Premier système IRIT-MyFamillyUp pour la compétition sur la reconnaissance des émotions Odyssey 2024

Adrien Lafore, Clément Pagés, Leila Moudjari, Sebastião Quintas, Isabelle Ferrané, Hervé Bredin, Thomas Pellegrini, Farah Benamara, Jérôme Bertrand, Marie-Françoise Bertrand, et al.

Premier système IRIT-MyFamillyUp pour la compétition sur la reconnaissance des émotions Odyssey 2024

Adrien Lafore[1, 2] Clément Pagés[1] Leila Moudjari[1] Sebastiao Quintas[1]
Isabelle Ferrané[1] Hervé Bredin[1] Thomas Pellegrini[1] Farah Benamara[1]
Jérôme Bertrand[2] Marie-Françoise Bertrand[2]
Véronique Moriceau[1] Jérôme Farinas[1]
(1) IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France
(2) My Family Up, Toulouse, France
prenom.nom@irit.fr
mf.bertrand@myfamilyup.com j.bertrand@myfamilyup.com

RÉSUMÉ

Dans cet article, nous présentons notre contribution à la tâche de classification des émotions dans la parole dans le cadre de notre participation à la campagne d'évaluation Odyssey 2024. Nous proposons un système hybride qui tire parti à la fois des informations du signal audio et des informations sémantiques issues des transcriptions automatiques. Les résultats montrent que l'ajout de l'information sémantique permet de dépasser les systèmes uniquement audio.

ABSTRACT

IRIT-MyFamilyUp system for the Odyssey 2024 Emotion Recognition Challenge.

In this paper, we present our contribution to emotion classification in speech as part of our participation in the Odyssey 2024 challenge. We propose a hybrid system that takes advantage of both audio signal information and semantic information from automatic transcriptions. The results show that adding semantic information allows surpassing systems based solely on audio.

MOTS-CLÉS : modélisation des émotions, Compétition Odyssey 2024, fusion texte et audio.

KEYWORDS: emotion modelling, Odyssey 2024 challenge, text and audio fusion.

1 Introduction

Un aidant familial ou proche aidant est une « personne qui vient en aide, de manière régulière et fréquente, à titre non professionnel, pour accomplir tout ou partie des actes ou des activités de la vie quotidienne d'une personne en perte d'autonomie, du fait de l'âge, de la maladie ou d'un handicap » (Loi n°2015-1776 article 51, 2015). En 2021, on estime à 11 millions le nombre d'aidants en France, soit un français sur six. D'après un rapport du ministère de l'économie (Ministère de l'économie des finances et de la relance, 2021), les principales observations sont :

— leur âge moyen est de 49 ans et 37% des aidants sont âgés de 50 à 54 ans ; 60% des aidants sont des femmes ;
— 69 % des aidants constatent un impact réel sur leur état moral,
— 53 % des aidants subissent des effets sur leur propre santé ;

— 50 % se sentent parfois seuls, non soutenus moralement ;
— 62 % se sont déjà retrouvés dans un état d'épuisement intense.

D'après l'OMS, 70% des jeunes aidants présentent des troubles anxio-dépressifs et 15 à 30% des personnes âgées souffrent de dépression. En 2021, 75% des français estiment qu'il faut du « courage » pour aller voir un psychologue (YouGov, 2019) et des enquêtes de la DREES [1] ont mis en évidence la demande des citoyens français pour des solutions de soutien psychologique personnalisées, professionnelles et accessibles par internet. La thérapie en ligne est souvent le seul soin possible, mais la majorité des applications de thérapie ne sont pas fiables dans une logique thérapeutique. C'est dans ce contexte que nous nous intéressons à l'identification des sentiments et états émotionnels dans la communication orale afin de développer un détecteur d'états émotionnels dans la parole, qui permettrait d'aider au diagnostic psychologique des proches aidants. Notre objectif est d'exploiter des informations audio et sémantiques afin de modéliser les états émotionnels par la détection de ces états dans la sémantique textuelle (parole retranscrite) et la détection d'émotions dans la prosodie.

Les émotions ont largement été étudiées dans un cadre théorique. On peut citer notamment les modèles de représentation de Plutchik (Plutchik, 1980) ou d'Ekman (Ekman & Journet, 2002) pour les plus connus.

Dans le cadre du Traitement Automatique des Langues (TAL), la détection des états émotionnels ou psychologiques a surtout été abordée dans le cadre de la détection des maladies mentales comme la dépression, les troubles de l'alimentation ou les troubles bipolaires (cf. (Harrigian *et al.*, 2021) pour une revue des collections de données et des tâches existantes). Les campagnes d'évaluation Computational Linguistics and Clinical Psychology (CLPsych) (Zirikly *et al.*, 2022) ou eRisk (Parapar *et al.*, 2023) se sont penchées en particulier sur la tâche de détection automatique de la dépression chez des utilisateurs des réseaux sociaux, avec un focus sur la détection "au plus tôt". Les modèles d'apprentissage développés pour la détection automatique sont à base soit de traits pour les plus performants (utilisation de pronoms personnels, sentiment positif ou négatif, temps des verbes, etc.) (Bae *et al.*, 2021; Molina *et al.*, 2023), soit d'apprentissage profond (cf. (Ríssola *et al.*, 2021) pour un panorama des méthodes automatiques existantes pour la classification des états mentaux sur les réseaux sociaux). Les travaux actuels dans ce domaine portent ainsi quasi uniquement sur des données écrites par des utilisateurs plutôt jeunes des réseaux sociaux.

En ce qui concerne le Traitement Automatique de la Parole (TAP), la détection d'émotion est un domaine de recherche qui est issu de l'analyse automatique de la prosodie. En effet, le champ des informations non verbales constitue la source des informations pour caractériser les émotions. Des compétitions internationales ont eu lieu depuis 2009 afin de faire avancer la connaissance sur cette problématique : The Interspeech 2009 Emotion Challenge (Schuller *et al.*, 2009) et The Interspeech 2010 Paralinguistic challenge (Schuller *et al.*, 2010). Les premiers systèmes étaient basés sur des systèmes discriminants alimentés par de nombreux paramètres extraits du signal audio, comme la boîte à outil OpenSmile (Eyben *et al.*, 2010). Les systèmes actuels se basent sur des architectures de réseaux de neurones profonds et ont permis d'obtenir de bonnes performances en reconnaissance des émotions sous forme de catégories (Tripathi *et al.*, 2018; Yenigalla *et al.*, 2018; Atmaja *et al.*, 2019; Yoon *et al.*, 2019). L'enjeu consiste maintenant à projeter les émotions dans un espace continu, ce qui permettra l'étude des états mentaux cognitifs (Atmaja & Akagi, 2020).

Le travail présenté dans cet article détaille une première contribution dans le cadre de notre participa-

1. https://drees.solidarites-sante.gouv.fr/sources-outils-et-enquetes/les-enquetes-capacites-aides-et-ressources-des-seniors-care

tion à la campagne d'évaluation Odyssey 2024 [2] portant sur la reconnaissance des émotions. Nous proposons ici un travail conjoint qui vise à utiliser des techniques de l'état de l'art pour combiner l'exploitation de données issues des retranscriptions d'extraits vidéo collectés sur internet et des données issues de la piste audio.

Dans les sections suivantes, nous présentons d'abord le corpus mis à notre disposition ainsi que la tâche de classification proposée. Ensuite, nous présentons en détail l'architecture de notre système. Enfin, nous détaillons et analysons les résultats obtenus.

2 Protocole expérimental

Nous présentons ici la campagne d'évaluation Odyssey 2024, à savoir : les données, la tâche proposée ainsi que les métriques d'évaluation utilisées.

2.1 Corpus

Les données mises à disposition sont des enregistrements en anglais issus du corpus MSP-Podcast (Lotfian & Busso, 2019), qui contient des segments audio provenant de podcasts en ligne. Les tours de parole ont été annotés par au moins 5 annotateurs selon les catégories d'émotion et leurs dimensions.

Les catégories d'émotions annotées dans ce corpus sont : Anger (colère), Contempt (mépris), Disgust (dégoût), Fear (peur), Happiness (bonheur), Neutral (neutre), Sadness (tristesse), Surprise, Other (autre), et No agreement (pas d'accord inter-annotateurs).

Les dimensions pour chaque émotion sont la *valence* (état positif ou négatif de l'individu), l'*arousal* (activité ou passivité de l'individu) et la *dominance* (contrôle faible à fort). Chacune de ces dimensions est annotée sur une échelle de 1 à 7.

Les données d'entraînement et de développement sont composées respectivement de 68 360 et 19 815 tours de parole annotés. Pour ces données, les transcriptions sont fournies ainsi que le genre des intervenants. Les données de l'ensemble de test sont constituées de 2 347 segments de parole venant de 187 personnes. Pour ces dernières, aucune transcription n'est fournie. De plus, les classes "Other" (O) et "No agreement" (X) ont été supprimées et la distribution des catégories d'émotion est équilibrée. Ainsi, nous avons aussi retiré ces deux classes du jeu de données d'entraînement et de développement. Le tableau 1 montre la distribution des données d'entraînement et de développement.

2.2 Tâche et métrique d'évaluation

La tâche à laquelle nous avons participé est celle de classification des émotions en 8 catégories : colère (A), mépris (C), dégoût (D), peur (F), bonheur (H), neutre (N), tristesse (S), et surprise (U).

La campagne d'évaluation n'autorise pas l'utilisation de modèles existants entraînés pour la détection d'émotions.

Les systèmes participants sont évalués selon les mesures classiques de précision, rappel, F1-score et

2. https://www.odyssey2024.org/emotion-recognition-challenge

Catégorie	Entraînement			Développement		
	Nombre	%	Durée totale	Nombre	%	Durée totale
Neutral (N)	25 106	36,72	39h43	5 667	28,60	08h47
No agreement (X)	13 709	20,05	22h02	4 013	20,25	06h32
Happiness (H)	13 440	19,66	22h09	3 340	16,86	05h14
Sadness (S)	3 882	5,68	06h13	1 101	5,56	01h44
Anger (A)	3 053	4,47	05h05	2 413	12,18	04h01
Surprise (U)	2 897	4,24	04h38	729	3,68	01h04
Contempt (C)	2 443	3,57	04h09	1 323	6,67	02h17
Disgust (D)	1 426	2,09	02h24	486	2,45	00h52
Other (O)	1 265	1,85	02h05	461	2,33	00h46
Fear (F)	1 139	1,67	01h46	282	1,42	00h26
TOTAL	68 360		110h14	19 815		32h06

TABLE 1 – Distribution des catégories d'émotion dans les données d'entraînement et de développement

accuracy. La distribution des classes dans les données de test étant équilibrée, la macro-F1 est utilisée pour classer les systèmes ; la macro-F1 étant la moyenne des F1-scores pour chacune des 8 classes (voir formule 1).

Pour comparer les performances de nos systèmes, nous avons durant ce challenge utilisé un jeu de données équilibré issu du jeu de développement fourni par les organisateurs de ce challenge. Nous l'avons construit par échantillonage de façon à calculer les performances de nos systèmes dans le même cadre que le jeu de Test qui lui aussi est équilibré.

$$\text{Macro F1} = \frac{1}{8}\sum_{i=1}^{8} 2 \times \frac{\text{précision}_i \times \text{rappel}_i}{\text{précision}_i + \text{rappel}_i} \tag{1}$$

3 Système proposé

Dans le cadre de la tâche de classification des émotions, nous proposons un premier système combinant les informations prosodiques et sémantiques à notre disposition. Dans ce premier système hybride, le but est de calculer les probabilités d'émotions pour chaque fichier en entrée (audio et texte) et d'utiliser la moyenne de ces probabilités pour notre prédiction. Les données de test pour ce challenge étant uniquement audio, nous avons dû utiliser un premier système de reconnaissance de la parole (cf. section 3.2) qui fournit les transcriptions au modèle sémantique entraîné pour cette tâche. En parallèle, le modèle audio (cf. section 3.1), lui aussi entraîné pour cette tâche, prend les segments audio fournis (d'une durée entre 3 et 11 secondes) pour calculer de son côté les probabilités d'émotion.

3.1 Modélisation acoustique

Le système dédié à la tâche de classification d'émotions en se basant uniquement sur l'audio a été développé et entraîné à partir de la librairie pyannote.audio (Bredin, 2023; Plaquet & Bredin, 2023), et

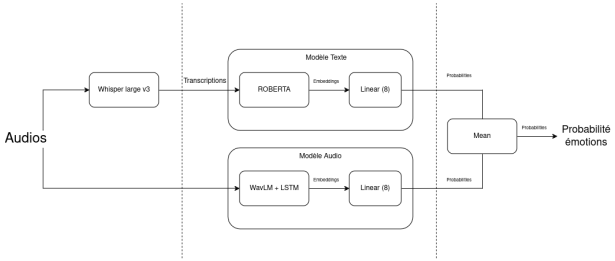


FIGURE 1 – Premier système hybride proposé

est inspiré de l'architecture de SSerRiouSS [3]. Le système prend en entrée des morceaux d'audio d'une durée de 10 secondes qui sont tout d'abord traités par un modèle WavLM-large (Chen *et al.*, 2022) pré-entraîné sur le corpus de LibriSpeech. La sortie fournie par ce module correspond à la moyenne pondérée des sorties de chacune des 12 couches composant le modèle WavLM-large, les poids utilisés pour cette moyenne étant appris durant l'entraînement. Cette sortie est ensuite injectée dans une pile de LSTM bidirectionnels (BLSTM) parcourant la séquence de caractéristiques produite par le modèle dans les deux sens. La séquence de trames issue de cette pile de BLSTM est ensuite passée à travers une couche linéaire dont le but est de faire la classification des émotions au niveau de chaque trame. L'étape suivante consiste à calculer la moyenne de ces classifications à l'aide d'une couche de mise en commun (mean pooling), la sortie de celle-ci étant finalement donnée à une fonction d'activation de type log-softmax, associant à chaque émotion possible une probabilité.

Ce modèle a été entraîné sur la partition d'entraînement du corpus MSP-PODCAST. Le nombre de BLSTM a été fixé à 2. Ce nombre correspond à la valeur par défaut dans l'architecture de SSeRiousSS, et également à celle pour laquelle les meilleurs résultats ont été obtenus. L'entraînement du système a été effectué à l'aide de l'optimiseur Adam, avec un taux d'apprentissage initial de 10e-3, ce dernier étant divisé par deux sans amélioration du F1-score macro sur 10 epochs consécutives, jusqu'à une valeur minimale de 10e-8. Le modèle WavLM-large a été gelé. La fonction de perte utilisée pour l'entraînement est une Vraisemblance logarithmique négative, pondérée avec des poids inversement proportionnels au nombre de représentants de chaque classe. La taille des lots a été fixée expérimentalement à 32.

3.2 Transcription de la parole

Des transcriptions manuelles ont été fournies pour les sous-ensembles d'entraînement et de développement, mais pas pour le jeu de test. Nous avons donc eu besoin d'utiliser un système de reconnaissance automatique de la parole (RAP) afin de fournir des transcriptions au modèle texte de classification d'émotions. Nous avons utilisé le système Whisper (Radford *et al.*, 2022), et en particulier le modèle whisper-large-v3. Ce modèle a été entraîné avec 680 000 heures de données supervisées multilingues et multitâches collectées sur le Web, un aspect qui le rend robuste à différents types d'accents et conditions acoustiques.

Le système de RAP génère des transcriptions forcément différentes des transcriptions manuelles fournies, et il peut en résulter une perte de performance si trop d'erreurs de transcription sont

3. Le code de cette architecture est accessible ici : https://github.com/pyannote/pyannote-audio/blob/develop/pyannote/audio/models/segmentation/SSeRiouSS.py

commises. Pour tenter d'évaluer la similarité entre les deux types de transcription, nous avons mesuré des taux d'erreur mot (TEM) sur les jeux d'entraînement et de développement. Avec le modèle whisper-large-v3, ce taux est d'environ 32% sur les deux jeux. Notons que nous avons testé d'autres modèles Whisper, plus petits (de *tiny* à *medium*), et le TEM augmentait inversement avec la taille du modèle.

Whisper fournit des transcriptions qui contiennent de la ponctuation et des majuscules pour les noms propres et autres acronymes. Si nous les normalisons (suppression des signes de ponctuation et de la casse), le TEM est drastiquement réduit à 12-13% sur les deux sous-ensembles. Nous avons choisi de garder la ponctuation et la casse cependant pour la modélisation des émotions, car a priori ces éléments donnent des informations probablement pertinentes pour cette tâche, comme par exemple un point d'exclamation qui peut exprimer de la surprise ou de la colère.

3.3 Modélisation des transcriptions

N'étant pas autorisés à utiliser des modèles dédiés à la détection d'émotions, nous avons entraîné et évalué plusieurs modèles de langue de l'état de l'art sur les transcriptions de référence des données d'entraînement et de développement : BERT (base, multilingual, large, etc.) (Devlin *et al.*, 2019) et RoBERTa-base (Liu *et al.*, 2019). Les meilleures performances ont été obtenues avec RoBERTa, c'est pourquoi nous avons fait le choix de ce modèle pour notre système hybride.

Le modèle RoBERTa-base est composé de 12 couches avec une taille cachée de 768 et un nombre de têtes d'attention de 12. Les couches intermédiaires (feedforward) ont une taille de 3072. Le modèle est construit sur un vocabulaire de 50 265 mots, y compris les jetons spéciaux pour le début et la fin de séquence (CLS et SEP). De plus, le modèle est sensible à la casse et utilise des jetons de masquage aléatoire (MLM) pour l'entraînement.

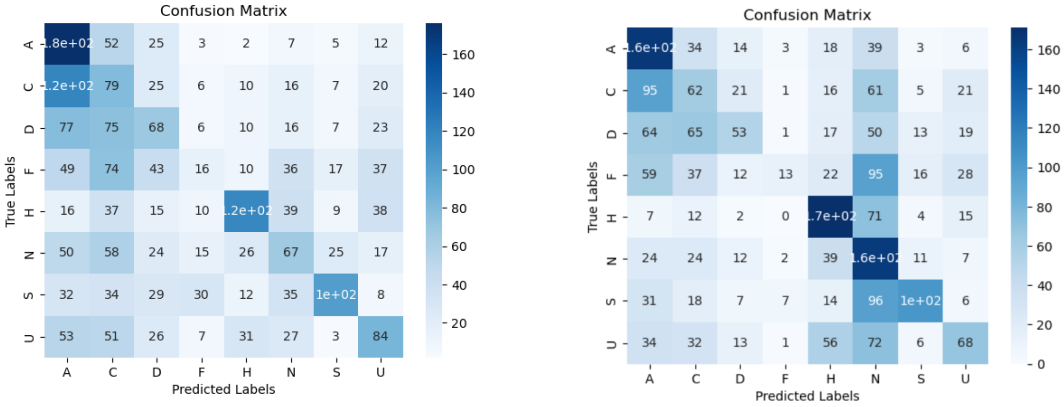
Nous avons adapté RoBERTa-base (125 millions de paramètres) aux données d'entraînement. Pour ceci, nous avons utilisé l'optimiseur Adam avec un taux d'apprentissage de $2e-5$ pendant 4 epochs. Les tailles de lots (batch sizes) ont été expérimentalement fixées à 64, et nous avons utilisé une fonction de perte d'entropie croisée pondérée (crossentropyweighted). Enfin, une couche linéaire a été ajoutée pour la tâche de classification des émotions. Comme pour le modèle acoustique, la sortie de cette couche est ensuite soumise à une fonction d'activation de type log-softmax, attribuant à chaque classe une probabilité.

3.4 Fusion des informations

Premièrement, comparons les résultats (sur notre jeu de développement équilibré) des systèmes acoustiques et textuels indépendamment pour établir plus tard du bénéfice de la fusion des deux informations pour notre tâche de classification.

	F1 Macro	**F1 Micro**	**Accuracy**
Système Audio	0.3073	0.3147	0.3147
Système Texte	0.2602	0.2836	0.2836

TABLE 2 – Résultats des deux systèmes sur notre jeu de développement équilibré



(a) Système audio

(b) Système texte

FIGURE 2 – Matrice de confusion des systèmes

Nous pouvons noter les différences entre les deux systèmes : l'émotion neutre est beaucoup mieux reconnue avec l'information textuelle tandis que le système audio semble plus performant en général.

Comme première expérimentation, nous proposons donc ici une méthode simple pour la fusion des résultats des deux modèles prosodique et sémantique. Nous départageons les deux modèles en faisant la moyenne de leurs résultats respectifs (probabilité pour chaque classe) afin de lisser les faiblesses ponctuelles de chacun (notamment lorsqu'un des deux systèmes est indécis).

4 Résultats et discussion

Le tableau 3 présente les résultats obtenus sur les ensembles de développement et de test.

À titre de comparaison, la campagne d'évaluation met à disposition une baseline (Goncalves *et al.*, 2024) présentée dans la figure 3. Leur méthode s'appuie uniquement sur un modèle audio constitué d'un encodeur WavLM (Chen *et al.*, 2022) pré-entraîné avec un taux d'apprentissage de $1e-5$, 20 epochs et une taille de lots de 32.

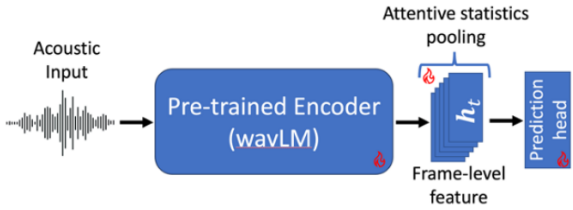


FIGURE 3 – Baseline de la campagne Odyssey 2024

Les meilleurs résultats obtenus par notre système se situent sur les émotions joie, tristesse et colère qui

Catégorie	Développement			Test		
	Précision	**Rappel**	**F1-score**	**Précision**	**Rappel**	**F1-score**
Neutral (N)	0.25	0.58	0.35			
Happiness (H)	0.48	0.61	0.54			
Sadness (S)	0.64	0.37	0.47			
Anger (A)	0.34	0.59	0.43			
Surprise (U)	0.40	0.24	0.30			
Contempt (C)	0.22	0.22	0.22			
Disgust (D)	0.40	0.19	0.25			
Fear (F)	0.46	0.05	0.08			
Accuracy	0,3537			**0.3511**		
Macro F1	0,3308			**0.3335**		
Baseline Accuracy	-			0,3272		
Baseline Macro F1	-			0,3113		

TABLE 3 – Résultats obtenus sur les données de développement et de test

ont des marqueurs sémantiques et prosodiques plus facilement identifiables et le neutre. La surprise, le mépris, le dégoût semblent elles difficilement identifiables par notre système. La peur, avec un F1-score de 0,08 est le point faible de notre système de classification.

Nous pouvons voir une petite baisse de précision lors de la prédiction sur l'ensemble de test (dont les labels ne sont encore pas publics) mais une amélioration majeure de la macro F1 permettant de dépasser la baseline de cette campagne.

5 Conclusion et perspectives

Dans cet article, nous avons présenté une première expérimentation pour la classification des émotions dans la parole. Le système hybride proposé repose sur un modèle acoustique et un modèle sémantique tous deux entraînés pour la tâche. Les résultats obtenus montrent qu'une fusion simple (moyenne des probabilités des deux modèles pour chaque classe) permet d'atteindre des résultats qui dépassent ceux d'un modèle acoustique seul. À court terme, nous envisageons de tester d'autres méthodes de fusion de l'audio et du texte, par exemple une concaténation des vecteurs de sortie des systèmes audio et texte, suivi d'une ou plusieurs couches permettant la classification. Cela pourrait nous permettre de mieux gérer les faiblesses de chaque système pour profiter au maximum de la dualité d'information que nous utilisons.

Remerciements

Ces travaux ont bénéficié d'un accès aux calculateur Jean Zay de l'IDRIS au travers des allocations de ressources AD011014274 et AD011013612R1 attribuées par GENCI.

Références

ATMAJA B. & AKAGI M. (2020). Dimensional speech emotion recognition from speech features and word embeddings by using multitask learning. *APSIPA Transactions on Signal and Information Processing*, **9**(1).

ATMAJA B., SHIRAI K. & AKAGI M. (2019). Speech emotion recognition using speech feature and word embedding. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2019*.

BAE Y. J., SHIM M. & LEE W. H. (2021). Schizophrenia detection using machine learning approach from social media content. *Sensors*, **21**(17), 5924.

BREDIN H. (2023). pyannote.audio 2.1 speaker diarization pipeline : principle, benchmark, and recipe. In *Proc. INTERSPEECH 2023*.

CHEN S., WANG C., CHEN Z., WU Y., LIU S., CHEN Z., LI J., KANDA N., YOSHIOKA T., XIAO X., WU J., ZHOU L., REN S., QIAN Y., QIAN Y., WU J., ZENG M., YU X. & WEI F. (2022). WavLM : Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing*, **16**(6), 1505–1518. arXiv :2110.13900 [cs, eess], DOI : 10.1109/JSTSP.2022.3188113.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*.

EKMAN P. & JOURNET N. (2002). *De l'universel au particulier*, In N. JOURNET, Éd., *La culture*, chapitre Le langage naturel des émotions, p. 29–37. Éditions Sciences Humaines : Auxerre.

EYBEN F., WÖLLMER M. & SCHULLER B. (2010). openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proc. ACM Multimedia (MM), ACM*.

GONCALVES L., SALMAN A., REDDY A., VELAZQUEZ L. M., THEBAUD T., GARCIA L. P., DEHAK N., SISMAN B. & BUSSO C. (2024). Odyssey 2024 - emotion recognition challenge. https://github.com/MSP-UTD/MSP-Podcast_Challenge.

HARRIGIAN K., AGUIRRE C. & DREDZE M. (2021). On the state of social media data for mental health research. In N. GOHARIAN, P. RESNIK, A. YATES, M. IRELAND, K. NIEDERHOFFER & R. RESNIK, Éds., *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology : Improving Access*, p. 15–24, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.clpsych-1.2.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). RoBERTa : A Robustly Optimized BERT Pretraining Approach. arXiv :1907.11692.

Loi n°2015-1776 article 51 (2015). Loi n° 2015-1776 du 28 décembre 2015 relative à l'adaptation de la société au vieillissement (article 51). JORF https://www.legifrance.gouv.fr/loda/article_lc/LEGIARTI000031706430.

LOTFIAN R. & BUSSO C. (2019). Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, **10, no. 4**.

MINISTÈRE DE L'ÉCONOMIE DES FINANCES ET DE LA RELANCE (2021). Guide ministériel du proche aidant. https://www.economie.gouv.fr/files/files/2021/guide_proche-aidant.pdf.

MOLINA A., HUANG X., HURTADO L.-F. & PLA F. (2023). ELiRF-UPV at eRisk 2023 : Early detection of pathological gambling using SVM. In *CLEF - CEUR-WS Working Notes*.

PARAPAR J., MARTIN-RODILLA, PATRICIA ANS LOSADA D. E. & CRESTANI F. (2023). Overview of eRisk at CLEF 2023 : Early Risk Prediction on the Internet. In *CLEF - CEUR-WS Working Notes*.

PLAQUET A. & BREDIN H. (2023). Powerset multi-class cross entropy loss for neural speaker diarization. In *Proc. INTERSPEECH 2023*.

PLUTCHIK R. (1980). A general psychoevolutionary theory of emotion. In *Theories of emotion*, p. 3–33. Elsevier.

RADFORD A., KIM J. W., XU T., BROCKMAN G., MCLEAVEY C. & SUTSKEVER I. (2022). Robust speech recognition via large-scale weak supervision. DOI : 10.48550/ARXIV.2212.04356.

RÍSSOLA E. A., LOSADA D. E. & CRESTANI F. (2021). A survey of computational methods for online mental state assessment on social media. *ACM Transactions on Computing for Healthcare*, **2**(2), 1–31.

SCHULLER B., STEIDL S. & BATLINER A. (2009). The INTERSPEECH 2009 emotion challenge. In *Proc. Interspeech 2009*, p. 312–315. DOI : 10.21437/Interspeech.2009-103.

SCHULLER B., STEIDL S., BATLINER A., BURKHARDT F., DEVILLERS L., MÜLLER C. & NARAYANAN S. S. (2010). The INTERSPEECH 2010 paralinguistic challenge. In *Proc. Interspeech 2010*, p. 2794–2797. DOI : 10.21437/Interspeech.2010-739.

TRIPATHI S., SAMARTH S. & HOMAYOON B. (2018). Multi-modal emotion recognition on IEMOCAP dataset using deep learning. arXiv preprint arXiv :1804.05788.

YENIGALLA P., KUMAR A., TRIPATHI S., SINGH C., KAR S. & VEPA J. (2018). Speech Emotion Recognition Using Spectrogram & Phoneme Embedding. In *Proc. Interspeech 2018*.

YOON S., BYUN S., DEY S. & JUNG K. (2019). Speech Emotion Recognition Using Multi-hop Attention Mechanism. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

YOUGOV I. (2019). Sondage Le Huffigton Post. Publié le 16/09/2019 dans le Huffington Post, https://d25d2506sfb94s.cloudfront.net/cumulus_uploads/document/20u9aq6czp/Copy%20of%20Results%20for%20YouGov%20%28Huf%20Post%20Psy%29%20159%2013.9.2019.pdf.

ZIRIKLY A., ATZIL-SLONIM D., LIAKATA M., BEDRICK S., DESMET B., IRELAND M., LEE A., MACAVANEY S., PURVER M., RESNIK R. & YATES A., Éds. (2022). *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, Seattle, USA. Association for Computational Linguistics.