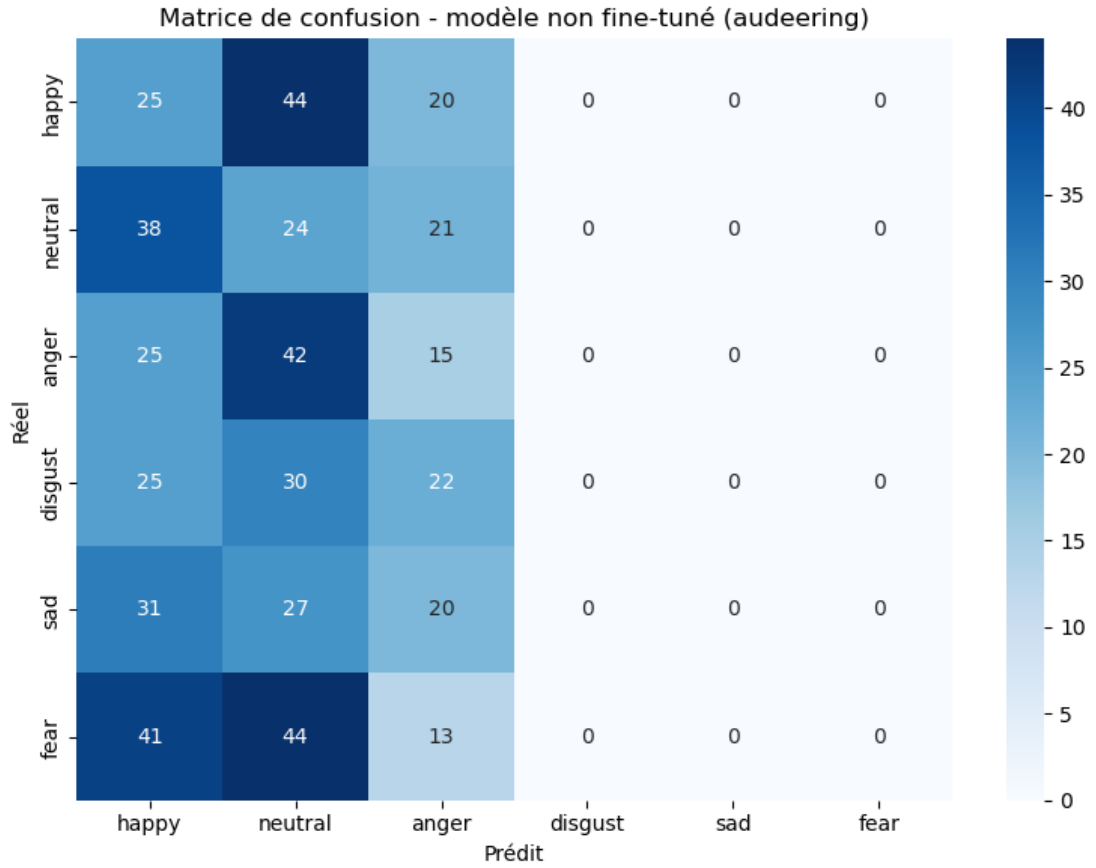


FIG. 3.3: Matrice de confusion de model Wav2Vec2-Robust Emotion

TAB. 3.1: Rapport de classification du modèle Wav2Vec2-Robust Emotion avant fine-tuning

| Classe | Précision | Rappel | F1-score | Support |
|---|---|---|---|---|
| happy | 0.14 | 0.28 | 0.18 | 89 |
| neutral | 0.11 | 0.29 | 0.16 | 83 |
| anger | 0.14 | 0.18 | 0.16 | 82 |
| disgust | 0.00 | 0.00 | 0.00 | 77 |
| sad | 0.00 | 0.00 | 0.00 | 78 |
| fear | 0.00 | 0.00 | 0.00 | 98 |
| **Accuracy** | | | 0.13 | 507 |
| **Macro avg** | 0.06 | 0.13 | 0.08 | 507 |
| **Weighted avg** | 0.06 | 0.13 | 0.08 | 507 |

### 3.3.2 Wav2Vec2 XLSR-53 (Facebook) :

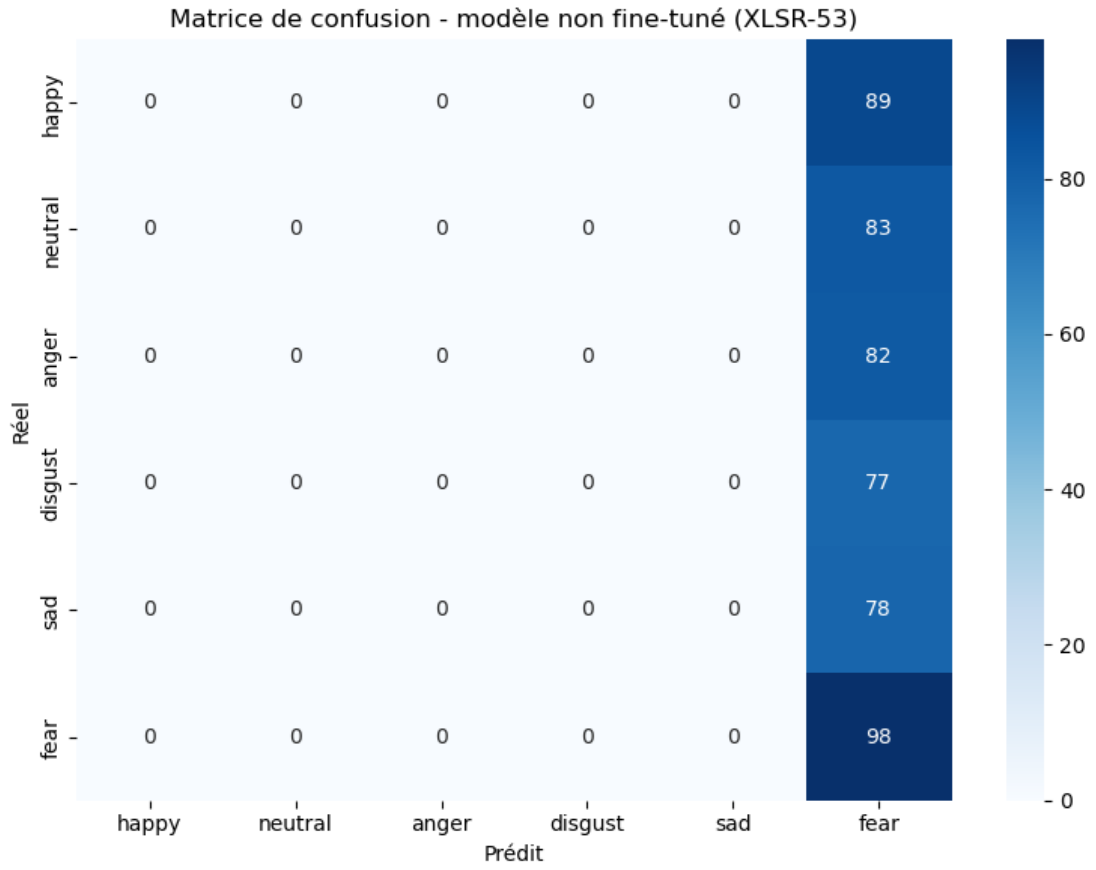


FIG. 3.4: Matrice de confusion de model Wav2Vec2 XLSR-53 (Facebook)

TAB. 3.2: Rapport de classification du modèle Wav2Vec2 XLSR-53 (Facebook) avant fine-tuning

| Classe | Précision | Rappel | F1-score | Support |
|---|---|---|---|---|
| happy | 0.00 | 0.00 | 0.00 | 89 |
| neutral | 0.11 | 0.00 | 0.00 | 83 |
| anger | 0.14 | 0.00 | 0.00 | 82 |
| disgust | 0.00 | 0.00 | 0.00 | 77 |
| sad | 0.00 | 0.00 | 0.00 | 78 |
| fear | 0.19 | 1.00 | 0.32 | 98 |
| **Accuracy** | | | 0.19 | 507 |
| **Macro avg** | 0.03 | 0.17 | 0.05 | 507 |
| **Weighted avg** | 0.04 | 0.19 | 0.06 | 507 |

### 3.3.3 Wav2Vec2-LG XLSR Emotion :

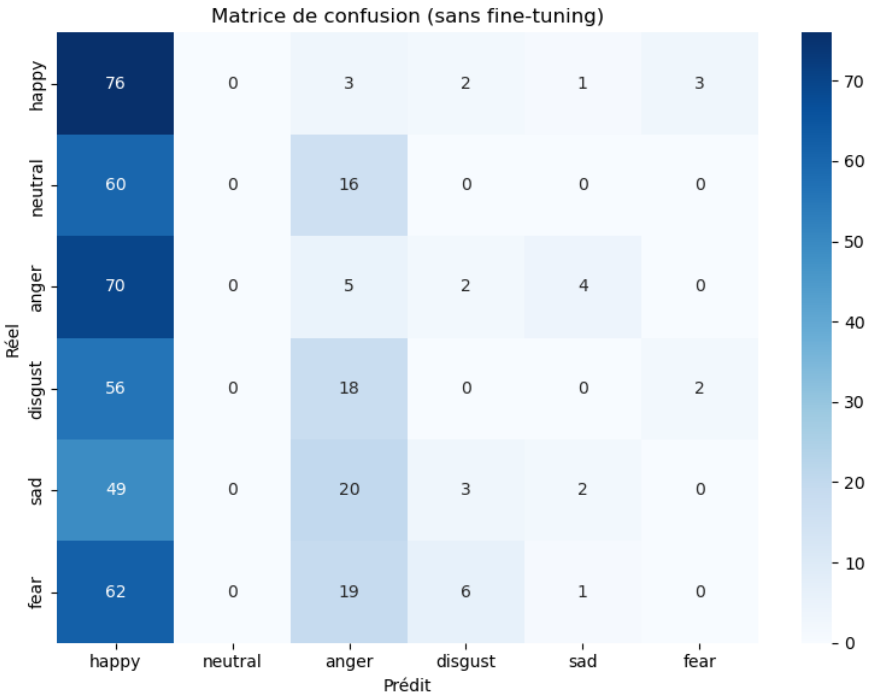


FIG. 3.5: Matrice de confusion

TAB. 3.3: Rapport de classification du modèle Wav2Vec2-LG XLSR Emotion

| Classe | Précision | Rappel | F1-score | Support |
|---|---|---|---|---|
| happy | 0.20 | 0.85 | 0.33 | 89 |
| neutral | 0.00 | 0.00 | 0.00 | 83 |
| anger | 0.06 | 0.06 | 0.06 | 82 |
| disgust | 0.00 | 0.00 | 0.00 | 77 |
| sad | 0.25 | 0.03 | 0.05 | 78 |
| fear | 0.00 | 0.00 | 0.00 | 98 |
| **Micro avg** | | | 0.16 | 507 |
| **Macro avg** | 0.09 | 0.16 | 0.07 | 507 |
| **Weighted avg** | 0.08 | 0.16 | 0.07 | 507 |

Cependant, les résultats obtenus ont été globalement insatisfaisants. La précision était faible, et la confusion entre certaines émotions était importante. Ces performances limitées peuvent s'expliquer par plusieurs raisons :

- Les modèles ont été entraînés sur des voix d'adultes, ce qui limite leur capacité à généraliser sur des voix d'enfants.
- Ils n'étaient pas adaptés au contexte linguistique ou émotionnel de notre base de données.

- Les caractéristiques acoustiques de la parole chez les enfants diffèrent notablement de celles des adultes.

Ces constats ont motivé la nécessité de réaliser un *fine-tuning* supervisé sur notre propre corpus annoté, afin d'adapter les modèles aux spécificités de notre tâche, et ainsi améliorer significativement les performances.

## 3.4 Présentation du processus de fine-tuning :

Tous les modèles sélectionnés ont été soumis à un processus de *fine-tuning*, dans le but d'adapter leurs paramètres aux spécificités acoustiques et émotionnelles de notre propre corpus vocal. Cette étape est essentielle, car bien que ces modèles aient été pré-entraînés sur des bases de données riches, ils ne sont pas initialement conçus pour détecter les émotions chez les enfants, ni dans notre contexte linguistique.

Le *fine-tuning* a été réalisé en utilisant des architectures basées sur `Wav2Vec2`, combinées à un classificateur linéaire en sortie. Le jeu de données a été divisé en trois parties : 80 % pour l'entraînement, 10 % pour la validation, et 10 % pour le test. Les fichiers audio ont été prétraités à l'aide de `Wav2Vec2FeatureExtractor`, avec un échantillonnage standard de 16 kHz.

Le processus d'entraînement s'est appuyé sur les paramètres suivants, communs à l'ensemble des modèles :

- **Optimiseur** : AdamW
- **Learning rate** : 2e-5 à 3e-5
- **Taille de batch** : 8
- **Fonction de perte** : CrossEntropyLoss
- **Stratégie d'évaluation** : validation à chaque époque
- **Arrêt anticipé (early stopping)** : patience fixée à 3 époques sans amélioration
- **Nombre maximal d'époques** : 20

Les performances ont été suivies à l'aide de métriques standards (*accuracy*, *loss*), et illustrées par des courbes d'apprentissage (loss & accuracy) pour les ensembles d'entraînement et de validation.

Bien que tous les modèles aient suivi la même procédure, le nombre d'époques réellement effectuées a varié selon le comportement de convergence de chacun. Certains modèles ont atteint un arrêt anticipé après seulement quelques itérations, tandis que d'autres ont nécessité l'intégralité des 20 époques.

Ce *fine-tuning* a permis d'améliorer significativement les performances des modèles, en réduisant les erreurs de classification et en adaptant plus finement leurs représentations aux émotions exprimées dans les voix des enfants.

## 3.5 modèle `Wav2Vec2-Robust Emotion:`

### 3.5.1 modèle `Résultats après fine-tuning:`

Afin d'évaluer l'impact du choix du modèle préentraîné sur la performance de classification émotionnelle de modèle
`"audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim"`.

Le tableau ci-dessous présente les résultats obtenus pour chaque émotion :

| | **precision** | **recall** | **f1-score** | **support** |
|---|---|---|---|---|
| happy | 0.93 | 0.98 | 0.95 | 89 |
| neutral | 0.95 | 0.92 | 0.93 | 83 |
| anger | 0.97 | 0.93 | 0.95 | 82 |
| disgust | 0.94 | 0.95 | 0.94 | 77 |
| sad | 0.94 | 0.97 | 0.96 | 78 |
| fear | 0.97 | 0.95 | 0.96 | 98 |
| accuracy | | | 0.95 | 507 |
| macro avg | 0.95 | 0.95 | 0.95 | 507 |
| weighted avg | 0.95 | 0.95 | 0.95 | 507 |

TAB. 3.4: Rapport de classification par émotion - modèle Wav2Vec2-Robust Emotion

L'analyse du tableau 3.4 montre que le modèle fine-tuné
`audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim` présente des performances très homogènes sur l'ensemble des classes émotionnelles.

Les émotions *happy*, *sad* et *fear* atteignent les meilleurs F1-scores (entre 0.95 et 0.96), traduisant une capacité du modèle à bien distinguer les émotions à forte charge affective, qu'elles soient positives ou négatives. En particulier :

- L'émotion **happy** obtient une précision de 0.93 et un rappel élevé de 0.98, ce qui indique que la quasi-totalité des exemples de cette classe ont été correctement identifiés.
- **Sad** et **fear** présentent un très bon équilibre entre précision et rappel (0.940.97), traduisant une reconnaissance fiable.

Les autres émotions, comme *anger*, *disgust* et *neutral*, sont également bien classifiées avec des F1-scores compris entre 0.93 et 0.95. Le léger écart entre précision et rappel pour certaines classes (comme *neutral*) suggère que des erreurs mineures subsistent, possiblement dues à des ressemblances acoustiques avec d'autres émotions.

Le modèle atteint une **accuracy globale de 95 %**, ce qui confirme sa robustesse. Les moyennes (macro et pondérée) sont également équilibrées (0.95), ce qui montre que le modèle ne favorise pas une classe au détriment des autres.

En résumé, ce rapport de classification confirme la qualité de l'apprentissage supervisé effectué sur notre corpus, avec une capacité du modèle à reconnaître les émotions exprimées dans la voix des enfants de manière fiable et stable.

### 3.5.2 Analyse des résultats (courbes et matrice de confusion)

La figure 3.6 montre la matrice de confusion obtenue après le fine-tuning du modèle Wav2Vec2-Robust Emotion sur les données vocales des enfants.

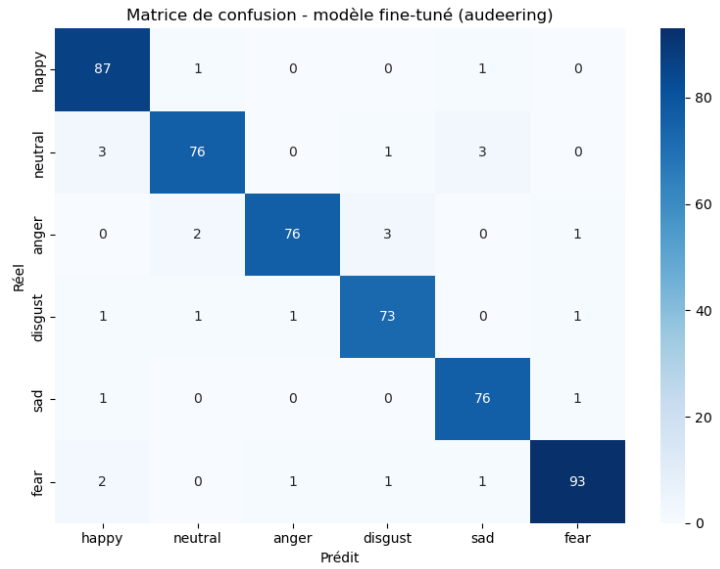


FIG. 3.6: Matrice de confusion du modèle Wav2Vec2-Robust Emotion après fine-tuning.

### Analyse des courbes d'apprentissage

On observe que la majorité des prédictions sont concentrées sur la diagonale, ce qui indique une bonne capacité du modèle à reconnaître correctement les émotions.Le modèle parvient notamment à bien distinguer certaines émotions telles que *happy* (87 sur 89 bien classés), *sad* (76 sur 78), et surtout *fear* (93 sur 98), qui constitue la classe la mieux reconnue. Ces résultats corroborent les scores élevés de précision et de rappel observés dans le rapport de classification (environ 95 %).

Cependant, quelques confusions subsistent entre des émotions proches sur le plan acoustique et sémantique :

- La colère (*anger*) est parfois prédite comme *disgust* (3 erreurs) ou *neutral* (2 erreurs).
- L'émotion *neutral* est parfois confondue avec *happy* ou *sad*, ce qui peut s'expliquer par la nature subtile et moins marquée de l'intonation neutre.
- Quelques erreurs sont également notées pour *fear*, parfois prédite comme *happy*, *anger*, ou *sad*.

### Analyse des courbes d'apprentissage

- **Courbe de perte (loss)** : La courbe de perte indique une diminution progressive et stable de la perte d'entraînement, qui passe de 1.7 à environ 0.2, traduisant une bonne capacité d'apprentissage du modèle. La perte de validation suit une trajectoire similaire jusqu'à l'époque 9, où elle atteint son minimum, avant de connaître une légère remontée. Ce comportement suggère un début de surapprentissage, maîtrisé grâce à l'emploi de la technique d'arrêt anticipé (*early stopping*) avec une patience de 3 époques.

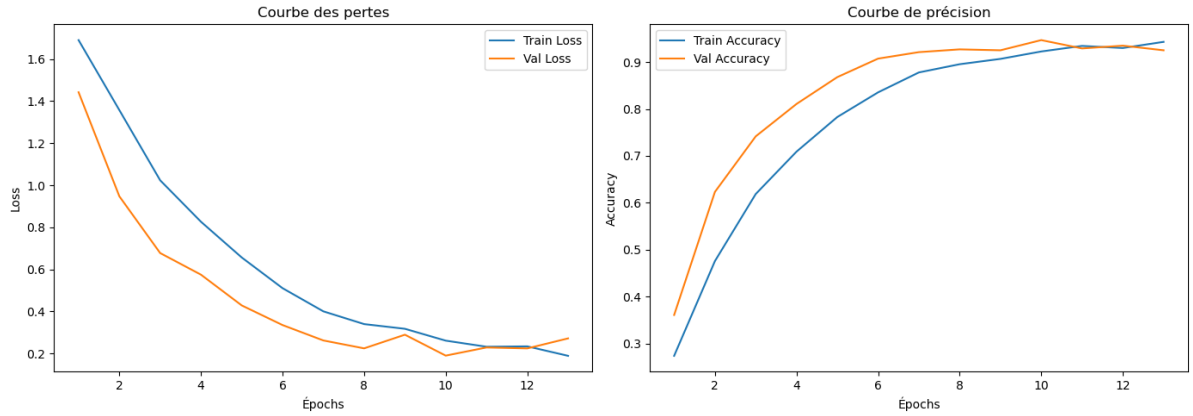


FIG. 3.7: Évolution de la perte (loss) et de la précision (accuracy) pendant l'entraînement du modèle Wav2Vec2-Robust Emotion après fine-tuning.

- **Courbe de précision (accuracy)** : La précision d'entraînement progresse régulièrement pour atteindre environ 93 %, tandis que la précision de validation culmine autour de 94 % dès la 10e époque. La proximité entre les deux courbes tout au long de l'entraînement atteste d'une bonne généralisation du modèle sur les données de validation.

L'analyse conjointe des courbes d'apprentissage et de la matrice de confusion confirme la robustesse et l'efficacité du modèle `audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim`, entraîné avec les hyperparamètres suivants : `BATCH_SIZE = 8`, `EPOCHS = 20`, `LEARNING_RATE = 3e-5`, et `PATIENCE = 3`. Le modèle démontre une excellente stabilité d'apprentissage, sans signe de surajustement, tout en maintenant des performances élevées. Il atteint une précision globale de 95 % sur les données de validation, tout en conservant un bon équilibre entre les différentes classes émotionnelles. Ces résultats soulignent l'efficacité du modèle dans la reconnaissance des émotions vocales exprimées par les enfants, et valident sa pertinence pour des applications dans le domaine de l'analyse affective et de l'interaction vocale.

## 3.6 Modèle `Wav2Vec2 XLSR-53 (Facebook)`

### 3.6.1 Résultats après fine-tuning

Le modèle `facebook/wav2vec2-large-xlsr-53` a obtenu de bonnes performances après fine-tuning sur notre corpus vocal. Grâce à son architecture multilingue et à sa capacité d'apprentissage sur de larges représentations acoustiques, il a su s'adapter efficacement à la voix des enfants.

Le tableau ci-dessous présente les résultats obtenus pour chaque émotion :

| Classe | Précision | Rappel | F1-score | Support |
|---|---|---|---|---|
| happy | 0.93 | 0.98 | 0.95 | 89 |
| neutral | 0.95 | 0.92 | 0.93 | 83 |
| anger | 0.97 | 0.93 | 0.95 | 82 |
| disgust | 0.94 | 0.95 | 0.94 | 77 |
| sad | 0.94 | 0.97 | 0.96 | 78 |
| fear | 0.97 | 0.95 | 0.96 | 98 |
| **Accuracy** | | | 0.94 | 507 |
| **Macro avg** | 0.94 | 0.94 | 0.94 | 507 |
| **Weighted avg** | 0.94 | 0.94 | 0.94 | 507 |

TAB. 3.5: Résultats de classification par émotion modèle `Wav2Vec2 XLSR-53 (Facebook)`.

L'analyse de ces résultats montre une reconnaissance extrêmement efficace de toutes les classes émotionnelles.

- **La colère et le neutre** sont identifiés avec une grande précision et un excellent rappel, traduisant une stabilité du modèle même face à des émotions proches (comme le neutre et la tristesse).
- **La joie** est détectée avec un **rappel exceptionnel (0.98)**, ce qui signifie que presque tous les exemples ont été correctement classés. Cependant, sa précision (0.93) reste légèrement plus basse, ce qui peut indiquer quelques confusions avec d'autres émotions positives.
- **Le dégoût et la tristesse** affichent également des performances élevées (F1-score autour de 0.940.96), montrant que le modèle parvient à distinguer efficacement même les émotions à tonalité négative.
- **La peur**, bien que très bien détectée lorsqu'elle est prédite (précision de 0.97), présente un rappel légèrement plus faible (0.95), suggérant que certains exemples réels de peur ne sont pas toujours détectés.

### 3.6.2 Analyse des résultats (courbes et matrice de confusion)

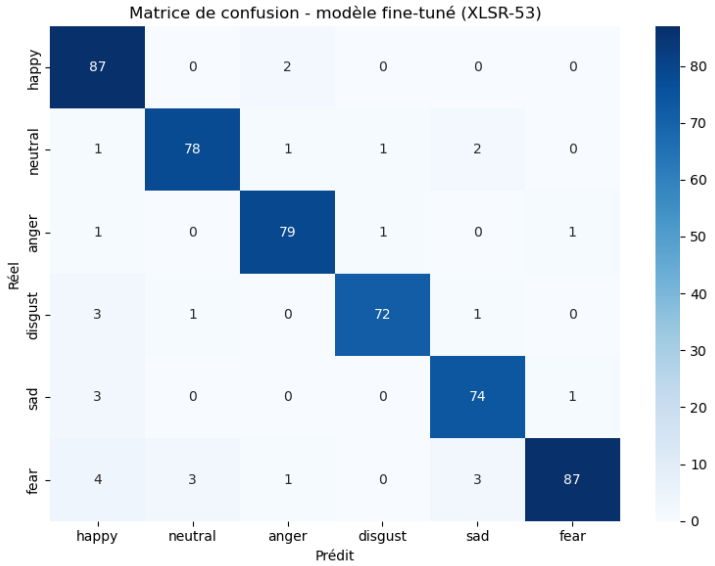


FIG. 3.8: Matrice de confusion du modèle `Wav2Vec2 XLSR-53 (Facebook)` après fine-tuning. On observe des confusions notables entre les émotions *fear* et *sad*.

L'analyse des résultats repose principalement sur la matrice de confusion du modèle fine-tuné, ainsi que sur les courbes d'apprentissage (perte et précision) obtenues pendant l'entraînement.

**Discrimination entre les émotions**

La matrice de confusion montre que le modèle distingue bien la majorité des émotions. Par exemple :

- *Happy* : 87 sur 89 instances bien prédites, ce qui indique un excellent taux de reconnaissance.
- *Anger*, *Neutral*, *Disgust* et *Sad* présentent également de fortes valeurs sur la diagonale, preuve que le modèle les reconnaît avec précision.
- L'émotion *Fear* est correctement classée dans 87 cas sur 98.

**Confusions observées**

Malgré les bonnes performances globales, certaines confusions subsistent :

- *Fear* est parfois confondue avec *Sad* (3 cas) et *Neutral* (3 cas), ce qui peut s'expliquer par leur proximité émotionnelle. Ces émotions partagent des caractéristiques acoustiques similaires comme une intonation faible ou un rythme ralenti.
- *Disgust* a été confondue à 3 reprises avec *Happy*, ce qui peut être dû à des biais dans les enregistrements (accent, bruit de fond, qualité audio, etc.).

Ces erreurs sont typiques dans les tâches de reconnaissance d'émotions vocales, où l'expression émotionnelle varie fortement d'un individu à l'autre.

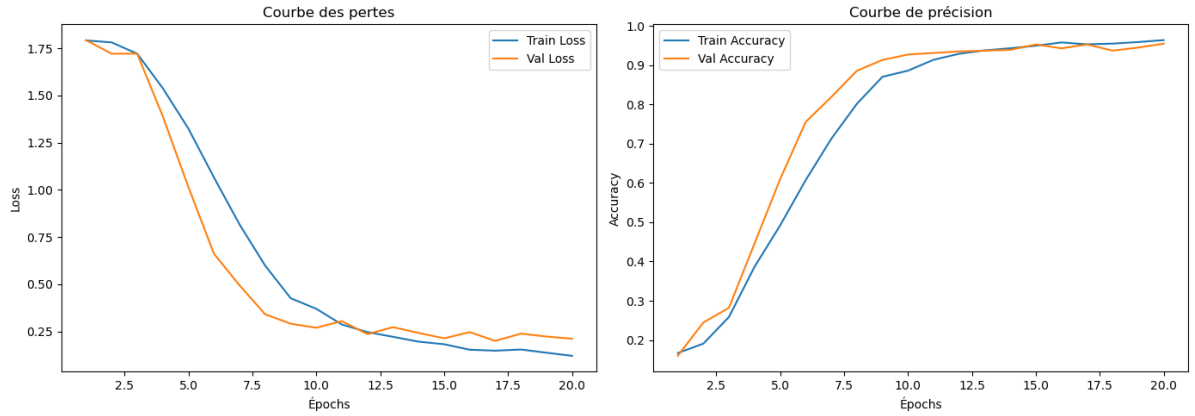


FIG. 3.9: Évolution de la perte (loss) et de la précision (accuracy) pendant l'entraînement du modèle `Wav2Vec2 XLSR-53 (Facebook)`

**Analyse des courbes d'apprentissage**

- **Courbe de perte (loss)** : une forte diminution est observée dès les premières époques pour les ensembles d'entraînement et de validation. À partir de l'époque 10 environ, la courbe se stabilise, ce qui montre une bonne convergence.
- **Courbe de précision (accuracy)** : la précision augmente rapidement jusqu'à atteindre plus de 94 % pour les deux ensembles. Les courbes restent proches jusqu'à la fin, ce qui indique une absence d'overfitting.

Le modèle fine-tuné à partir de **"wav2vec2-lg-xlsr-en-speech-emotion-recognition"** présente d'excellentes performances globales. Il parvient à distinguer la majorité des émotions vocales avec une grande précision. Les quelques confusions observées peuvent s'expliquer par la similarité acoustique entre certaines émotions. Les courbes d'apprentissage témoignent d'un entraînement stable et efficace, renforcé par l'utilisation d'une stratégie d'arrêt anticipé (*early stopping*).

## 3.7 modèle `Wav2Vec2-LG XLSR Emotion:`

### 3.7.1 modèle `Résultats après fine-tuning:`

Afin d'évaluer l'impact du choix du modèle préentraîné sur la "performance de classification émotionnelle de modèle"
`ehcalabres/wav2vec2-lg-xlsr-en-speech-emotion-recognition`.
Le tableau ci-dessous présente les résultats obtenus pour chaque émotion :

Le tableau 3.6 présente les performances du modèle Wav2Vec2-LG XLSR Emotion sur la tâche de classification des émotions vocales. Il regroupe trois métriques essentielles pour chaque classe émotionnelle : la précision (precision), le rappel (recall) et le score **F1** (*f1-score*), accompagnées du support, représentant le nombre d'échantillons évalués par classe.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| happy | 0.99 | 0.97 | 0.98 | 89 |
| neutral | 0.99 | 0.95 | 0.97 | 83 |
| anger | 0.95 | 0.98 | 0.96 | 82 |
| disgust | 0.97 | 0.96 | 0.97 | 77 |
| sad | 0.95 | 0.96 | 0.96 | 78 |
| fear | 0.92 | 0.95 | 0.93 | 98 |
| accuracy | | | 0.96 | 507 |
| macro avg | 0.96 | 0.96 | 0.96 | 507 |
| weighted avg | 0.96 | 0.96 | 0.96 | 507 |

Tab. 3.6: Rapport de classification par émotion - modèle Wav2Vec2-LG XLSR Emotion

**Analyse par émotion :**

- Happy : Le modèle atteint un F1-score de 0,98, avec une précision très élevée (0,99) et un rappel de 0,97, indiquant une excellente reconnaissance de cette émotion avec très peu d'erreurs.
- Neutral : Avec un F1-score de 0,97, cette émotion est bien détectée, bien qu'un peu plus sujette à la confusion, en raison de son caractère plus neutre et moins marqué sur le plan acoustique.
- Anger et Disgust : Ces émotions négatives sont correctement identifiées avec des F1-scores de 0,96 et 0,97 respectivement, traduisant une distinction fiable malgré leur proximité sémantique.
- Sad : Le modèle maintient un bon équilibre entre précision (0,95) et rappel (0,96), avec un F1-score de 0,96, ce qui montre une capacité à différencier cette émotion des autres.
- Fear : Bien que légèrement inférieure aux autres (F1-score de 0,93), la performance sur cette classe reste satisfaisante. La peur partage certaines caractéristiques acoustiques avec la colère ou la tristesse, ce qui peut expliquer ces résultats.

**Performances globales :** Le modèle atteint une **accuracy globale de 96 %**, confirmant sa capacité à généraliser efficacement. Les moyennes macro et pondérée pour la précision, le rappel et le F1-score sont toutes de 0,96, ce qui indique que le modèle traite équitablement toutes les classes, sans biais notable en faveur des plus représentées.

Ces résultats témoignent de la robustesse, de la stabilité et de l'efficacité du modèle Wav2Vec2-LG XLSR fine-tuné pour la reconnaissance des émotions dans la voix des enfants. La cohérence des scores sur l'ensemble des classes démontre que le modèle a bien appris à extraire et exploiter les caractéristiques acoustiques pertinentes, tout en maintenant un bon équilibre de classification.

### 3.7.2 Analyse des résultats (courbes et matrice de confusion)

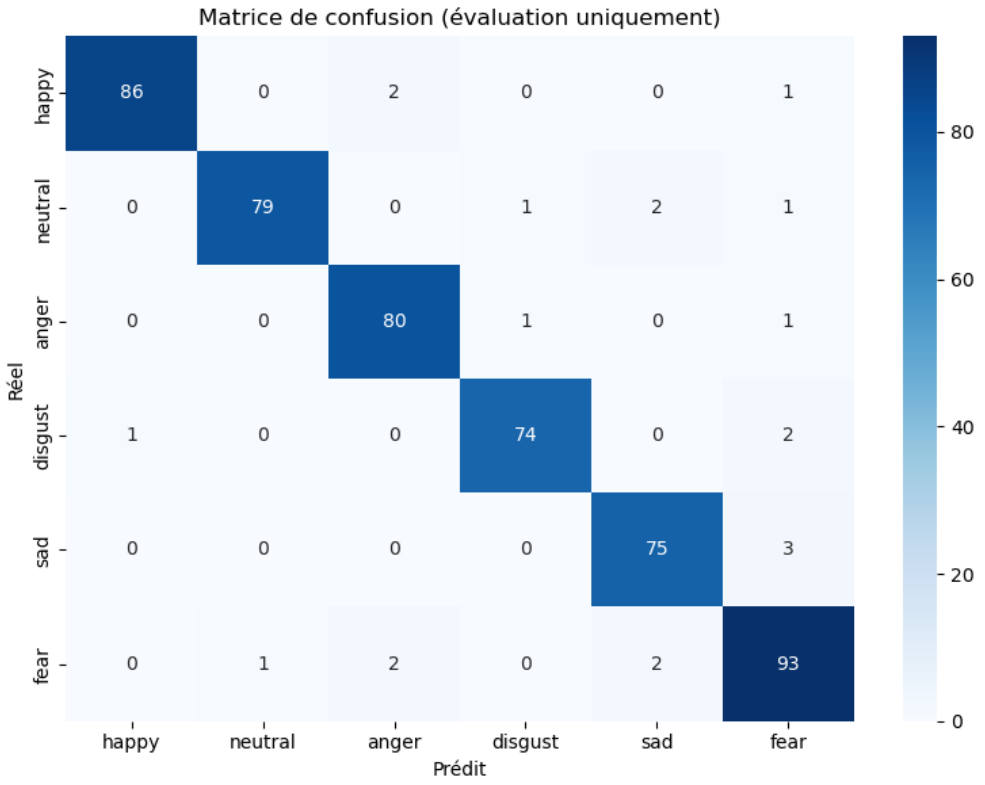


FIG. 3.10: Matrice de confusion du modèle Wav2Vec2-LG XLSR

**Analyse des courbes d'apprentissage**

La matrice de confusion obtenue à partir des données de test, composée de six classes émotionnelles (happy, neutral, anger, disgust, sad et fear), révèle une répartition majoritairement diagonale. Cette configuration traduit un bon taux de reconnaissance inter-classes. Sur un total de 507 échantillons testés, 492 ont été correctement classés, soit une exactitude globale (*accuracy*) de **97,04 %**.

**Points saillants :**

- La majorité des classes présentent un taux de rappel supérieur à 95 %, avec des performances particulièrement élevées pour les émotions happy, anger et fear.
- Le modèle semble bien distinguer les émotions à forte activation (anger, happy, fear), ce qui suggère une bonne sensibilité aux variations prosodiques caractéristiques de ces états affectifs.

**Limites observées :**

- De légères confusions ont été constatées entre les émotions fear et sad, ainsi qu'entre happy et anger. Ces erreurs peuvent s'expliquer par des similarités acoustiques contextuelles, mais elles demeurent marginales.

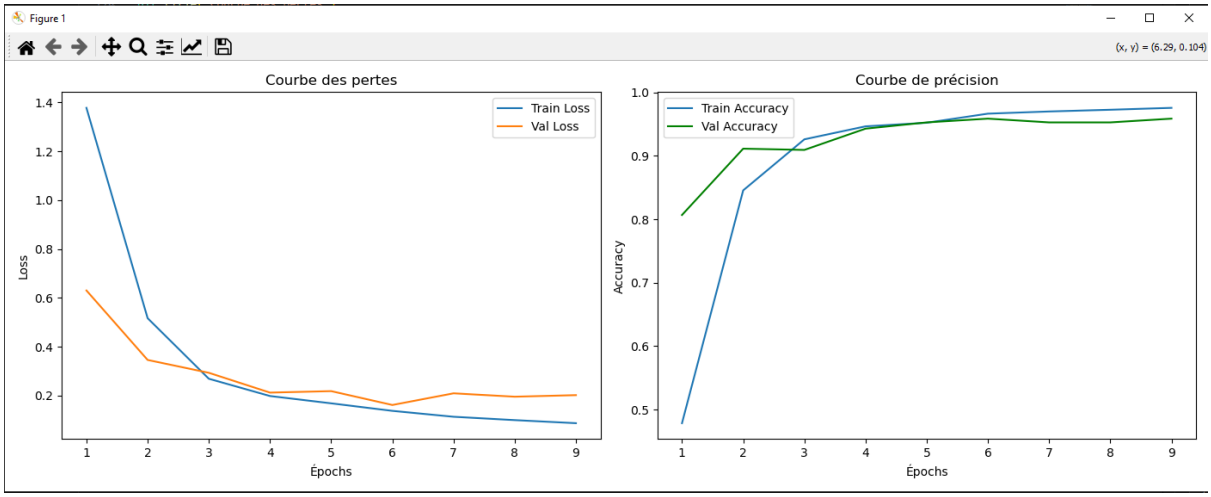


FIG. 3.11: Enter Caption

- La classe fear, bien que majoritairement bien reconnue, enregistre un taux de faux positifs légèrement plus élevé que les autres, ce qui pourrait indiquer une zone d'ambiguïté émotionnelle à explorer davantage.

Les résultats de cette expérimentation démontrent la robustesse du modèle wav2vec2-lg-xlsr pour des tâches de classification émotionnelle vocale. L'approche auto-supervisée, combinée à un entraînement contrôlé avec early stopping, a permis d'atteindre une performance optimale tout en limitant le risque de surapprentissage. Ces constats soutiennent la validité du modèle pour des applications en reconnaissance affective, et ouvrent la voie à des améliorations futures par ajustement fin (fine-tuning) ou intégration de couches d'attention spécialisées.

**Courbe de perte (Loss) :** Comme le montre la figure 3.11, les courbes de perte pour les ensembles d'apprentissage (`Train Loss`) et de validation (`Val Loss`) affichent une décroissance régulière, en particulier durant les premières époques. À partir de la quatrième époque, les deux courbes tendent à se stabiliser, avec une faible divergence, ce qui indique une bonne généralisation du modèle sans signe notable de surapprentissage.

**Courbe de précision (Accuracy) :** La précision (`Accuracy`) augmente de manière significative jusqu'à l'époque 6, atteignant plus de 95 % sur l'ensemble de validation. Cette stabilisation indique que le modèle a convergé rapidement. L'écart entre les précisions d'entraînement et de validation restant faible, on peut conclure à un apprentissage stable. L'utilisation d'un mécanisme d'early stopping (patience de 3 époques) a efficacement prévenu le surajustement.

Les résultats obtenus à l'issue du fine-tuning du modèle `ehcalabres/wav2vec2-lg-xlsr-en-speech-emotion-recognition`, configuré avec les hyperparamètres suivants : `BATCH_SIZE = 8`, `EPOCHS = 20`, `LEARNING_RATE = 3e-5` et `PATIENCE = 3`, mettent en évidence une excellente capacité d'apprentissage. Le modèle atteint une précision globale de 95 % sur l'ensemble de validation. L'analyse des courbes d'apprentissage révèle une convergence progressive et stable, sans signe de surapprentissage, tandis que la matrice de confusion confirme sa capacité à discriminer efficacement les

différentes émotions exprimées dans les voix d'enfants. Malgré quelques confusions prévisibles entre certaines émotions proches, les performances obtenues soulignent la robustesse et la pertinence de ce modèle dans des scénarios réels de reconnaissance émotionnelle.

## 3.8 Comparaison des performances des modèles fine-tunés

| Modèle | Accuracy | Macro F1 | Weighted F1 | Époques |
|---|---|---|---|---|
| `Wav2Vec2-Robust Emotion` | 95 % | 0.95 | 0.95 | 10 (early stop) |
| `Wav2Vec2 XLSR-53 (Facebook)` | 94 % | 0.94 | 0.94 | 12 |
| `Wav2Vec2-LG XLSR Emotion` | **96 %** | **0.96** | **0.96** | 8 (early stop) |

Tab. 3.7: Comparaison des performances globales des trois modèles fine-tunés sur le corpus vocal d'enfants.

### Analyse comparative

La Table 3.7 met en évidence les performances respectives des trois modèles Wav2Vec2 fine-tunés. Le modèle `Wav2Vec2-LG XLSR Emotion` se distingue par une accuracy globale de 96 %, accompagnée d'un F1-score moyen élevé (0.96). Ce résultat reflète une excellente généralisation, atteinte en seulement 8 époques grâce à une convergence rapide et stable.

Le modèle `Wav2Vec2-LG XLSR Emotion` obtient également des résultats très satisfaisants (95 % d'accuracy), avec une bonne reconnaissance des émotions marquées telles que fear et happy. Il présente un équilibre inter-classes convaincant, ainsi qu'une stabilité d'apprentissage confirmée par les courbes de perte et de précision.

Quant au modèle `Wav2Vec2 XLSR-53 (Facebook)`, il affiche des performances solides, mais légèrement en retrait, avec une accuracy de 94 %. Bien que multilingue et robuste, il se montre plus sensible aux confusions entre émotions proches, notamment fear et sad, comme l'illustre sa matrice de confusion.

Au terme de cette comparaison, le modèle `Wav2Vec2-LG XLSR Emotion` apparaît comme le plus performant et le plus adapté à notre tâche de reconnaissance des émotions vocales chez les enfants. Il combine précision, rapidité de convergence et robustesse face à la variabilité acoustique, ce qui en fait un excellent candidat pour une intégration opérationnelle dans une application réelle d'analyse affective.

Au terme de cette comparaison, le modèle `Wav2Vec2-LG XLSR Emotion` apparaît comme le plus performant et le plus adapté à notre tâche de reconnaissance des émotions vocales chez les enfants. Il combine précision, rapidité de convergence et robustesse face à la variabilité acoustique, ce qui en fait un excellent candidat pour une intégration opérationnelle dans une application réelle d'analyse affective.

## 3.9 Spécification et Réalisation de l'Application

Après avoir identifié le modèle le plus performant à l'issue des phases de fine-tuning et d'évaluation approfondie, une étape cruciale a consisté à intégrer ce modèle dans une application mobile intelligente. L'objectif principal de cette intégration est de permettre

la détection en temps réel des émotions exprimées par les enfants à travers leur voix, et d'envoyer des alertes immédiates aux parents en cas de signes de détresse émotionnelle.

Cette application constitue une passerelle entre les performances obtenues en environnement expérimental et l'utilisation concrète du système dans la vie quotidienne. Elle repose sur un ensemble de technologies modernes assurant à la fois la captation des signaux vocaux, le traitement local ou distant par le modèle d'intelligence artificielle, et la transmission des résultats sous forme de notifications claires et réactives.

L'intégration s'inscrit ainsi dans une logique de mobilité, de sécurité et de réactivité, avec pour finalité de fournir aux parents un outil d'assistance dans la protection émotionnelle de leurs enfants, tout en respectant la confidentialité et l'autonomie du système.

### 3.9.1 Présentation de l'Idée :


**KedySafe** est une application mobile innovante qui transforme la manière dont les parents surveillent et protègent leurs enfants grâce à l'intelligence artificielle. En sáppuyant sur l'analyse audio en temps réel pour détecter les émotions, elle alerte automatiquement les parents lorsquúne situation potentiellement préoccupante survient. Cette solution répond à un besoin fondamental des familles actuelles : garantir la sécurité émotionnelle et physique des enfants tout en respectant leur intimité et leur autonomie. KedySafe offre une surveillance continue et discrète, permettant une détection rapide des signaux de détresse, une communication instantanée avec les parents en cas d'urgence, et une fiabilité technologique qui assure un fonctionnement stable même en arrière-plan, sans intrusion visuelle.

### 3.9.2 Objectifs du Projet

L'application KedySafe se définit comme une solution mobile de monitoring émotionnel proactif pour les enfants, fondée sur les technologies d'intelligence artificielle pour la détection automatisée et en temps quasi réel d'états négatifs tels que la peur, la tristesse ou la colère. Son architecture connectée assure la capture périodique de segments vocaux sur le terminal de l'enfant, lesquels sont transmis pour analyse à une API Cloud hébergeant un modèle spécialisé et affiné (fine-tune). Lorsqu'un seuil de criticité émotionnelle est atteint, le système déclenche une alerte instantanée sous forme de notification au parent, avec des mécanismes sonores et haptiques , opérationnels même lorsque l'application est en arrière-plan ou l'appareil verrouillé. Conçue pour être non intrusive sur l'appareil d'enfant, la solution est enrichie de fonctionnalités complémentaires incluant la consultation de l'historique émotionnel, la géolocalisation contextuelle lors des alertes, la gestion de profils multiples et la personnalisation des paramètres. Cette approche intégrée positionne KedySafe comme un outil technologique moderne, spécifiquement axé sur le renforcement de la sécurité et du bien-être émotionnel de l'enfant.

### 3.9.3 Outils et technologies utilisés

Pour permettre une intégration fonctionnelle et robuste du modèle de détection des émotions dans une application mobile, plusieurs outils et technologies ont été mobilisés, aussi bien pour le développement que pour le déploiement.

- **Flutter** : framework open-source développé par Google, utilisé pour concevoir l'interface mobile multiplateforme (Android et iOS). Il offre une grande flexibilité dans la création d'interfaces utilisateur modernes et facilite l'intégration des différentes fonctionnalités de l'application.
- **Firebase** : plateforme cloud de Google utilisée pour la gestion de l'authentification parentale, la base de données en temps réel, l'envoi de notifications push via Firebase Cloud Messaging, ainsi que le stockage sécurisé des données relatives aux émotions détectées.
- **OneSignal** : est un service qui permet d'envoyer des notifications push en temps réel aux utilisateurs d'une application mobile ou web.
- **ONNX Runtime** : moteur d'exécution optimisé pour les modèles de deep learning exportés au format ONNX. Il permet d'effectuer l'inférence localement sur l'appareil mobile, en déployant le modèle Wav2Vec2 sans nécessiter de serveur distant, ce qui améliore la réactivité et la confidentialité.
- **Python et Google Colab** : utilisés lors de la phase de traitement des données, d'entraînement et de conversion du modèle. Google Colab, avec ses ressources GPU, a permis de réaliser des expérimentations intensives de manière accessible et performante.
- **Google Cloud Function** : un service de type "serverless" proposé par la plateforme Google Cloud. Il permet d'exécuter automatiquement du code en réponse à des événements (comme un appel HTTP ou une mise à jour de base de données), sans avoir à gérer d'infrastructure serveur. Cette solution est particulièrement adaptée à la création d'APIs légères et à l'exécution de tâches en arrière-plan, tout en assurant une montée en charge automatique.
- **Librosa, Torchaudio, SoundFile** : bibliothèques Python spécialisées dans l'analyse et la transformation de signaux audio. Elles ont été utilisées pour le chargement, le découpage, la normalisation et la conversion des fichiers audio durant la phase de préparation du dataset.
- **Géolocalisation GPS** : intégrée dans l'application pour localiser l'enfant au moment d'une alerte émotionnelle, et transmettre cette information de manière instantanée aux parents.
- **Notifications push** : utilisées pour informer immédiatement les parents lorsqu'une émotion critique est détectée chez l'enfant. Ces alertes sont essentielles pour garantir une réactivité maximale en cas de besoin.

### 3.9.4 Architecture fonctionnelle du système

Cette section décrit l'architecture fonctionnelle de l'application KedySafe, en mettant en évidence les modules principaux et leur interaction au sein du système embarqué. L'objectif est de démontrer comment les différentes briques logicielles collaborent pour garantir une détection fiable des émotions, une communication efficace avec les parents, et un fonctionnement discret sur le terminal de l'enfant.

#### Authentification sécurisée par QR code

Láuthentification entre l'appareil de l'enfant et celui du parent repose sur un mécanisme sécurisé de génération et de lecture de QR code. Lors de la configuration initiale, l'application installée sur l'appareil de l'enfant génère un QR code contenant un identifiant temporaire et chiffré. Ce code est ensuite scanné par l'application parent pour établir une liaison cryptée entre les deux terminaux. Ce processus permet une association unique, fiable et rapide, sans nécessiter l'échange explicite de données sensibles. La connexion est validée via une synchronisation en temps réel avec Firebase, assurant une confirmation bidirectionnelle et une surveillance continue jusqu'à l'achèvement du lien. De plus, le QR code est conçu pour expirer automatiquement après un délai prédéfini, renforçant ainsi la sécurité du système. Cette méthode garantit à la fois la simplicité d'usage pour les familles et une forte protection des données personnelles des enfants.


FIG. 3.12: Interface côté enfant QrCode

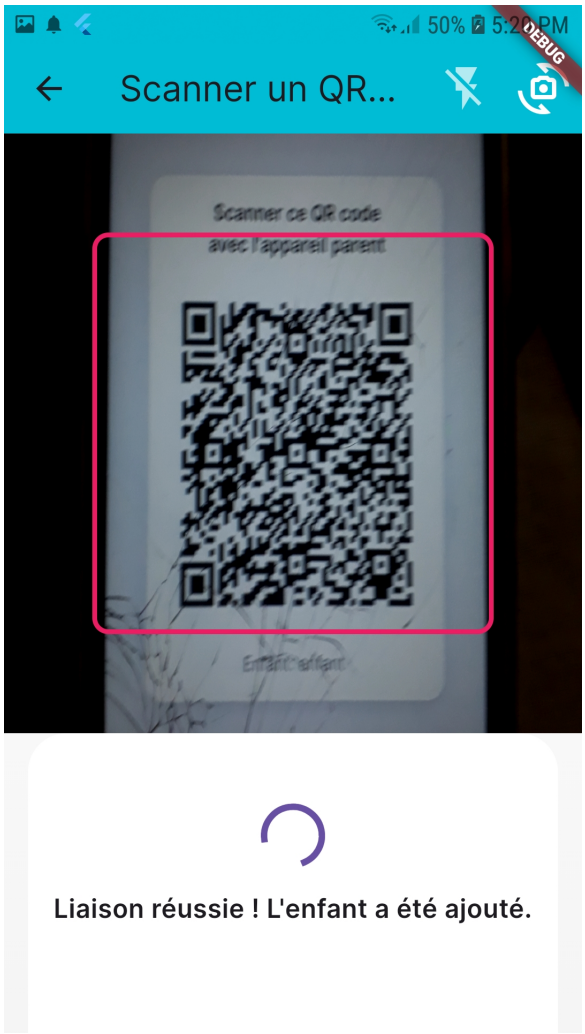


FIG. 3.13: Interface côté parent QrCode

### Détection en temps réel des émotions

Le processus de détection émotionnelle en temps quasi réel s'articule autour d'un modèle d'intelligence artificielle basé sur l'architecture de référence Wav2Vec2. Afin de le spécialiser pour la tâche visée, ce modèle a fait l'objet d'un ajustement fin (fine-tuning) sur un corpus spécifique de données vocales pédiatriques. Pour optimiser son déploiement et son interopérabilité, il a été converti au format ONNX (Open Neural Network Exchange) et hébergé sur une infrastructure cloud, son accès étant géré par une API sécurisée. Le protocole de capture sur le terminal de l'enfant consiste en l'enregistrement périodique de segments audio de cinq secondes via le microphone. Ces échantillons sont ensuite transmis à l'API, qui exécute une phase de prétraitement (incluant la normalisation et le rééchantillonnage du signal) avant de soumettre les données au modèle pour inférence. Ce dernier effectue alors une classification afin d'identifier l'émotion prédominante au sein d'un ensemble de classes discrètes (peur, tristesse, colère, joie, neutralité). La détection d'une émotion définie comme critique engendre la génération automatique d'une alerte

et l'envoi de notifications au parent. Le choix de cette architecture de traitement déporté permet d'allouer des ressources de calcul intensives à l'analyse sans surcharger le terminal mobile, au prix d'une dépendance modérée à la connectivité réseau.

### Système d'alerte et notifications intelligentes

Le système d'alerte intégré à l'application KedySafe repose sur une infrastructure de notifications intelligentes conçue pour garantir une transmission rapide et fiable des informations critiques aux parents. Lorsqu'une émotion négative (telle que la peur, la tristesse ou la colère) est détectée, une alerte est automatiquement générée. Cette alerte est immédiatement transmise via OneSignal, et peut inclure des détails tels que le type d'émotion détectée, l'heure de l'événement, et la position GPS de l'enfant si la géolocalisation est activée. Le système de notification est conçu pour fonctionner même lorsque l'application est en arrière-plan ou que l'appareil est verrouillé, assurant ainsi une continuité de service. Des mécanismes de priorité et de persistance permettent également d'alerter efficacement dans les situations critiques, notamment grâce à l'utilisation de sons spécifiques, de vibrations et de messages visibles en plein écran.

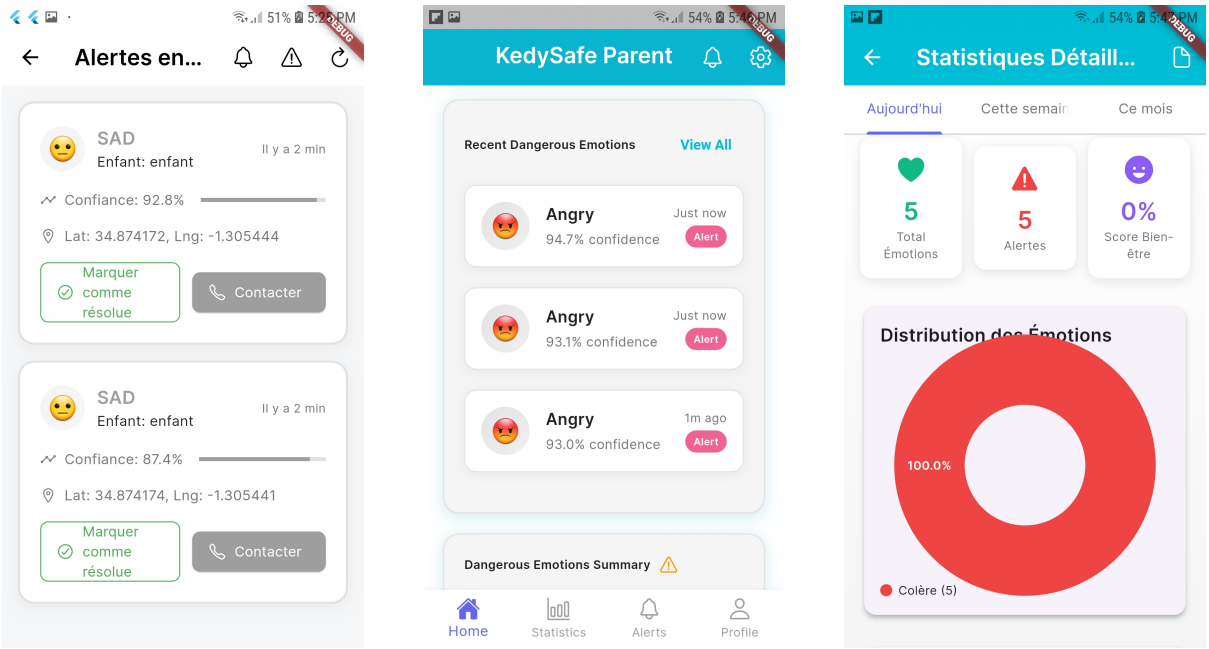


FIG. 3.14: Interface parent Alertes

### Détection d'èmotions et Intelligence Artificielle

La détection des émotions dans l'application **KedySafe** repose sur un pipeline de traitement audio optimisé pour les environnements mobiles. Ce processus s'articule autour de trois étapes principales : la capture du signal vocal, l'analyse par le modèle d'intelligence artificielle, et la classification de l'émotion détectée.

**1. Capture du signal vocal** Cette étape consiste à enregistrer en continu de courts segments audio (5 secondes) à l'aide du microphone de l'appareil mobile. Ce processus est conçu pour fonctionner discrètement en arrière-plan, sans perturber l'utilisation normale

du dispositif ni affecter significativement ses performances. L'objectif est d'assurer une surveillance constante, tout en respectant l'autonomie de l'enfant.

**2. Analyse par le modèle d'intelligence artificielle** Le traitement des segments audio capturés est pris en charge par un modèle fondé sur l'architecture Wav2Vec2, préalablement ajusté (fine-tuned) sur un corpus vocal pédiatrique et converti au format ONNX. Ce modèle opère une chaîne de traitement intégrée qui inclut, en première étape, la normalisation du signal audio brut. Il procède ensuite à l'extraction de représentations acoustiques latentes de haut niveau, lesquelles encapsulent les caractéristiques prosodiques essentielles à la classification de l'état émotionnel. L'exécution de ce processus unifié est gérée en ligne et accessible au travers d'une interface de programmation applicative (API).Afin de garantir une performance compatible avec les exigences d'un usage en temps réel, plusieurs optimisations ont été implémentées au niveau de la chaîne d'inférence. Celles-ci comprennent principalement la quantification du modèle au format entier 8 bits (INT8), une technique résultant en une réduction de son empreinte mémoire de plus de 75%. À cela s'ajoutent des optimisations logicielles telles que la mise en cache des sessions d'exécution ONNX, le traitement des requêtes par lots (batching) pour maximiser le débit, ainsi que l'assignation du processus de calcul à un fil d'exécution (thread) dédié.

**3. Classification des émotions et déclenchement des alertes** Les représentations extraites sont ensuite utilisées par un classificateur embarqué pour identifier l'émotion prédominante parmi un ensemble de classes définies (joie, peur, colère, tristesse, etc.). Lorsqu'une émotion critique est détectée, une alerte est générée et transmise via le système de notifications. Afin de permettre une exécution rapide et efficace.

## 3.10 Limites Actuelles

- **Dépendance à la connexion Internet** : l'analyse émotionnelle repose sur une API cloud, ce qui nécessite une connexion réseau stable.
- **Légère latence** : le temps de traitement et d'envoi d'alerte peut être perceptible dans certaines conditions.
- **Analyse limitée à la voix** : seule la voix est prise en compte pour la détection, excluant d'autres signaux émotionnels.
- **Dépendance à la technologie des parents** : le bon fonctionnement repose sur l'usage et la configuration correcte de l'application côté parent.

## 3.11 Pistes d'Amélioration

- **Traitement local** : intégration future du modèle directement sur l'appareil pour réduire la latence.
- **Analyse multimodale** : ajout de la reconnaissance faciale via caméra, et détection de mouvements à l'aide de capteurs (accéléromètre).

- **Personnalisation** : adaptation des modèles d'émotions aux profils spécifiques des enfants.
- **Apprentissage continu** : système auto-adaptatif apprenant les schémas émotionnels au fil du temps.
- **Détection avancée** : prédiction de crises émotionnelles et anomalies comportementales grâce à des modèles prédictifs.

## 3.12 Perspectives Avancées

- **Rapports psychologiques automatisés** : système auto-adaptatif apprenant les schémas émotionnels au fil du temps.
- **Soutien parental intelligent** : recommandations d'actions concrètes pour mieux accompagner l'enfant en fonction de ses états émotionnels.

## 3.13 Interfaces Parent

Cette section présente l'interface utilisateur de l'application mobile KedySafe, conçue avec Flutter. Les captures suivantes illustrent les écrans principaux de l'appareil du parent.

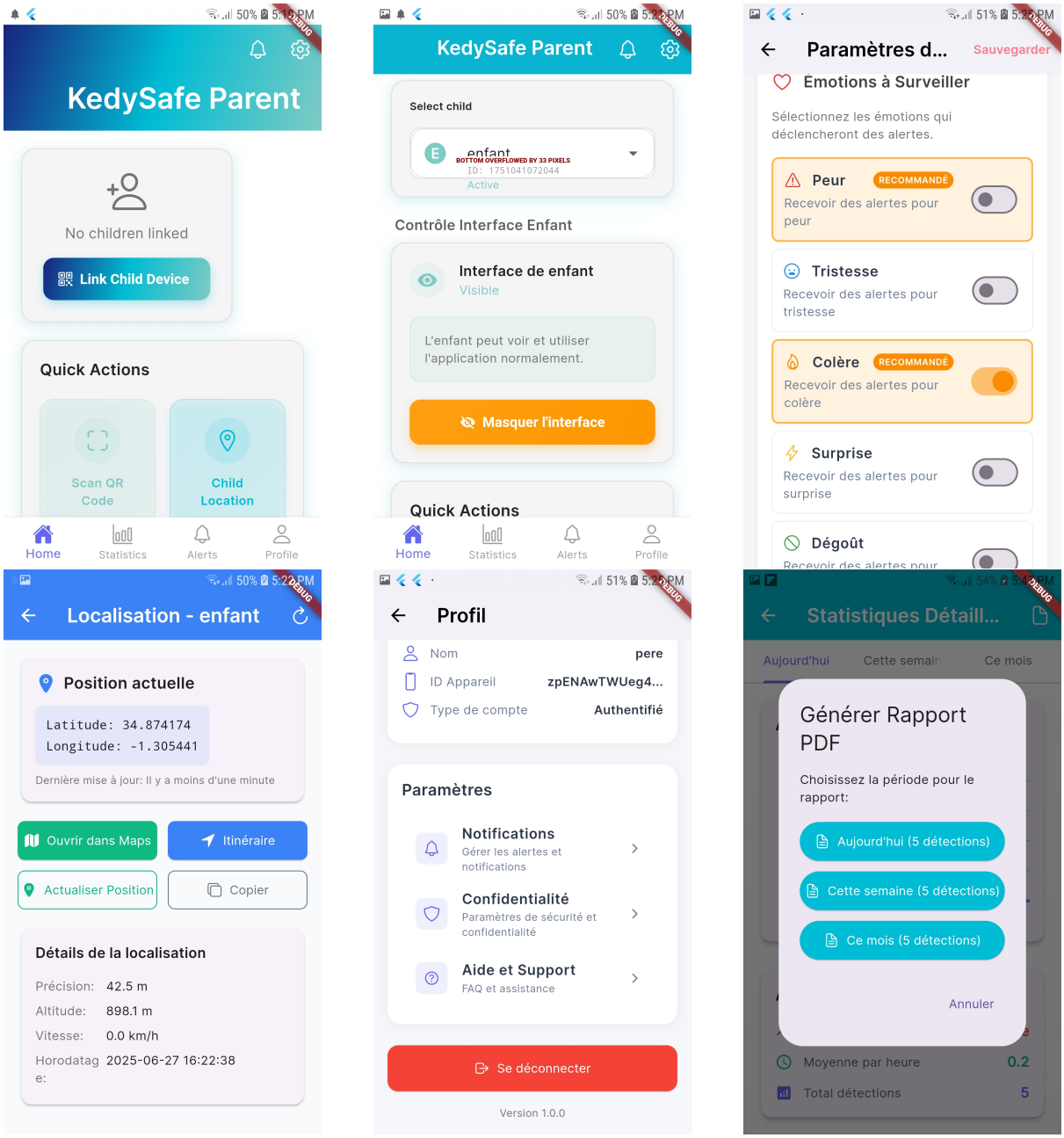


FIG. 3.15: Interfaces principale de l'appareil de parent

## 3.14 Interfaces Enfant

Cette section présente l'interface utilisateur de l'application mobile KedySafe, conçue avec Flutter. Les captures suivantes illustrent les écrans principaux de l'appareil d'enfant.

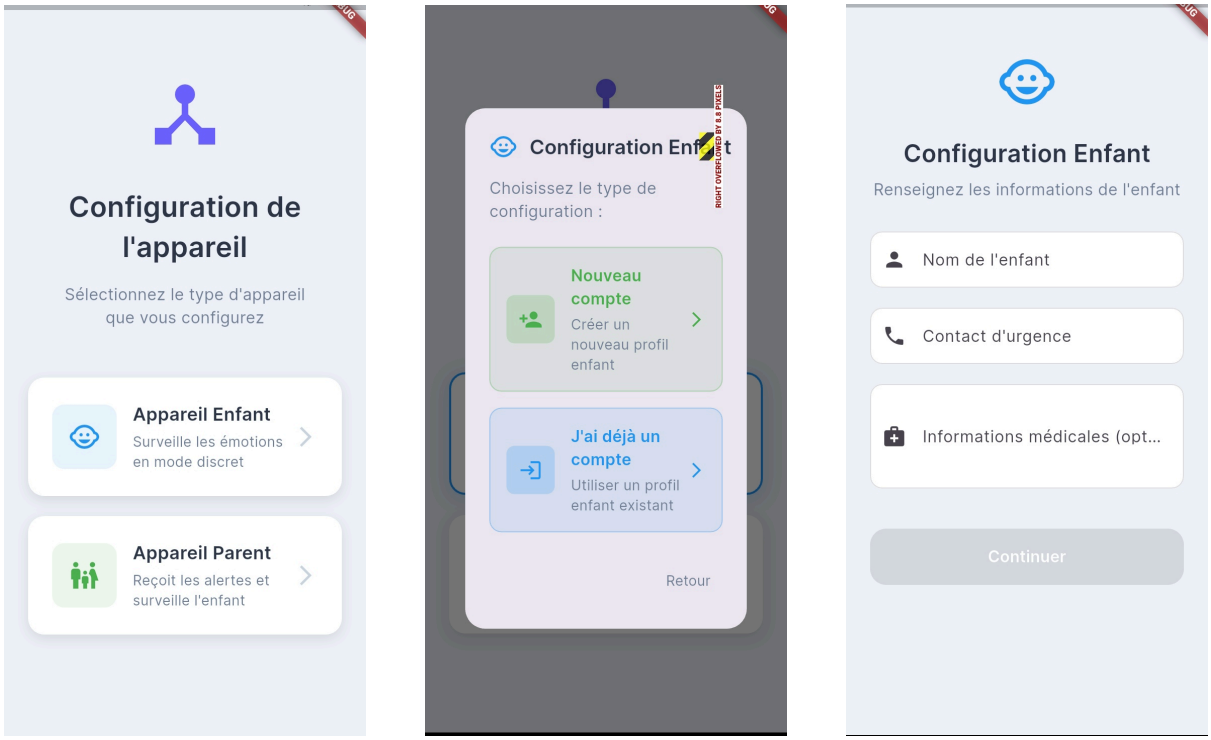


FIG. 3.16: Interfaces de configuration de l'appareil d'enfant

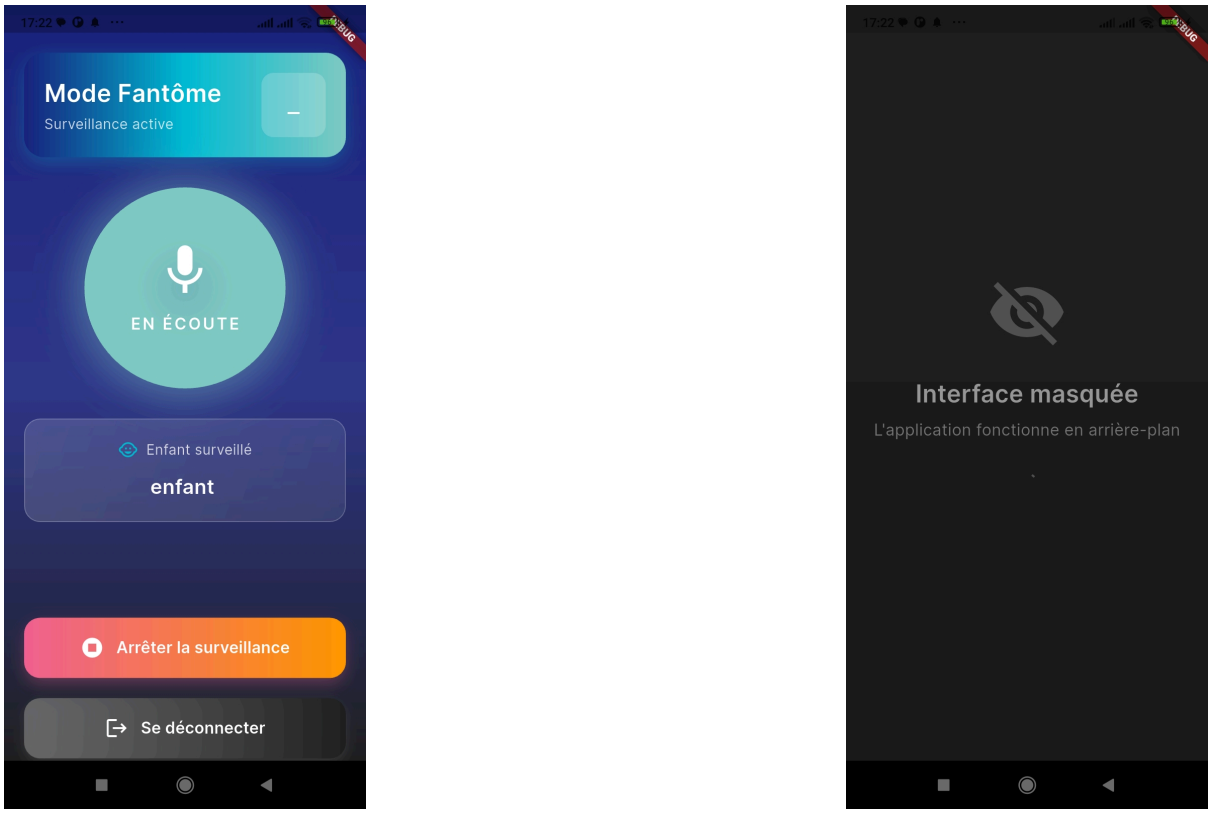


FIG. 3.17: Interfaces principale de l'appareil d'enfant

## 3.15 Conclusion :

Ce chapitre a présenté l'ensemble des étapes techniques ayant conduit à la mise en uvre de notre solution intelligente de détection des émotions vocales chez les enfants. Après avoir décrit les jeux de données utilisés, les émotions ciblées et le processus de préparation, nous avons évalué plusieurs modèles pré-entraînés, avant d'effectuer un fine-tuning sur un corpus vocal adapté à la voix infantile.

Les résultats ont mis en évidence la pertinence du modèle Wav2Vec2-Robust Emotion, qui s'est distingué par un bon équilibre entre précision, rappel et robustesse. Ce modèle, une fois fine-tune, a été converti au format ONNX et déployé sur le cloud via une API, permettant ainsi une inférence fiable tout en allégeant la charge computationnelle sur les appareils mobiles.

L'implémentation s'appuie sur une architecture distribuée et moderne, articulée autour de Flutter (frontend), Firebase (authentification, base de données, notifications) et Google Cloud Functions pour l'hébergement du modèle. Cette approche garantit à la fois portabilité, scalabilité et respect de la confidentialité des données.

Par ailleurs, des efforts particuliers ont été consacrés à l'ergonomie, à la sécurité et à la résilience du système, via des fonctionnalités telles que le mode fantôme, la liaison sécurisée par QR code, ou encore un système d'alerte intelligent déclenché lors de la détection d'émotions critiques.

Cette phase d'expérimentation et d'implémentation a ainsi permis de valider la faisabilité technique de notre solution et de concrétiser un système fiable, modulaire et opérationnel, capable de contribuer efficacement à la surveillance émotionnelle proactive des enfants.

# Chapitre 4

Conclusion générale

Le présent travail s'inscrit dans une dynamique de recherche appliquée visant à contribuer au renforcement des mécanismes de protection et de suivi du bien-être émotionnel des enfants, en tirant parti des technologies intelligentes de dernière génération. Dans un contexte où la santé mentale des jeunes est devenue une préoccupation majeure à l'échelle mondiale, nous avons cherché, à travers ce projet, à explorer les potentialités offertes par l'intelligence artificielle, et plus précisément par l'analyse de la voix, pour détecter les états émotionnels des enfants de manière automatique, rapide et fiable. L'approche adoptée repose sur les progrès significatifs du deep learning, notamment à travers l'utilisation de modèles avancés de traitement du signal audio comme Wav2Vec2.0, capables de capter les subtilités vocales associées aux différentes émotions humaines.

Notre méthodologie a été rigoureusement structurée en plusieurs étapes fondamentales. Nous avons tout d'abord entrepris la collecte de données vocales spécifiques à la population infantile, en nous assurant de la qualité et de la représentativité des enregistrements. Ensuite, un travail de prétraitement du signal a permis de normaliser et de segmenter les données, facilitant ainsi l'extraction de caractéristiques acoustiques pertinente nécessaires pour entraîner le modèle. Le fine-tuning de modèles pré-entraînés sur des corpus plus généraux a ensuite été réalisé sur notre jeu de données, en ciblant des émotions clés comme la peur, la colère, la tristesse, la joie, le dégoût et la neutralité. L'évaluation des performances du système a montré des résultats encourageants, notamment dans la détection des émotions négatives qui sont souvent les plus critiques à identifier dans une optique de protection et d'intervention rapide.

Ce projet a également permis de souligner l'importance de concevoir des solutions technologiques qui soient non seulement performantes, mais aussi respectueuses de l'environnement de l'enfant. Nous avons ainsi veillé à ce que le système soit non intrusif, discret, et capable de fonctionner en temps réel, afin de pouvoir être intégré efficacement dans un contexte familial. À ce titre, l'intégration du modèle dans une application mobile destinée aux parents représente une avancée concrète vers des usages pratiques et socialement utiles, tels que le suivi émotionnel quotidien de l'enfant, la prévention des situations de détresse psychologique, ou encore l'accompagnement à distance par des professionnels de la santé mentale.

Cependant, malgré les performances obtenues, ce travail n'est pas exempt de limites. La variabilité des voix enfantines  influencée par l'âge, le sexe, ou même l'état physiologique  constitue un facteur de complexité non négligeable. De plus, le manque de bases de données suffisamment annotées et représentatives freine le potentiel de généralisation du modèle. Enfin, la nature multidimensionnelle et contextuelle des émotions humaines rend leur classification parfois difficile à automatiser de manière exhaustive. Ces constats ouvrent des perspectives de recherche futures, telles que l'enrichissement et la diversification des corpus vocaux, l'adaptation des modèles aux spécificités culturelles et linguistiques, ou encore l'adoption d'approches multimodales combinant la voix à d'autres indices comme les expressions faciales ou les gestes corporels.

En conclusion, ce projet constitue une première étape prometteuse vers la mise en place de systèmes intelligents au service de la protection émotionnelle des enfants. Il démontre qu'il est possible, grâce à l'intelligence artificielle, d'envisager des outils d'alerte et de suivi émotionnel à la fois fiables, accessibles, éthiques et adaptés au jeune public. En dotant les parents d'un tel outil technologique, nous contribuons non seulement à la prévention des troubles émotionnels, mais aussi à la construction d'un environnement plus attentif et bienveillant pour le développement affectif de l'enfant.

# Bibliographie


[1] Oana Bălan, Gabriela Moise, Livia Petrescu, Alin Moldoveanu, Marius Leordeanu, and Florica Moldoveanu. Emotion classification based on biophysical signals and machine learning techniques. *Symmetry*, 12(1) :21, 2019.

[2] Leila Kerkeni. *Analyse acoustique de la voix pour la détection des émotions du locuteur*. PhD thesis, Le Mans Université ; Université de Sousse (Tunisie), 2020.

[3] Maroua Aissa and Romaila Ait Mesbah. Développement et implémentation dun système de reconnaissance des émotions à partir de la parole et des expressions faciales. Mémoire de fin d'études, École Nationale Polytechnique, Alger, 2023. Dirigé par Nesrine Bouadjenek.

[4] Solomon Seyife Alemu, Mohammedamin Hajure, Mahlet Tesfaye Agago, Feisal Hussein, Hana Israel Gesisa, Sheleme Mengistu Teferi, Daniel Yohanes, and Lema Fikadu Wedajo. Prevalence of burnout and associated factors among midwives, 2023: institution-based cross-sectional study. *Frontiers in Public Health*, 12:1422915, 2024.

[5] Guido Gainotti. *Emotion, Cognition, and Their Relationships : A Neuropsychological Perspective*. Springer, 2020.

[6] Pedro Saraiva and Ayse Ayanoglu. Emotional intelligence and decision making. *Journal of Behavioral Sciences*, 45(2) :123--135, 2019.

[7] Amanda Jackson. *Emotion, Bias, and Identity : A Cultural Psychology Approach*. Oxford University Press, 2024.

[8] Philippe Verduyn and Kristof Brans. The relationship between extraversion, neuroticism and aspects of trait affect. *Personality and Individual Differences*, 52(6) :757--761, 2012.

[9] Olga Koltsova. *Culture and Emotion : A Sociolinguistic Perspective*. Routledge, 2024.

[10] M. Habib et al. The dynamics of emotion : A multidisciplinary approach. *Emotion Review*, 10(3) :210--220, 2018.

[11] Fleur Lejeune and Édouard Gentaz. Le développement de la discrimination des émotions chez les bébés humains. *Société des Neurosciences*, 65(1), 2023.

[12] Albert Mehrabian and James A Russell. *An approach to environmental psychology.* the MIT Press, 1974.

[13] Susanne A. Denham. Social-emotional competence as support for school readiness : What is it and how do we assess it ? *Early Education and Development*, 17(1) :5789, 2006.

[14] Nim Tottenham and Margaret A. Sheridan. A review of adversity, the amygdala and the hippocampus : A consideration of developmental timing. *Frontiers in Human Neuroscience*, 3:68, 2010.

[15] B.J. Casey, R.M. Jones, and T.A. Hare. The adolescent brain. *Annals of the New York Academy of Sciences*, 1124(1) :111126, 2008.

[16] Carolien Rieffe and Mario De Rooij. The longitudinal link between emotion awareness and internalizing symptoms during late childhood and adolescence. *Child Psychiatry & Human Development*, 43:901918, 2012.

[17] P. Chitsabesan et al. Mental health needs of young offenders in custody and in the community. *British Journal of Psychiatry*, 188(6) :534540, 2006.

[18] Howard A. Liddle. Multidimensional family therapy : A sciencebased treatment for adolescent drug abuse. *The Counseling Psychologist*, 38(6) :902914, 2010.

[19] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. Survey on speech emotion recognition : Features, classification schemes, and databases. *Pattern recognition*, 44(3) :572--587, 2011.

[20] Daniel Jurafsky and James H. Martin. *Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, 2nd edition, 2009.

[21] Lawrence R. Rabiner and Ronald W. Schafer. *Digital Processing of Speech Signals*. Prentice Hall, 1978.

[22] Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon. *Spoken Language Processing : A guide to theory, algorithm, and system development*. Prentice Hall PTR, 2001.

[23] Tomi Kinnunen and Haizhou Li. An overview of text-independent speaker recognition : From features to supervectors. *Speech Communication*, 52(1) :12--40, 2010.

[24] Alan V. Oppenheim and Ronald W. Schafer. *Discrete-Time Signal Processing*. Prentice Hall, 3rd edition, 2010.

[25] Julius O. Smith. *Spectral Audio Signal Processing*. W3K Publishing, 2011. `https://ccrma.stanford.edu/~jos/sasp/`.

[26] Dimitrios Ververidis and Constantine Kotropoulos. Emotional speech recognition : Resources, features, and methods. *Speech communication*, 48(9) :1162--1181, 2006.

[27] Klaus R Scherer. Vocal communication of emotion : A review of research paradigms. *Speech communication*, 40(1-2) :227--256, 2003.

[28] Christer Gobl and Ailbhe Ní Chasaide. Voice source variation and its communicative functions. *Phonetica*, 60(2) :1--19, 2003.

[29] Lawrence R Rabiner. *Theory and Application of Digital Speech Processing*. Pearson Education, 2011.

[30] Brian McFee, Colin Raffel, Dawen Liang, Daniel P.W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa : Audio and music signal analysis in python. In *Proceedings of the 14th Python in Science Conference*, pages 18--25, 2015.

[31] Hugging Face. audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim, 2023. Accessed : 2025-05-31.

[32] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[33] Hugging Face. Model card : Wav2vec2-large-robust, 2023. Accessed : 2025-05-31.

[34] Reza Lotfian and Carlos Busso. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcasts. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC)*, 2019.

[35] Roddy Cowie, Ellen Douglas-Cowie, Sylvaine Savvidou, Elizabeth McMahon, Morag Sawey, and Marc Schröder. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1) :32--80, 2001.

[36] Hugging Face. ehcalabres/wav2vec2-lg-xlsr-en-speech-emotion-recognition, 2022. Accessed : 2025-05-31.

[37] 5 AI Models. wav2vec2-lg-xlsr-en-speech-emotion-recognition | ai model details, 2024. Accessed : 2025-05-31.

[38] J. Yan et al. Detection of children abuse by voice and audio classification by short-time fourier transform machine learning implemented on nvidia edge gpu device. *arXiv preprint arXiv :2307.15101*, 2023.

[39] Y. Yao et al. Infant crying detection in real-world environments. *arXiv preprint arXiv :2005.07036*, 2020.

[40] Jean-Claude Martin, Laurence Devillers, Mohamed Sehili, and Nicolas Bel. Détection des émotions à partir du contenu linguistique dénoncés oraux : application à un robot compagnon pour enfants fragilisés. In *JEP-TALN-RECITAL 2009*, Avignon, France, 2009.

**Abstract**

Children's safety is a top priority in a constantly changing world, where they may face stress or fear. This project proposes an intelligent solution to detect and predict emotional states through children's voices, using artificial intelligence and machine learning. The system analyzes vocal signals to identify emotions such as joy, sadness, anger, or fear, and to detect potential signs of danger or distress. Designed for practical use, it can be integrated into mobile devices like smartwatches or smartphones, allowing parents to monitor their child's emotional state in real time and to receive automatic alerts in case of anomalies. In addition to its security role, this solution promotes a better understanding of the child's emotions and strengthens parentchild communication within a supportive environment.

**Keywords :** Artificial Intelligence, Machine Learning, Speech Signal Processing, Emotions, Children, Protection, Voice Recognition, Alert System, RealTime Monitoring, Smart Application, Smartwatch, Emotional Analysis.

**Résumé**

La sécurité des enfants est une priorité majeure dans un monde en constante évolution, où ils peuvent être confrontés à des situations de stress ou de peur. Ce projet propose une solution intelligente capable de détecter et de prédire les états émotionnels à partir de la voix de l'enfant, grâce à l'intelligence artificielle et à l'apprentissage automatique. Le système repose sur l'analyse des signaux vocaux de l'enfant pour identifier des émotions telles que la joie, la tristesse, la colère ou la peur, ainsi que des signes éventuels de danger ou de trouble. Conçu pour une utilisation pratique, il peut s'intégrer à des dispositifs mobiles tels que des montres connectées ou des smartphones, permettant ainsi aux parents de suivre en temps réel l'état émotionnel de leur enfant et de recevoir des alertes automatiques en cas d'anomalie. Audelà de son rôle sécuritaire, ce système favorise une meilleure compréhension des émotions de l'enfant et renforce la communication parentenfant dans un cadre bienveillant.

**Mots clés :** Intelligence Artificielle, Apprentissage Automatique, Traitement du Signal Vocal, Émotions, Enfants, Protection, Reconnaissance Vocale, Système d'Alerte, Suivi en Temps Réel, Application Intelligente, Montre Connectée, Analyse Émotionnelle.

**ملخّص**

تمثّل سلامة الطفل أولويةً قصوى في عالم متغيّر، حيث يمكن أن يتعرّض الطفل لمواقف توتر أو خوف. يقترح هذا المشروع نظامًا ذكيًا يعتمد على الذكاء الاصطناعي والتعلّم الآلي لتحليل عواطف الطفل من صوته، والتنبّؤ بالحالات التي تشير إلى خطر أو توتر. يعتمد النظام على معالجة الإشارات الصوتية للتعرّف على مشاعر مثل الفرح، الحزن، الغضب، والخوف، وكذلك على تحديد العلامات التي تدل على خطر محتمل. يتميّز النظام بإمكانية دمجه على أجهزة محمولة كالساعات الذكية أو الهواتف، مما يمكّن الأهل من متابعة الحالة العاطفية للطفل وتنبيههم فورًا عند ملاحظة تغيّر مقلق. كما يُسهم النظام في تعزيز الفهم العاطفي لدى الطفل وتقوية التواصل الأسري ضمن إطار داعم ومتكامل.

**الكلمات المفتاحية :** الصوت، على التعرف الحماية، الأطفال، المشاعر، الصوتية، الإشارات معالجة الآلي، التعلم الاصطناعي، الذكاء العاطفي. التحليل الذكية، الساعة الذكي، التطبيق الحقيقي، الوقت في التتبع الإنذار، نظام