Laurence Devillers
Sorbonne-Université, CNRS-LISN (France)

Théo Deschamps-Berger
Université Paris-Saclay, CNRS-LISN (France)

Lori Lamel
Université Paris-Saclay, CNRS-LISN (France)

Les émotions ‹*in the wild*› des appelants d'un centre d'appels d'urgence : vers un système de détection des émotions dans la voix

1. INTRODUCTION

Ce papier présente une étude sur la détection des émotions à partir du signal audio et montre le mélange des émotions dans un corpus collecté dans un centre d'appels d'urgence à Paris (CEMO). Lors de l'enregistrement de données ‹*in the wild*›, il est fréquent de trouver de nombreuses émotions mixtes ou mélangées dépendantes du contexte des situations d'interaction. Nous décrivons les approches récentes en neurosciences et sciences affectives qui remettent en question les théories classiques des émotions, ainsi que le corpus CEMO et son annotation, la méthodologie d'apprentissage et, enfin, l'analyse des émotions mixtes et l'évaluation des principales macro-émotions présentes pour les appelants : *peur*, *tristesse*, *positif* et *neutre*. La classe *peur* dans ce corpus comprend, par exemple, des émotions plus « fines » telles que la panique ou l'anxiété ; la classe *tristesse*, la résignation et le désespoir ; et la classe *positive*, le soulagement et l'intérêt. Le but de ce travail est de comparer des systèmes de reconnaissance automatique des émotions vocales à l'aide de *Transformers* à des systèmes plus classiques de détection pour des émotions non mixtes. Une première étude est menée en détection sur les émotions mixtes. Cette recherche est menée dans le cadre de la chaire hors 3IA HUMAAINE – *Human-Machine Affective Interaction & Ethics* pilotée par L. Devillers au LISN-CNRS.

2. LES AVANCÉES DANS LES RECHERCHES SUR LES ÉMOTIONS EN SCIENCES AFFECTIVES

2.1. Un sujet pluridisciplinaire

Depuis les débuts de la science psychologique, les chercheurs se demandent comment le cerveau génère des émotions. Les neurosciences affectives étudient les émotions comme des états fonctionnels biologiques mais elles examinent également l'expérience consciente de l'émotion appelée *sentiment*, ainsi que notre capacité à attribuer des émotions aux autres, notre capacité à parler d'elles et les comportements qu'elles engendrent. Ces états fonctionnels peuvent également provoquer des expériences conscientes, et nos souvenirs de leurs effets contribuent à notre connaissance sémantique. À la suite de l'explosion de ces travaux, l'émotion est maintenant considérée comme un facteur explicatif déterminant du comportement humain. Les théories sont nombreuses. K. R. Scherer, psychologue spécialisé dans l'étude des émotions, considère l'état émotionnel comme un processus de variations épisodiques dans plusieurs composantes de l'organisme en réponse à des événements évalués comme importants pour l'organisme (Scherer, Schorr & Johnstone 2001). Selon cette définition, l'émotion est constituée de cinq composantes : l'évaluation cognitive, les changements psychophysiologiques, l'expression motrice, les tendances à l'action et le sentiment subjectif. Avec l'essor des études en neurosciences affectives ces dernières décennies, des études interdisciplinaires avec des conceptions d'imagerie cérébrale plus écologiquement valides devraient être de plus en plus utilisées pour mieux comprendre ces différents phénomènes. Des chercheurs comme L. F. Barrett, s'appuyant sur les neurosciences, ont avancé la théorie de la construction des émotions (Barrett 2017). Elle considère que l'émotion est un phénomène qui émerge dans la conscience à un moment à partir d'ingrédients comme l'intéroception et les concepts et la réalité sociale. Selon cette théorie, les émotions ne possèdent pas d'empreintes spécifiques et universelles ; elles varient en fonction des contextes et des cultures. Elles se forment par la combinaison de signaux physiques et de concepts, acquis grâce à la flexibilité du cerveau. L'émotion est vue comme une réalité sociale, façonnée par le consensus dans une société. L. F. Barrett prône l'étude de la diversité des caractéristiques émotionnelles. Le cerveau, recevant en continu des informations sensorielles internes et externes, doit leur attribuer un sens. Ainsi, une émotion comme la peur peut se construire à partir de diverses caractéristiques physiques. Ce processus de catégorisation utilise des concepts appris depuis l'enfance et tient compte du contexte, avec un rôle important accordé au langage pour exprimer l'émotion. La dimension sociale permet, en outre, la communication émotionnelle et l'influence sociale. Dans le dialogue, les émotions se construisent aussi pendant l'interaction en fonction des réponses de l'autre (Damasio 2018). Selon K. R. Scherer, les émotions résultent d'une évaluation cognitive des stimuli en termes de pertinence pour les besoins et les objectifs personnels, et les différentes composantes fonctionnent en coordination pour produire une réponse émotionnelle intégrée. Il reconnaît,

à la différence de L. F. Barrett, certaines bases universelles des émotions mais il accepte également des variations dues aux différences individuelles et culturelles.

Les progrès récents dans le domaine des neurosciences affectives et de l'*affective computing* suggèrent une convergence croissante entre ces deux disciplines du domaine des sciences affectives. Des études telles que « *How should neuroscience study emotions?* » de R. Adolphs (2017) mais aussi l'utilisation d'outils d'apprentissage avancés comme les *Transformers* pour la détection des émotions (Deschamps-Berger, Lamel & Devillers 2023) ont permis des avancées notables. Cependant, ces développements remettent en question certaines des méthodologies et théories traditionnelles dans la recherche sur les émotions, suggérant une réévaluation de la compréhension et de l'étude des émotions dans ces champs scientifiques.

2.2. L'expression des émotions est-elle universelle ?

La théorie classique des émotions, notamment celle de P. Ekman (1973), postule que des *stimuli* spécifiques déclenchent automatiquement six émotions universelles (colère, peur, dégoût, surprise, tristesse et bonheur), identifiables par leurs « empreinte » distinctes. Dans son article de 1994, J. A. Russell (1994) a exprimé son désaccord avec le concept d'*émotions universelles* proposé par P. Ekman et d'autres chercheurs. Selon J. A. Russell, l'idée d'expressions émotionnelles universelles, en particulier celles identifiées par les expressions faciales, n'est pas aussi claire ou universellement applicable que ce que l'on pensait auparavant. Il a suggéré que l'interprétation des expressions faciales des émotions peut varier considérablement d'une culture à l'autre, remettant en question l'idée que certaines expressions émotionnelles sont universellement reconnues et comprises. Les avancées de R. Adolphs et D. J. Anderson (2018) mettent en effet en évidence le rôle significatif du contexte, de l'apprentissage et de la culture dans leur formation et leur expression. Ces avancées remettent en question la validité des représentations classiques des émotions universelles. Les deux dernières décennies de recherche en neurosciences convergent vers l'hypothèse selon laquelle tous les événements mentaux, émotionnels ou autres, sont générés sous forme de prédictions, ne réagissant pas aux mots émotionnels comme descripteurs unitaires de l'expérience mais plutôt aux caractéristiques mentales de l'émotion qui peuvent varier d'une instance à l'autre du même mot (Hoemann, Gendron & Barrett 2017).

2.3. Quelle représentation pour les émotions ‹in the wild› ?

La méthode expérimentale de P. Ekman (1973), basée sur des photographies de visages exprimant des émotions et une liste restreinte de choix d'étiquettes émotionnelles, est critiquée pour son manque de robustesse, notamment parce qu'elle oriente les réponses. Aujourd'hui, l'approche dimensionnelle (valence, arousal) est de plus en plus privilégiée par rapport à l'approche par étiquette verbale

dans le domaine de l'*affective computing*, comme l'indiquent M. Wöllmer *et al.* (2008). La détection des émotions traditionnellement fondée sur des catégories verbales présente des limites car il s'agirait plutôt d'un *continuum*. Cependant, il est difficile de travailler sur les mélanges émotionnels si l'on ne se réfère pas à des étiquettes verbales. Les méthodes de collecte en laboratoire offrent des avantages liés à un environnement contrôlé, des protocoles bien définis et un lexique prédéterminé. Toutefois, dans les environnements de déploiement réels, on observe souvent une variabilité complexe qui crée un écart entre les bases de données disponibles publiquement et les applications de conversation naturelle. K. R. Scherer a également apporté des contributions à la compréhension de cette complexité. Ses travaux se concentrent sur la manière dont les émotions sont vécues, exprimées et perçues, et il a exploré le concept des *émotions mélangées* qui réfère à la coexistence de plusieurs émotions différentes en même temps. Dans ses travaux, K. R. Scherer (Scherer, Schorr & Johnstone 2001) a souvent abordé les émotions à travers le prisme de son modèle appelé *Component Process Model* (CPM), qui décrit les émotions comme le résultat de plusieurs composantes interactives, notamment la cognition, la physiologie, l'expression motrice et la motivation. Ce modèle est particulièrement utile pour comprendre les émotions mélangées, car il reconnaît que les différentes composantes peuvent réagir à différents aspects d'une situation, produisant ainsi une expérience émotionnelle complexe et nuancée. À mesure que les expressions émotionnelles deviennent plus naturelles et ambiguës, l'utilisation d'une étiquette unique peut s'avérer trop limitative. Il pourrait être nécessaire de développer de nouvelles méthodes, telles que l'utilisation de *soft-label* offrant des annotations plus riches qui tiennent compte de l'incertitude de la perception humaine ou de la variabilité des annotateurs (Han *et al*. 2017 ; Chou *et al*. 2022). Il y a peu d'études des émotions mixtes montrant des approches théoriques (Hoemann, Gendron & Barrett 2017) et il existe encore de nombreuses questions inexplorées concernant les émotions mixtes.

2.4. Les indices caractérisant les émotions

Nous n'avons pas d'instrument pour mesurer directement une émotion mais seulement des mesures comportementales indirectes. Celles-ci consistent en une observation et une quantification des réponses dans la voix, les mots prononcés, le visage, la posture, les gestes ou encore les actions. Les premières méthodes de modélisation reposent sur la conception de caractéristiques fondées sur les connaissances linguistiques et paralinguistiques qui capturent leurs modulations acoustiques dans la parole émotionnelle (Eyben *et al.* 2016 ; Devillers, Vidrascu & Lamel 2005). Leur représentation dans la voix a été empiriquement documenté par la mesure de la phonation et de l'articulation, à partir de paramètres dans le domaine temporel (p. ex : débit de parole), fréquentiel (f_o et formants), de l'amplitude (intensité ou énergie), de la qualité vocale (*jitter* ; variations à court terme de f_o et *shimmer* ; variations à court terme de l'intensité) et de la distribution

spectrale (MFCCs, *Mel Frequency Cepstral Coefficients* sont des paramètres caractérisant l'enveloppe spectrale de la voix). Ces descripteurs sont généralement combinés à des mesures statistiques pour encoder les trajectoires temporelles et sont ensuite représentés dans un vecteur de taille fixe. De nombreuses autres mesures sont également utilisées, comme la fréquence cardiaque, la pression artérielle, les niveaux d'hormones de stress sécrétées dans le sang, la sudation, ainsi que les changements dans l'activité cérébrale. Cet ensemble de mesures permet d'affirmer qu'il y a bien un changement d'état interne en présence d'un *stimulus*. L'expérience de détection du stress lors de prises de parole en public (Hua et *al.* 2016 ; Giraud *et al.* 2013), montre la difficulté de trouver des indices universaux pour tous les sujets. Les méthodes permettant de mesurer ces paramètres de manière de plus en plus précise, en particulier l'activité cérébrale et le comportement, sont en constante amélioration.

2.5. Les solutions récentes *end-to-end* et les modèles pré-entraînés

Le terme *end-to-end* fait généralement référence à des systèmes intégrés capables de gérer l'ensemble du processus de détection des émotions, des données initiales à l'analyse et à la réponse finale. L'article de recherche de G. Trigeorgis *et al.* (2016) intitulé « *Adieu features?* » suggère l'utilisation de réseaux neuronaux profonds, combinant des techniques convolutionnelles et récurrentes, pour la reconnaissance des émotions à partir de la parole. Cette approche *end-to-end* signifie que le système est conçu pour traiter directement les données brutes de la parole sans nécessiter de sélection manuelle de caractéristiques (*features*), marquant potentiellement un changement significatif dans les méthodes traditionnelles de reconnaissance des émotions par la parole. Les solutions *end-to-end* sont devenues plus performantes que les approches expertes (Etienne *et al*. 2018 ; Deschamps-Berger, Lamel & Devillers 2021). Les modèles pré-entraînés représentent également une avancée significative dans la reconnaissance des émotions, grâce à leur capacité à traiter et à généraliser à partir de vastes ensembles de données. Ce sont des réseaux de neurones profonds, fondés sur des mécanismes d'attention (Cheng, Dong & Lapata 2016), qui sont pré-entraînés sur de larges ensembles de données non étiquetées de manière auto-supervisée. Ces modèles permettent de capturer efficacement des corrélations de faits et de traiter un large éventail de sujets et de formes. En 2017, A. Vaswani *et al*. ont présenté l'architecture *Transformer* dans « *Attention is all you need* » (Vaswani *et al*. 2017), une architecture largement répliquée et adaptée pour créer des encodeurs performants tels que BERT (*Bidirectional Encoder Representations from Transformers*) (Devlin *et al*. 2019) et ses variantes dont FlauBERT en français (Le *et al*. 2020). Des encodeurs auto-supervisés, pré-entraînés sur de vastes corpus de texte peuvent être adaptés pour une grande variété de tâches. Cette approche a été étendue à l'audio avec WAV2VEC 2.0 (Baevski *et al.* 2020), qui introduit une méthode auto-supervisée de discrétisation des représentations audio. Ce modèle génère des représentations latentes de la parole qui peuvent être ajustées pour des tâches comme la reconnaissance automatique de la parole (ASR). La communauté

en apprentissage automatique des émotions s'accorde désormais sur l'utilité d'adapter ces modèles pour traiter des tâches en aval plutôt que d'apprendre des modèles à partir de zéro.

3. CORPUS ET ANNOTATION

Le corpus ‹*in the wild*› CEMO pour C(orpus)EMO(tions) est un corpus de données d'urgence, collecté en 2003 (avec un accord entre le LIMSI-CNRS et le SAMU) (Devillers, Vidrascu & Lamel 2005). Le schéma d'annotation est décrit dans de précédentes recherches (Devillers, Vidrascu & Lamel 2005 ; Vidrascu & Devillers 2005). Les centres d'appels d'urgence sont ouverts 24 heures sur 24, 7 jours sur 7. Les Agents de Régulation Médicale (ARM) sont formés pour évaluer rapidement les besoins de l'appelant, son état émotionnel, le niveau de crise et l'urgence de la situation pour déterminer une prise en charge adaptée : assistance ambulancière ou médicale, soins psychiatriques ou redirection du patient vers son médecin traitant. L'utilisation de ces données respecte les conventions éthiques assurant l'anonymat des données et respectant le RGPD (Règlement Général sur la Protection des Données).

Les données collectées ont été annotées par plusieurs personnes avec un schéma original permettant de représenter plusieurs émotions par segment. Les annotateurs ont eu la possibilité de choisir une émotion majeure et une émotion mineure pour chaque segment extrait parmi 21 étiquettes fines qui sont ensuite regroupées en 7 macro-classes : *peur* (anxiété, stress, peur, panique, déception, embarras), *tristesse* (résignation, déception, tristesse, désespoir), *colère* (impatience, agacement, colère froide et chaude), *positif* (soulagement, intérêt, compassion, amusement), *douleur*, *surprise* et *neutre*. L'annotation a été effectuée en 2 temps : une première annotation (Vidrascu & Devillers 2005) a été menée par deux psycholinguistes ; une seconde, suivant le même schéma d'annotation, a été menée récemment sur le reste du corpus.

Le corpus CEMO de notre étude regroupe toutes les annotations. Comme nous utilisons des catégories d'annotations en majeur et mineur, nous avons opté pour une approche du coefficient Kappa plus souple, permettant de considérer les annotateurs comme étant en accord dès lors qu'une émotion identique est présente soit en majeur, soit en mineur. La classification émotionnelle des segments est déterminée par un vote majoritaire, où les étiquettes émotionnelles majeures comptent double par rapport aux mineures. En situation d'égalité, le segment est classé comme exprimant des émotions mélangées. Pour donner une idée de l'inter-annotation, sur la première partie du corpus (Devillers, Vidrascu & Lamel 2005), l'accord inter-annotateur sur les annotations de macro-classes pour la première partie (Vidrascu & Devillers 2005) a un indice Kappa de 0,54 pour les appelants et sur la seconde partie annotée récemment de 0,61 pour les appelants.

La partie du corpus des données CEMO (21 h) comporte la voix de 1 127 appelants soit 26 649 segments annotés par au moins 2 annotateurs. Le corpus complet CEMO (appelants, agents) (Devillers, Vidrascu & Lamel 2005) contient 52 097 segments annotés, issus de 967 conversations et de 1 437 locuteurs. Le corpus CEMO a été segmenté en tours de parole ou en parties de tours de parole entre les locuteurs des conversations téléphoniques du SAMU. Ces interactions impliquent un ARM et une personne faisant l'appel. L'appelant peut être soit le patient lui-même, soit un tiers (famille, ami, collègue, voisin, inconnu). L'ensemble très large des types d'appelants (âge, sexe, origine), des accents (régionaux, étrangers), des différentes qualités vocales (altérations dues à l'alcool/médicaments, un rhume, etc.) en fait également un corpus très diversifié.

Les émotions sont des expériences activement construites. En effet, les émotions mixtes ou mélangées peuvent être perçues comme un épisode d'événements émotionnels distincts et liés (Hoemann, Gendron & Barrett 2017). Les prédictions se répercutent en cascade dans le cerveau jusqu'aux régions viscéromotrices, motrices et sensorielles primaires, contrôlant l'action et créant l'ensemble sensoriel complexe de l'expérience. Toutefois, nous ne percevons pas le mécanisme d'ajustement itératif continuellement à l'œuvre dans le cerveau ; nous ne faisons l'expérience que de ses résultats finaux, vivant les émotions comme des entités séparées susceptibles de se manifester simultanément, telles que la peur et le soulagement.

Les émotions comme l'*anxiété*, la *peur*, l'*inquiétude*, le *stress* sont les émotions majeures de ce corpus et ont fait l'objet d'une importante littérature. La différence entre *peur* et *anxiété* continue d'ailleurs d'être débattue par les neuroscientifiques (LeDoux 1996). Certains pensent connaître la différence en regardant le contexte social et interactionnel mais on ignore encore s'il s'agit d'états cérébraux distincts. Dans notre étude, *anxiété*, *stress*, *panique*, *déception*, *embarras* et *peur* sont incluses dans la macro *peur*. Parmi les annotations des appelants, 23 % sont des segments avec des émotions macro mélangées (*peur/positif*, *peur/tristesse*, *peur/neutre*), indiquant que le vote majoritaire sur les annotations n'a pas abouti à une émotion prédominante.

Tableau 1 : Les dix émotions et mélanges d'émotions les plus représentés dans les segments des appelants [a]

CEMO	Segments	Appelants
Total	*26 649*	*1 127*
PEU	8 993	1 036
NEU	7 560	908
POS	2 129	646
TRI	1 254	357
PEU/NEU	2 472	814

Tableau 1 : (suite)

PEU/POS	725	285
PEU/TRI	700	317
POS/NEU	685	418
TRI/NEU	277	171
AUTRES	1 854	1 076

a. (PEU)R, (NEU)TRE, (POS)ITIF, (TRI)STESSE, AUTRES (somme des classes restantes)

Dans cette étude, on se limitera aux 26 649 segments des appelants comme détaillé dans le Tableau 1. L'analyse de ce corpus montre de nombreuses combinaisons d'émotions (23 %), par exemple, la peur associée à la tristesse ou à des sentiments positifs comme le soulagement. Cette coexistence de peur peut être attribuée à l'expérience vécue par l'appelant, tandis que la tristesse ou le soulagement pourrait refléter ses réactions en réponse à l'interaction avec l'agent. Le contexte donné de la situation aide clairement à reconnaître les émotions. Le but de l'agent est de répondre à tous les appelants en essayant de déterminer l'aspect critique de la situation et l'urgence de l'action du centre mais sa façon de répondre peut engendrer des réactions émotionnelles de l'appelant qui se superposent à son émotion première. La prise en compte du contexte est très complexe. Savoir qu'une personne est tombée aide à comprendre sa douleur, savoir que c'est son fils qui appelle aide à comprendre la tristesse et la compassion mélangée dans la voix. Le contexte inclut donc la situation, la tâche, la personnalité de l'appelant, sa confiance dans l'agent, etc. Certaines de ces situations se répètent et engendrent des comportements particuliers.

Tableau 2 : Détails du sous-ensemble « CEMO simple » [a]

CEMO simple	PEU	NEU	POS	TRI	*Total*
#Segment audio	1 132	1 132	1 132	1 132	*4 528*
#Locuteurs	547	525	452	330	*918*
#Dialogues	507	481	433	305	*816*
Temps total	1h12min	51min	45min	1h9min	*3h57min*
Temps moyen	3,8s	1,6s	2,4s	3,7s	*2,9s*

a. (PEU)R, (NEU)TRE, (POS)ITIF, (TRI)STESSE, Total (nombre total de segments)

Tableau 3 : Détails du sous-ensemble « CEMO mixte » [a]

CEMO mélangée	PEU/NEU	PEU/POS	PEU/TRI	*Total*
#Segment audio	637	637	637	*1 911*
#Locuteurs	399	259	294	*652*
#Dialogues	380	248	278	*602*

Tableau 3 : (suite)

Temps total	27min	36min	43min	*1h46min*
Temps moyen	2,5s	3,4s	4,1s	*3,3s*

a. (PEU)R/(NEU)TRE, (PEU)R/(POS)ITIF, (PEU)R/(TRI)STESSE, Total (nombre total de segments)

Deux sous-ensembles ciblant les segments de 0,5 à 15 secondes ont été considérés :

– le premier, appelé « CEMO simple », détaillé dans le Tableau 2, regroupe les étiquettes émotionnelles uniques. Les segments sélectionnés portent sur les 4 émotions les plus fréquentes (cf. Tab. 1) pour les appelants (*peur*, *positif*, *tristesse*, *neutre*) ;
– le second, appelé « CEMO mixte », détaillé dans le Tableau 3, réunit les émotions de *peur* mélangées les plus fréquentes (*peur/neutre*, *peur/positif*, *peur/tristesse*) – cf. Tableau 1.

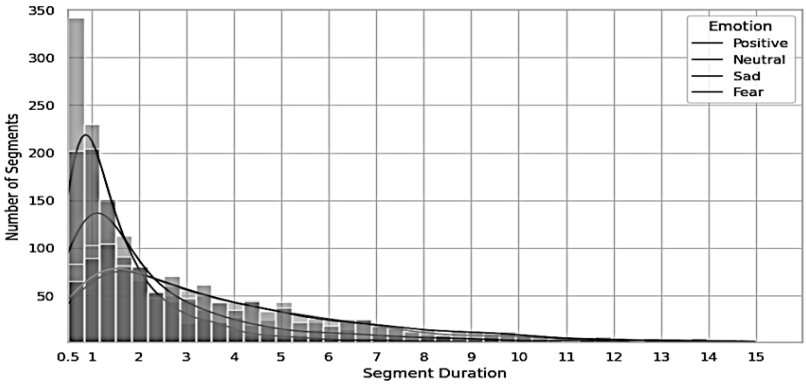

Figure 1 : Histogramme et estimation de densité par noyau des durées de segments par classe dans « CEMO simple »

La Figure 1 présente un histogramme combiné avec une estimation de densité par noyau pour illustrer la distribution des durées de segments des classes émotionnelles de « CEMO simple ». L'axe des abscisses représente la durée des segments en secondes, tandis que l'axe des ordonnées indique le nombre de segments correspondant à chaque intervalle de durée. Chaque couleur dans l'histogramme correspond à une classe émotionnelle spécifique : *positive*, *neutre*, *triste* et *peur*. Les barres superposées montrent le nombre de segments de chaque classe émotionnelle dans des plages de durée spécifiques. L'estimation de densité par noyau, qui est superposée à l'histogramme sous la forme de courbes lisses, est une méthode non paramétrique pour estimer la fonction de densité de probabilité d'un ensemble de données. Cette technique est particulièrement utile lorsque l'on souhaite obtenir une représentation lisse et continue de la distribution des données ; il est intéressant de constater que les distributions

pour les paires de classes émotionnelles *positif/neutre* et *peur/tristesse* montrent des similitudes.

4. MÉTHODOLOGIE

Nous nous sommes intéressés ici à la modalité acoustique des émotions. Une approche multimodale combinant les transcriptions et l'audio pourrait être plus performante (Deschamps-Berger, Lamel & Devillers 2023). Cette section présente les différentes approches explorées pour la détection des émotions à partir de données audio. Nous avons exploré à la fois des systèmes « classiques » à partir de connaissances expertes, qui reposent sur des caractéristiques prosodiques ou acoustiques extraites à partir du signal ou de spectrogrammes, ainsi que des systèmes plus récents fondés sur l'adaptation de modèles pré-entraînés. Toutes les expériences utilisent le même type de classifieur à base de réseaux de neurones dont les caractéristiques sont détaillées dans T. Deschamps-Berger, L. Lamel et L. Devillers (2023). Dans le processus d'optimisation de nos entraînements, nous avons adopté une stratégie de validation croisée comportant cinq plis. Cette méthode implique de diviser la totalité des données en cinq parties égales, communément appelés « plis ». À chaque itération, quatre plis servent à l'entraînement du modèle, tandis que le cinquième est réservé au test. Il est important de noter que, au sein du pli d'entraînement, une portion réduite est spécifiquement allouée à la validation. Ce mécanisme de validation croisée garantit que chaque pli des données est utilisé une fois en tant qu'ensemble de test, assurant ainsi une évaluation de l'ensemble des données. Nous veillons également à ce que les locuteurs soient distincts au sein des ensembles d'entraînement, de validation et de test, afin d'éliminer toute forme de biais liée à l'identité d'un locuteur dans la répartition des données. Le critère de sélection du meilleur modèle à l'issue de chaque phase d'entraînement repose sur sa performance sur l'ensemble de validation. Chaque modèle est ensuite soumis à une évaluation sur son ensemble de test, puis les prédictions de tous les modèles sont combinées pour recalculer globalement les scores d'évaluation, offrant une mesure collective de la performance sur l'ensemble des données de test. Les métriques d'évaluation sont l'*Accuracy* pour les émotions non mixtes. L*'Accuracy* est le pourcentage de prédictions correctes par rapport au total des prédictions, fournissant ainsi une mesure claire de la capacité du modèle à classer correctement chaque échantillon. Dans l'analyse des émotions mélangées, nous utilisons le *Score de Hamming* pour mesurer l'exactitude des prédictions du modèle par rapport aux étiquettes réelles, où une moyenne élevée indique une meilleure performance. Parallèlement, la *Perte de Hamming* évalue la fréquence des erreurs de prédiction, en incluant à la fois les fausses prédictions et les omissions. Un score proche de zéro signifie moins d'erreurs et donc une performance optimale du modèle.

4.1. Prétraitement de l'audio

Les données audio sont normalisées en utilisant une z-normalisation effectuée sur l'ensemble des corpus d'entraînement, de validation et de test. Cette normalisation ajuste chaque segment audio pour qu'il ait une moyenne de zéro et une variance unitaire, assurant ainsi une uniformité et minimisant les biais dus à des variations d'amplitude entre les enregistrements.

4.2. Approches classiques à partir de caractéristiques expertes

Deux approches « classiques » ont été testées à partir d'extraction d'indices experts pour détecter les émotions dans l'audio.

1. Modèle avec extractions d'indices prosodiques et acoustiques :

– **Caractéristiques d'entrée** : l'ensemble eGeMAPSv02 d'Opensmile (Eyben *et al*. 2016) est composé de 88 attributs prosodiques et acoustiques. Cet ensemble comporte des caractéristiques testées par des experts qui ont sélectionné un ensemble pertinent d'indices pour l'analyse des émotions. Voici les premiers 18 descripteurs prosodiques et acoustiques bas niveau (LLDs) de eGeMAPS, classés en paramètres de fréquence, d'énergie/amplitude et spectraux :
  - Fréquence : Pitch, *Jitter*, fréquence des trois premiers Formants et bande passante du premier Formant,
  - Énergie/amplitude : *shimmer*, *loudness*, ratio harmoniques-bruit (HNR),
  - Spectral : ratio Alpha, index de Hammarberg, pente spectrale (0-500 Hz et 500-1500 Hz), énergie relative des trois premiers Formants, et différences harmoniques (H1-H2, H1-A3),
  - Les 60 autres descripteurs sont décrits dans Eyben *et al*. (2016) ;
– **Encodeur** : une adaptation à des données audio du modèle convolutionnel VGG-16 (Simonyan & Zisserman 2015) est basée sur une succession de couches de convolutions et de normalisations. Cette version modifiée applique la convolution à une dimension sur les caractéristiques acoustiques, permettant une analyse des signaux temporels audio.

2. Modèle avec des spectrogrammes de Mel avec CNN Temporel et BLSTM, (Deschamps-Berger, Lamel & Devillers 2021) :

– **Caractéristiques d'entrée** : nous avons utilisé des spectrogrammes de Mel comme caractéristiques d'entrée. Pour les configurer, nous avons appliqué une *Transformée de Fourier à court terme* (STFT) avec une taille spécifique de 200 points. Cette taille correspond à une fenêtre de 25 millisecondes lorsqu'elle est appliquée à un signal audio échantillonné à 8 kHz. De plus, nous avons défini un intervalle de 80 points (équivalent à 10 millisecondes à 8 kHz) entre chaque fenêtre STFT. Cette approche nous permet de capturer les nuances temporelles et fréquentielles du signal audio, essentielles pour l'analyse des émotions dans la parole ;

– **Encodeur** : un réseau de neurones convolutif est conçu pour traiter les séquences temporelles suivi d'un BLSTM (*Bidirectional Long Short-Term Memory* ; Hochreiter & Schmidhuber 1997) qui prend en compte les données dans les deux directions temporelles, permettant une meilleure compréhension du contexte.

4.3. Modèles pré-entraînés sur l'audio

WAV2VEC 2.0 (Baevski *et al*. 2020), un modèle auto-supervisé, a été utilisé pour générer des représentations acoustiques. Le modèle 3k large doté de 311 millions de paramètres, issu de l'organisation leBenchmark (Evain *et al*. 2021) a été sélectionné pour ses similitudes entre sa base de données d'entraînement et les besoins spécifiques de notre étude et justifié dans des résultats précédents par la performance sur le corpus CEMO (Deschamps-Berger, Lamel & Devillers 2022). La base d'entraînement du modèle inclut des caractéristiques essentielles telles que des données en français, des dialogues spontanés, des données téléphoniques et un contenu émotionnel.

Procédure d'adaptation (*fine-tuning*). L'utilité de modèles pré-entraînés est de pouvoir adapter ceux-ci à des tâches en aval, comme la détection des émotions. Les couches de convolution et les projections linéaires au sein des couches d'attention à têtes multiples du modèle WAV2VEC 2.0 sont initialement adaptées à la reconnaissance automatique de la parole. Lors de la procédure d'adaptation les paramètres du modèle WAV2VEC 2.0 sont ajustés avec un entraînement sur les données CEMO. Cette adaptation permet de calibrer le modèle pour extraire des caractéristiques pertinentes à la détection des émotions dans la voix.

4.4. Métriques pour la classification des émotions simples *vs* mixtes

Dans le contexte de notre travail sur la classification des émotions, où nous traitons à la fois des cas de classification simple (une seule émotion par échantillon) et de classification multi-étiquettes (émotions mixtes – plusieurs émotions par échantillon), nous avons sélectionné des fonctions de perte spécifiques adaptées à chaque scénario :

– **Pour la classification simple des émotions**, nous avons opté pour la **fonction de perte d'entropie croisée**. Cette fonction mesure l'écart entre les probabilités prédites par notre modèle pour chaque classe (émotion) et les étiquettes réelles. Elle est idéale pour les tâches de classification où chaque échantillon appartient à une seule catégorie, car elle pénalise les prédictions qui s'éloignent de la véritable classe. L'objectif est de minimiser cette perte, ce qui signifie que les prédictions de notre modèle se rapprochent de plus en plus des étiquettes réelles.
– **Pour la classification multi-étiquettes des émotions**, nous utilisons **la somme des entropies croisées binaires de chaque classe**. Contrairement à la classification simple, où un échantillon appartient à une seule classe, la classification multi-étiquettes permet à un échantillon d'être associé à plusieurs classes en

même temps. L'entropie croisée binaire est calculée séparément pour chaque classe comme si chaque prédiction était un problème de classification binaire (la classe est présente ou non). Ensuite, nous additionnons les entropies croisées binaires obtenues pour toutes les classes pour obtenir la perte totale. Cette approche permet de traiter individuellement la présence ou l'absence de chaque émotion dans les échantillons, assurant ainsi que le modèle peut apprendre à reconnaître plusieurs émotions simultanément.

L'évaluation de la performance des modèles dans des tâches de classification multi-étiquettes nécessite des métriques capables de saisir la complexité des prédictions exactes et partielles. Nous utilisons trois métriques principales :

– **Ratio de Correspondance Exacte (Exact Match Ratio – MR)** : cette métrique évalue la proportion de prédictions parfaitement exactes, où toutes les étiquettes prédites correspondent entièrement aux étiquettes réelles. Elle est calculée comme suit :

$$\text{Exact Match Ratio, MR} = \frac{1}{n}\sum_{i=1}^{n} I\left(y_i = \hat{y}_i\right)$$

Cette métrique ne fait pas de distinction entre les réponses totalement incorrectes et les réponses partiellement correctes, donc nous utilisons également le *Score de Hamming* et la *Perte de Hamming*.

– **Score de Hamming ou *Accuracy* (Hamming Score – HS)** : l'*Accuracy* pour chaque classe émotionnelle est définie comme la proportion des étiquettes correctes prédites par rapport au nombre total (prédit et réel) d'étiquettes pour cette classe. Le *Score de Hamming* est défini comme la moyenne de ces *Accuracy* :

$$\text{Accuracy} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i \cap \hat{y}_i|}{|y_i \cup \hat{y}_i|}$$

– **Perte de Hamming (Hamming Loss – HL)** : elle indique combien de fois en moyenne la pertinence d'un exemple par rapport à l'étiquette d'une classe est mal prédite. Par conséquent, la *Perte de Hamming* prend en compte l'erreur de prédiction (une étiquette incorrecte est prédite) et l'erreur manquante (une étiquette pertinente n'est pas prédite), normalisées sur le nombre total de classes et le nombre total d'exemples. La *Perte de Hamming* est définie comme suit :

$$\text{Hamming Loss} = \frac{1}{nL}\sum_{i=1}^{n}\sum_{j=1}^{L} I\left(y_i^j \neq \hat{y}_i^j\right)$$

Idéalement, la *Perte de Hamming* devrait être égale à 0, ce qui signifierait qu'il n'y a pas d'erreur.

5. EXPÉRIENCES

Nous avons mené une série d'expérimentations pour évaluer la capacité d'un système basé sur l'architecture *Transformer* à détecter des états émotionnels non mixtes et mixtes. Nous avons également comparé l'approche *Transformer* pré-entraîné avec des approches « classiques » employant des réseaux de convolutions et des réseaux de neurones récurrents.

5.1. Expériences sur les émotions simples

Le modèle a d'abord été ajusté sur des segments où l'émotion exprimée est clairement identifiée par les annotateurs dans le sous-ensemble appelé « CEMO simple » limité à 4 macro-classes : *peur*, *neutre*, *positif*, *tristesse*.

Tableau 4 : Résultats des expériences, présentant l'*Accuracy* (ACC en %) sur le sous-ensemble « CEMO simple », décomposée par taux de rappel par émotion (en %) [a]

Features	Encodeur	ACC	PEU	NEU	POS	TRI
App. "classiques" :						
eGeMAPSv02	VGG-16 [27]	37.1	32.7	62.3	6.5	47.0
MelSpectr.	CNN + BLSTM [19]	37.9	56.5	56.2	7.1	31.6
Transformer :						
Audio brut	w2v2-Fr-3K param. figés	35.9	16.8	66.4	23.4	37.0
Audio brut	w2v2-Fr-3K + CEMO	56.8	51.0	70.1	55.0	51.0

a. Peur (PEU), Neutre (NEU), Positif (POS), Tristesse (TRI), Approches (app.), paramètres (param.)

« CEMO-simple » comprend un grand nombre de locuteurs et une grande diversité d'expressions émotionnelles souvent mélangées mais aussi pondérées dans un corpus de taille réduite (21h). Nous avons exploré des approches « classiques » pour la reconnaissance des émotions détaillées dans le Tableau 4 qui ont donné des résultats mitigés :

- l'ensemble de paramètres GeMAPS v0.2, qui comprend 88 caractéristiques, a été utilisé avec le modèle de réseau de neurones convolutif VGG-16, adapté en une dimension. Cette approche a abouti à une *Accuracy* de 37.1 % ;
- suivant les travaux précédents adoptés dans T. Deschamps-Berger, L. Lamel et L. Devillers (2021), nous avons utilisé des Mel spectrogrammes combinés à des réseaux convolutifs temporel suivis de réseaux récurrents (Temporal CNN-BLSTM). Cette technique a permis d'obtenir une *Accuracy* du même ordre à 37.9 %.

Nous avons donc cherché à améliorer ces résultats en utilisant une approche de type *Transformer* spécifique à l'audio (WAV2VEC2) pré-entraîné sur un grand corpus en français :

- nous avons sélectionné le modèle WAV2VEC2 pré-entraîné sur 3 000 heures de français (Evain *et al.* 2021), figé les paramètres du modèle et entraîné un

classifieur sur le sous-ensemble « CEMO simple ». Cette expérience a produit une *Accuracy* de 35.9 % ;

– dans un second temps, nous avons entièrement adapté (*fine-tuning*) le modèle WAV2VEC2 pré-entraîné sur 3 000 heures de français (Evain *et al*. 2021) avec le sous-ensemble « CEMO simple », obtenant 56.8 % d'*Accuracy*.

Les résultats, rapportés dans le Tableau 4, montrent de meilleurs résultats avec **WAV2VEC2-FR-3K** adapté sur CEMO en termes d'*Accuracy* globale face aux approches « classique ». Ils illustrent également la valeur ajoutée du pré-entraînement sur 3 000 heures de français du modèle WAV2VEC2 ainsi que l'importance de son adaptation au corpus d'étude. Il y a également une nette amélioration dans la reconnaissance des émotions spécifiques (*peur*, *neutre*, *positif*, *tristesse*). Par exemple, le rappel des émotions de *peur* passe de 16.8 % avec le WAV2VEC2-FR-3K figé à 51 % avec le modèle adapté, démontrant l'effet bénéfique d'une adaptation ciblée. De même, la reconnaissance des émotions *positives* et de la *tristesse* sont améliorées significativement, passant de 23.4 % et 37 % à 55 % et 51 % d'*Accuracy*.

5.2. Expériences sur les émotions mixtes

Nous avons sélectionné un corpus équilibré d'émotions mixtes liées à la *peur*, les plus pertinentes et fréquentes dans les 23 % du corpus total que nous avons appelé « CEMO mixte ». Nous avons testé la capacité du modèle *Transformer* à les détecter (cf. Tab. 5). Cette phase visait à évaluer la robustesse du modèle à la détection des mélanges émotionnels, une tâche proche mais néanmoins plus complexe que la prédiction des émotions simples.

Tableau 5 : Résultats des expériences sans et avec adaptation de WAV2VEC2-FR-3K sur « CEMO mixte », présentant le taux de correspondance exacte (MR), le *Score de Hamming* (HS) et la *Perte de Hamming* (HL) sur le sous-ensemble CEMO mélangée, décomposée par taux de rappel par émotion (en %) [a]

Corpus	Adaptation	MR (%)	HS (%)	HL	PEU	NEU	POS	TRI
CEMO mixte	Sans	29.3	62.3	0.3	80.4	62.5	76.1	77.6
CEMO mixte	Avec	49.9	69.7	0.2	100	80.9	49.1	20.4

a. Peur (PEU), Neutre (NEU), Positif (POS), Tristesse (TRI)

Le test réalisé sur des émotions mélangées montre une réduction significative du taux de correspondance exacte (MR), qui atteint 29.3 %, par rapport à un taux d'*Accuracy* de 56.8 % observé pour les émotions simples. Cette baisse marque l'impact de la complexité accrue des émotions mélangées sur la performance du modèle. Cependant, il est important de noter que la métrique MR, par sa nature stricte, ne nous indique pas si le modèle est loin ou non des annotations réelles. Le *Score de Hamming* de 62.3 % et la *Perte de Hamming* de 0,3 nous indiquent

des mesures plus nuancées de la performance du modèle sur les émotions mélangées, reflétant sa capacité à prédire correctement les éléments individuels d'un ensemble d'émotions, même en cas de correspondance partielle.

Nous avons ensuite adapté (*fine-tuning*) le modèle, initialement entraîné sur les émotions simples, pour détecter les émotions mélangées du corpus mixte, afin d'affiner la discrimination des états émotionnels co-occurrents par le modèle. Cette adaptation a montré un taux de correspondance exacte (MR), qui s'élève désormais à 49.9 %, et du *Score de Hamming* (HS), qui atteint 69.7 %. Cependant, cette amélioration est biaisée par le fait que tous les échantillons du corpus mixte contiennent de la peur. Cette observation suggère une adaptation superficielle du modèle aux complexités des émotions mixtes. De nouveaux tests avec plus de classes mélangées seront menés.

6. CONCLUSION

Cette étude a exploré la reconnaissance des émotions à travers des approches « classique » et des modèles de type *Transformer* pré-entraîné, avec une attention portée aux émotions simples et mixtes. Les résultats montrent que le *Transformer* WAV2VEC2 pré-entraîné sur le français, combiné à l'adaptation sur nos données, donne de meilleurs résultats que les approches « classiques » en termes *d'Accuracy* pour la reconnaissance des émotions simples. En revanche, la tâche de reconnaissance des émotions mixtes a révélé des défis supplémentaires, marqués par une baisse de performance en termes de taux de correspondance exacts. Une première tentative d'ajustement du modèle aux émotions mixtes a révélé les défis associés aux mesures sur ce type d'émotions. Ces observations mettent en lumière l'importance qu'il y a à développer des stratégies d'annotation contextuelle, d'entraînement et d'adaptation plus sophistiquées pour aborder la complexité des émotions ‹*in the wild*›. Elles suggèrent également la nécessité d'une analyse plus détaillée pour comprendre les mécanismes par lesquels les modèles apprennent à reconnaître des états émotionnels complexes.

En conclusion, cette recherche souligne le potentiel des modèles *Transformer* dans le domaine de la reconnaissance automatique des émotions. Toutefois, elle révèle également leurs limites face à la complexité des émotions mixtes et l'importance d'intégrer un contexte supplémentaire, lié soit au dialogue, à la personne ou à la situation, et ainsi de pouvoir correctement détecter les nuances des expressions émotionnelles humaines qui sont plus souvent qu'on ne le pense mélangées dans les données ‹*in the wild*›.

Références

ADOLPHS R. & ANDERSON D. J. (2018), *The Neuroscience of Emotion: A New Synthesis*, Princeton (NJ), Princeton University Press.

ADOLPHS R. (2017), "How should neuroscience study emotions? By distinguishing emotion states, concepts, and experiences", *Social Cognitive and Affective Neuroscience* 12 (1), 24-31.

BAEVSKI A. *et alii* (2020), "Wav2vec 2.0: A framework for self-supervised learning of speech representations", *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver BC, Canada), Red Hook (NY), Curran Associates Inc., 12449-12460.

BARRETT L. F. (2017), "The theory of constructed emotion: An active inference account of interoception and categorization", *Social Cognitive and Affective Neuroscience* 12 (1), 1-23.

CHEN M. & ZHAO X. (2020), "A multi-scale fusion framework for bimodal speech emotion recognition", *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech 2020)* (Shanghai, China), ISCA, 374-378.

CHENG J., DONG L. & LAPATA M. (2016), "Long short-term memory-networks for machine reading", *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (Austin, Texas), 551-561, ACL, arXiv:1601.06733v7.

CHOU H.-C. *et alii* (2022), "Exploiting annotators' typed description of emotion perception to maximize utilization of ratings for speech emotion recognition", *Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)* (Virtual), Danvers (MA), IEEE, 7717-7721.

DAMASIO A. R. (2018), *The Strange Order of Things: Life, Feeling, and the Making of Cultures*, First edition, New York, Pantheon Books.

DESCHAMPS-BERGER T., LAMEL L. & DEVILLERS L. (2021), "End-to-End speech emotion recognition: Challenges of real-life emergency call centers data recordings", *Proceedings of the 9th International Conference on Affective Computing and Intelligent Interaction (ACII 2021)* (Nara, Japan), Danvers (MA), IEEE, arXiv: 2110.14957v1.

DESCHAMPS-BERGER T., LAMEL L. & DEVILLERS L. (2022), "Investigating transformer encoders and fusion strategies for speech emotion recognition in emergency call center conversations", *ICMI'22 Companion: Companion Publication of the 2022 International Conference on Multimodal Interaction* (Bengaluru, India), New York (NY), Association for computing Machinery, 144-153.

DESCHAMPS-BERGER T., LAMEL L. & DEVILLERS L. (2023), "Exploring attention mechanisms for multimodal emotion recognition in an emergency call center corpus", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)* (Rhodes Island, Greece), Danvers (MA), IEEE, arXiv: 2306.07115v1.

DEVILLERS L., VIDRASCU L. & LAMEL L. (2005), "Challenges in real-life emotion annotation and machine learning based detection", *Neural Networks* 18 (4), 407-422.

DEVLIN J. *et alii* (2019), "BERT: Pre-training of deep bidirectional transformers for language understanding", *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Minneapolis, Minnesota), 4171-4186, ACL, arXiv: 1810.04805v2.

EKMAN P. (1973), "Universal facial expressions in emotion", *Studia Psychologica* 15 (2), 140-147.

ETIENNE C. *et alii* (2018), "CNN+LSTM architecture for speech emotion recognition with data augmentation", *Proceedings of the Workshop on Speech, Music and Mind (SMM 2018)* (Hyderabad, India), ISCA, arXiv:1802.05630v2.

EVAIN S. *et alii* (2021), "LeBenchmark: A reproducible framework for assessing self-supervised representation learning from speech", *Proceedings of the Annual Conference of the*

International Speech Communication Association (Interspeech 2021) (Brno, Czechia), 1439-1443, ISCA, arXiv:2104.11462.

EYBEN F. *et alii* (2016), "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing", *IEEE Transactions on Affective Computing* 7 (2), 190-202.

GIRAUD T. *et alii* (2013), "Multimodal expressions of stress during a public speaking task: Collection, annotation and global analyses", *Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII 2013)* (Geneva, Switzerland), Danvers (MA), IEEE, 417-22.

HAN J. *et alii* (2017), "From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty", *MM'17: Proceedings of the 25th ACM International Conference on Multimedia*, New York (NY), Association for Computing Machinery, 890-897.

HOCHREITER S. & SCHMIDHUBER J. (1997), "Long short-term memory", *Neural Computation* 9 (8), 1735-1780.

HOEMANN K., GENDRON M. & BARRETT L. F. (2017), "Mixed emotions in the predictive brain", *Current Opinion in Behavioral Sciences* 15, 51-57.

HUA J. *et alii* (2016), "Predicting a failure of public speaking performance using multidimensional assessment", *International Journal of Sports Science and Coaching* 4 (4), 197-209.

LE H. *et alii* (2020), "FlauBERT: Unsupervised language model pre-training for French", *Proceedings of the Twelfth Language Resources and Evaluation Conference* (Marseille, France), 2479-2490, ELRA, arXiv:1912.05372v4.

LEDOUX J. E. (1996), *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*, New York (NY), Simon & Schuster.

RUSSELL J. A. (1994), "Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies", *Psychological Bulletin* 115 (1), 102-141.

SCHERER K. R., SCHORR A. & JOHNSTONE T. (2001), *Appraisal Processes in Emotion: Theory, Methods, Research*, Oxford, Oxford University Press.

SIMONYAN K. & ZISSERMAN A. (2015), "Very deep convolutional networks for large-scale image recognition", *3rd International Conference on Learning Representations (ICLR 2015)* (San Diego, USA), arXiv:1409.1556v6.

TRIGEORGIS G. *et alii* (2016), "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network", *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)* (Shanghai, China), Danvers (MA), IEEE, 5200-5204.

VASWANI A. *et alii* (2017), "Attention is all you need", *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach California, USA), Red Hook (NY), Curran Associates Inc., 6000-6010.

VIDRASCU L. & DEVILLERS L. (2005), "Detection of real-life emotions in call centers", *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech 2005)* (Lisbon, Portugal), ISCA, 1841-1844.

WÖLLMER M. *et alii* (2008), "Abandoning emotion classes – Towards continuous emotion recognition with modelling of long-range dependencies", *Proceedings of the 9th Annual Conference of the International Speech Communication Association (Interspeech 2008)* (Brisbane, Australia), ISCA, 597-600.

ABSTRACTS

Laurence Devillers, Théo Deschamps-Berger, Lori Lamel, *Emotions* ‹in the Wild› *of Emergency Call Center Callers: Towards a Speech Emotion Recognition System*

This contribution presents a study on the detection of emotions and mixtures of emotions in a corpus collected from an emergency call center in Paris (CEMO). Our corpus, recorded ‹in the wild›, is rich in voice diversity (age, accent, number of speakers) and is annotated with an original scheme that represents up to two emotions per segment. Tests with systems using audio-specific Transformers adapted to the CEMO corpus on a portion of CEMO's unmixed emotions obtained a detection score (*Accuracy)* of 56.7% for 4 classes (fear, neutral, positive, sadness) surpassing those obtained with more classical approaches based on expert prosodic features. Additional tests were carried out on a portion of the CEMO corpus with mixed emotions, highlighting some of the outstanding challenges, in particular how to take into consideration the context of the interaction.

Keywords : emotions ‹*in the wild*›, emergency call center, transformer, classification of mixed emotion

RÉSUMÉS

Laurence Devillers, Théo Deschamps-Berger, Lori Lamel, *Les émotions* ‹in the wild› *des appelants d'un centre d'appels d'urgence : vers un système de détection des émotions dans la voix*

Cette contribution présente une étude sur la détection d'émotions et de mélanges d'émotions dans un corpus collecté dans un centre d'appels d'urgence à Paris (CEMO). Notre corpus, enregistré ‹in the wild›, est riche en diversité vocale (âge, accent, nombre de locuteurs) et est annoté avec un schéma original qui représente jusqu'à deux émotions par segment. Des tests avec des systèmes utilisant des *Transformers* audio spécifiques adaptés à CEMO sur une partie des émotions non mixtes ont permis d'obtenir un score de détection (*Accuracy*) de 56.7 % pour 4 classes (peur, neutre, positif, tristesse) surpassant ceux obtenus avec des approches plus classiques basées sur des caractéristiques prosodiques expertes. Des tests supplémentaires ont été effectués sur une partie de CEMO avec des émotions mixtes, mettant en évidence certains des défis à relever, en particulier la prise en compte du contexte de l'interaction.

Mots-clés : émotions ‹*in the wild*›, centre d'appels d'urgence, *transformer*, classification des émotions mélangées