MULTI-LINGUAL SPEECH EMOTION RECOGNITION

Younis Mouacha Youssef Ben Seddik

Université Laval

ABSTRACT

Cette recherche explore l'utilisation d'un modele pré-entraînés basé sur Wav2Vec 2.0 pour la reconnaissance des émotions dans la parole multilingue (SER). Alors que la communication humaine s'étend sur une multitude de langues, identifier avec précision les émotions dans des contextes linguistiques variés représente un défi majeur. Notre étude a impliqué l'adaptation de ces modèles avancés à des ensembles de données vocales multilingues, en se concentrant sur les spécificités nous permettant de comprendre les comportements de nos systèmes a généraliser sur des langues qui lui sont inconnues. Ces découvertes ouvrent la voie à des applications pratiques dans divers secteurs d'activité, soulignant le potentiel de l'intelligence artificielle pour une compréhension émotionnelle interculturelles.

Index Terms— Reconnaissance des émotions, Multilingue, Wav2Vec2.0, SER, Cross-Corpus, Multitâche

1. INTRODUCTION

La reconnaissance des émotions dans la parole (SER) dans un contexte multilingue est un défi majeur dans le domaine de l'intelligence artificielle et du traitement du langage naturel. Alors que les systèmes monolingues de SER ont prouvé leur efficacité, l'extension de ces méthodes à plusieurs langues pose de nouvelles questions complexes. Notre recherche s'inscrit dans cette perspective, avec une attention particulière portée sur les capacités des modèles d'apprentissage semi-supervisé (SSL) pré-entraînés pour le traitement de la parole. Nous explorons la faisabilité de créer un système SER multilingue qui puisse rivaliser avec les méthodes monolingues en termes de performances. Cette exploration implique une étude approfondie des modèles SSL pour une classification précise des émotions dans différentes langues (colère, joie, neutralité, et tristesse). Un aspect important de notre recherche est l'étude de l'interprétabilité des modèles pré-entraînés dans ce cadre multilingue, visant à comprendre comment ces modèles traitent et réagissent aux nuances linguistiques diverses. Pour ce faire, nous examinons la performance des modèles multilingues sur des langues non rencontrées durant leur entraînement, ce qui est essentiel pour évaluer leur adaptabilité et généralisabilité. Deplus, un aspect de notre recherche va porter sur la quantité de données labellisées nécessaires dans une langue inconnue pour que le modèle fonctionne de manière satisfaisante, éclairant ainsi sur l'efficacité des modèles SSL dans des scénarios de données limitées. Finalement, nous envisageons de mettre en œuvre notre méthode basée sur l'apprentissage multi-tâches avec gâting. Cette approche vise à surmonter les défis posés par les corpus déséquilibrés dans un contexte multilingue, en espérant améliorer davantage la robustesse et l'efficacité de nos systèmes SER.

2. DATASETS

Table 1. Corpus utilisés

Dataset Name	Langue	Samples	Length	Ref
IEMOCAP	en-US	5,531	12h	[1]
SUBESCO	bn-BN	4,000	4.47h	[2]
MESD	es-MX	574	6.85m	[3]
EMOUERJ	pt-BR	377	18.90m	[4]
EMODB	de-DE	339	15.89m	[5]
EMOVO	it-IT	336	16.76m	[6]
Oréau	fr-FR	254	12.25m	[7]

Dans le cadre de notre étude, nous mobilisons une sélection diversifiée d'ensembles de données pour enrichir notre analyse et renforcer la généralisabilité de nos résultats. Les datasets tels que IEMOCAP, SUBESCO, EMODB, MESD, EMOVO, Oréau, et EMOUERJ ont été sélectionnés pour couvrir un large éventail de langues et de nuances émotionnelles. Chacun de ces ensembles de données est spécifique à une langue et comprend des échantillons d'acteurs masculins et feminins exprimant différentes émotions, ce qui nous permet de capturer diverses nuances dans les expressions d'emotions. Cependant, il est important de considérer certaines limitations inhérentes. Principalement, le fait que comme l'ensemble des réactions émotionnelles dans ces datasets soient actées, cela pourrait introduire un biais dans le sens ou les émotions exprimées ne reflètent pas toujours la complexité et la subtilité des émotions naturelles vécues dans la vie réelle.

3. ETAT DE L'ART

Les recherches en détection de sentiments à travers la parole multilingue ont connu des avancées. Plusieurs approches ont été explorées pour relever ce défi complexe. L'état actuel de la recherche en reconnaissance des émotions dans la parole (SER) multilingue met en évidence des progrès à travers diverses approches innovantes. Sharma [8] présente un système SER basé sur le modèle pré-entraîné wav2vec 2.0, qui se distingue par son efficacité dans l'analyse multilingue et multitâches, surpassant les modèles basés sur PANN dans de multiples langues. Parallèlement, Zhang et al. [9] développent un cadre SER multilingue qui utilise un modèle de pré-entraînement pour extraire des représentations de parole de haut niveau, aboutissant à une précision améliorée dans une variété de langues. En complément, Wang et al. [10] introduisent un modèle multi-domaine avec un mécanisme multi-gating et une recherche d'architecture neuronale, offrant des améliorations significatives en termes de précision pour des langues moins représentées. Ces études [8][9][10] illustrent l'importance de l'adaptation des modèles SER aux spécificités linguistiques et soulignent l'efficacité des approches basées sur des modèles pré-entraînés et des structures de classification innovantes.

4. MÉTHODOLOGIE

4.1. Choix du modele

Dans notre recherche, nous avons opté pour deux architectures principales : Wav2Vec2 [11] et HuBERT [12]. Ces modèles, relevant de la famille des Transformers, se distinguent par leur capacité à traiter efficacement les signaux audio bruts, les transformant en représentations utiles grâce à une architecture d'apprentissage semi-supervisée. Ils excellent dans l'extraction de détails acoustiques et d'informations contextuelles, convertissant ainsi l'audio en vecteurs riches en informations. La particularité des Transformers réside dans leur mécanisme d'attention, qui permet au modèle de se concentrer sur différentes parties de la séquence d'entrée pour chaque élément généré dans la séquence de sortie. Ce processus facilite la compréhension des relations complexes et contextuelles au sein de notre signal audio. Ces modèles partagent une architecture de base similaire. Au cœur de cette architecture se trouve un encodeur de caractéristiques, constitué d'un réseau neuronal convolutif multicouche (CNN), qui transforme l'audio brut en représentations latentes de la parole. Ces représentations latentes sont par la suite acheminées vers un réseau d'encodeurs Transformer empilés, produisant des représentations contextualisées. Ces architectures avancées transforment des données audio complexes en vecteurs représentatifs précis, qui sont ensuite exploités par une tête de classification spécifiquement conçue pour diverses tâches.

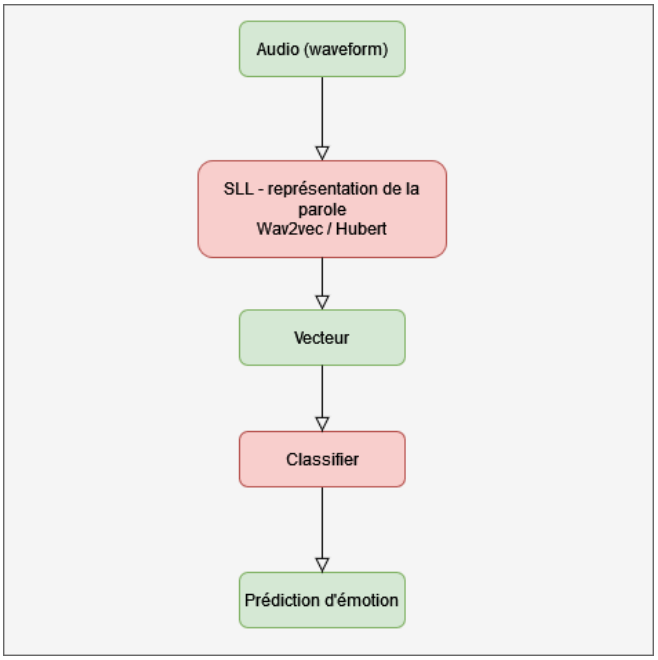

Fig. 1. Framework général

4.2. Politique d'entrainement

Un superset a été constitué en combinant les différents ensembles d'entraînement, de validation et de test issus de l'ensemble des corpus disponibles. Pour chaque dataset ainsi formé, une répartition de 70% pour l'entraînement, 15% pour la validation et 15% pour le test a été mise en place, garantissant une distribution homogène des données à travers toutes les phases de notre étude. Au cours du processus de préparation de ces datasets pour notre étude, nous avons dû effectuer certains prétraitements pour garantir la cohérence et la comparabilité entre les différentes sources. Ces prétraitements incluaient la normalisation des formats audio, l'unification des échelles de notation émotionnelle, la suppression de catégories émotionnelles spécifiques qui n'étaient pas uniformément représentées à travers tous les ensembles de données, et le resampling de l'ensemble du superset à 16000 kHz, un format nécessaire pour l'utilisation de nos modèles transformer. Pour notre étude de classification, nous avons opté pour la fonction de perte d'entropie croisée, privilégiée pour sa compétence à traiter de manière adéquate les scénarios impliquant de multiples classes.

$$H(y, \hat{y}) = -\sum_{i} y_i \log(\hat{y}_i) \quad (1)$$

Table 2. Hyperparamètres pour l'entrainement

Hyperparamètre	Valeur
Num train epochs	15
Batch size	32
Learning rate	2e-5
Lr scheduler	linear
Optimizer	ADAM

4.3. Experimentations

4.3.1. Finetuning et Comparaison de Wav2Vec2 et HuBERT pour la Reconnaissance Émotionnelle

L'expérimentation prévue implique le finetuning des modèles Wav2Vec2 [11] et HuBERT [12] pour la reconnaissance des émotions dans des données audio, en se concentrant sur un super ensemble de données qui comprend l'intégralité des corpus disponibles. Le finetuning sur ce super ensemble vise à affiner les capacités des deux modeles en les adaptant aux nuances et variabilités présentes dans les divers ensembles de données. Après cette phase d'ajustement, une évaluation comparative des performances des deux modèles sera réalisée. Le modèle qui montrera les meilleures performances sur le superset sera sélectionné pour des expérimentations futures. Cette approche garantit que le modèle choisi est le plus efficace pour traiter une gamme étendue et diverse de données audio émotionnelles.

4.3.2. Dataset Dropout

La meilleure baseline ayant été établie lors des tests précédents, le modèle sélectionné fera l'objet d'une évaluation plus poussée. L'approche adoptée consiste à entraîner l'architecture sur l'ensemble des corpus disponibles, à l'exception d'un seul, choisi alternativement pour chaque cycle de test. Cette technique vise à mesurer la capacité du modèle à s'adapter et à maintenir des performances élevées sur des ensembles de données entièrement nouveaux. Le but est d'évaluer son aptitude à s'accommoder à différentes langues ou contextes, donnant ainsi une mesure concrète de sa polyvalence et de sa fiabilité dans diverses situations.

4.3.3. Evolution de la performance dans un contexte Multilingue

La phase d'expérimentation est conçue pour tester le système dans un contexte où les données sont limitées. Utilisant le corpus SUBESCO, l'expérience consiste à ajouter graduellement des échantillons de ce corpus au jeu d'entraînement, afin d'évaluer l'impact sur les performances du jeu de test correspondant. Cette approche permettra de déterminer le nombre minimum d'échantillons requis pour obtenir des performances acceptables.

4.3.4. Reconnaissance Multitâche des Émotions

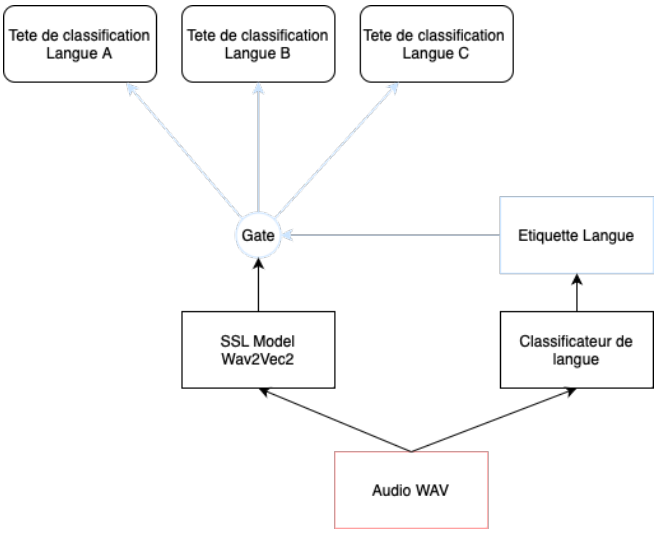

Fig. 2. Architecture Multitâche

Pour tenter de surmonter le défi du déséquilibre des corpus dans la reconnaissance des émotions à partir de données audio, une méthode de multitask learning s'appuyant sur un layer commun basé sur Wav2Vec2 est adoptée. Chaque langue est associée à un label spécifique, qui active la tête de classification correspondante pour l'analyse des émotions. Dans la phase de test, l'identification automatique de la langue de l'audio entrant devient essentielle. À cet effet, un modèle de classification de langue est intégré. Cette étape est fondamentale dans des contextes réels où la langue de l'audio n'est pas connue à l'avance. Le classificateur de langue détermine l'étiquette de la langue, activant ainsi la tête de classification appropriée. Cette approche renforce la praticité et l'applicabilité du modèle dans des situations réelles, permettant une reconnaissance des émotions indépendamment de la langue utilisée.

5. RÉSULTATS

Table 3. Finetuning et Comparaison de Wav2Vec2 et HuBERT pour la Reconnaissance Émotionnelle

Model	Accuracy	Precision	Recall	F1-Score
wav2vec2.0	0.747	0.747	0.748	0.747
huBERT	0.728	0.729	0.727	0.725

6. DISCUSSIONS

L'utilisation de wav2vec 2.0 se révèle plus efficace que celle de Hubert, démontrant des performances améliorées (Table 3).

Table 4. Dataset Dropout avec wav2vec2.0

Dropout	Accuracy	Precision	Recall	F1-score
EMODB	0.795	0.805	0.804	0.791
emoUERJ	0.602	0.597	0.598	0.595
SUBESCO	0.508	0.506	0.507	0.502

Table 5. Evolution de la performance dans un contexte Multilingue avec SUBESCO

Samples	Accuracy	Precision	Recall	F1-score
0	0.508	0.506	0.507	0.502
100	0.5516	0.546	0.553	0.542
500	0.631	0.636	0.632	0.624
1000	0.655	0.652	0.656	0.646
2000	0.703	0.7038	0.705	0.697
3600	0.746	0.769	0.747	0.744

Les résultats, présentés dans la Table 4, mettent en évidence des tendances dans l'expression des émotions à travers différentes langues. L'analyse du dataset EMODB en allemand révèle des similitudes avec d'autres langues européennes, particulièrement l'anglais, en raison de leur héritage germanique commun et de leurs influences historiques et géographiques proches. Cependant, le dataset emoUERJ en portugais affiche des performances inférieures, possiblement due à la représentation restreinte des langues latines dans le superset. Cette tendance est encore observable avec le dataset subesco en bengali, où les résultats sont moins probants, illustrant les écarts significatifs entre les structures linguistiques européennes et asiatiques, et soulignant l'impact de leurs origines géographiques et linguistiques distinctes.

Comme indiqué dans la Table 5, l'augmentation de la taille du dataset entraîne généralement une amélioration des scores, avec une moyenne d'environ 0.05. Toutefois, une exception est observée lors du passage de 500 à 1000 échantillons, où l'augmentation du score est de seulement 0.02. Cette tendance suggère l'intérêt de poursuivre les recherches avec un plus grand nombre d'échantillons de subesco, afin de déterminer à quel point l'accroissement des échantillons cesse d'influencer significativement les scores.

La Table 6 présente les performances de diverses configurations de modèles pour la reconnaissance multitâche des émotions. Le modèle wav2vec2.0, utilisé comme référence, a enregistré une précision de 0.747. Lors de l'intégration du composant MTGate et des étiquettes de vérité terrain au modèle wav2vec2.0, la précision observée diminue à 0.670. En ce qui concerne la combinaison de wav2vec2.0, MTGate et les étiquettes issues du classificateur de langue, il est important de noter que nous n'avons pas pu réaliser cette configuration en raison de contraintes liées à l'accès aux ressources de calcul nécessaires pour entraîner le classificateur de langue.

Table 6. Reconnaissance Multitâche des Émotions

Model	Acc
wav2vec2.0 (Baseline)	0.747
wav2vec2.0 + MTGate + Ground Truth labels	0.670
wav2vec2.0 + MTGate + Classifier	N/A

Table 7. Comparaison des performances avec SUPERB Benchmark sur IEMOCAP

Model	Acc
wav2vec2.0	0.543
SUPERB-wav2vec2.0 [13]	0.656
SUPERB-huBERT [13]	0.676

Note : Le benchmark SUPERB a été entraîné et testé uniquement sur IEMOCAP, tandis que nous avons été entraînés sur l'ensemble des corpus et testés sur une partie d'IEMOCAP.

7. CONCLUSION

Cette étude a abordé l'interprétabilité des modèles d'apprentissage semi-supervisé dans un contexte de traitement de données multilingues limitées. Elle a examiné comment ces modèles gèrent la diversité linguistique en présence de données restreintes, visant à comprendre leur capacité à maintenir une précision acceptable. En parallèle, l'architecture d'apprentissage multitâche avec techniques de gating a été explorée, visant à renforcer la flexibilité et l'adaptabilité des modèles dans un contexte multilingue complexe avec des corpus débalancés.

8. TRAVAUX FUTURS

Pour explorer de nouvelles voies d'amélioration de la classification des émotions dans un contexte multilingue, plusieurs approches pourraient être envisagées. L'une des stratégies consiste à développer des têtes de classification spécifiques pour chaque famille linguistique, en tirant parti des similarités observées entre les langues au sein de ces groupes. Cette méthode pourrait être appliquée pour regrouper et traiter différemment les langues asiatiques, latines, germaniques, etc., exploitant ainsi les caractéristiques communes et les nuances propres à chaque groupe. Une autre piste intéressante est la réduction ciblée de la taille des corpus majoritaires. Nos études ont montré que des ensembles de données moins volumineux peuvent souvent produire des résultats aussi précis que ceux obtenus avec des ensembles plus importants. En complément, l'adoption de modèles axés sur les données, comme le démontre Tang et al. [14], propose un modèle DNN basé sur l'apprentissage par transfert et wav2vec 2.0, offrant des avancées notables en SER inter-langues. Ce modèle, enrichi d'une couche de normalisation Deep-WCCN, excelle tant dans les contextes intra-langues qu'inter-langues, surpas-

sant les approches acoustiques traditionnelles et établissant un nouveau standard en SER.

9. REFERENCES

[1] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, et al., "Iemocap: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.

[2] Sadia Sultana, M. Shahidur Rahman, M. Reza Selim, and M. Zafar Iqbal, "Sust bangla emotional speech corpus (subesco): An audio-only emotional speech corpus for bangla," *PLOS ONE*, vol. 16, no. 4, pp. 1–27, 04 2021.

[3] Mathilde Marie Duville, Luz María Alonso-Valerdi, and David I Ibarra-Zarate, "Mexican emotional speech database (mesd)," 2022.

[4] R. G. Bastos Germano, M. Pompeu Tcheou, F. da Rocha Henriques, and S. Pinto Gomes Junior, "emouerj: an emotional speech database in portuguese," 2021.

[5] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss, "A database of German emotional speech.," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Lisbon, Portugal, 2005, vol. 5, pp. 1517–1520, ISCA.

[6] Giovanni Costantini, Iacopo Iaderola, Andrea Paoloni, and Massimiliano Todisco, "EMOVO corpus: an Italian emotional speech database," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, Eds., Reykjavik, Iceland, May 2014, pp. 3501–3504, European Language Resources Association (ELRA).

[7] Leila Kerkeni, Catherine Cleder, Youssef Serrestou, and Kosai Raoof, "French emotional speech database - oréau," 2020.

[8] Mayank Sharma, "Multi-lingual multi-task speech emotion recognition using wav2vec 2.0," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6907–6911.

[9] Zhaohang Zhang, Xiaohui Zhang, Min Guo, Wei-Qiang Zhang, Ke Li, and Yukai Huang, "A multilingual framework based on pre-training model for speech emotion recognition," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2021, pp. 750–755.

[10] Zihan Wang, Qi Meng, HaiFeng Lan, XinRui Zhang, KeHao Guo, and Akshat Gupta, "Multilingual speech emotion recognition with multi-gating mechanism and neural architecture search," 2022.

[11] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020.

[12] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," 2021.

[13] Shu wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee, "Superb: Speech processing universal performance benchmark," 2021.

[14] Duowei Tang, Peter Kuppens, Luc Geurts, and Toon van Waterschoot, "End-to-end transfer learning for speaker-independent cross-language speech emotion recognition," 2023.