Modélisation auto-supervisée de la parole affective spontanée

Ziyi Tong

Modélisation Auto-Supervisée de la Parole Affective Spontanée

Ziyi
TONG

Sous la direction de Fabien RINGEVAL

Laboratoire : Laboratoire d'informatique de Grenoble

UFR LLASIC
Département I3L

Mémoire de master 2 mention Sciences du Langage – 30 crédits

Parcours : Industrie de la langue (IDL)

Année universitaire 2020-2021

Modélisation Auto-supervisée de la Parole Affective Spontanée

Ziyi
TONG

Sous la direction de Fabien RINGEVAL

Laboratoire : Laboratoire d'informatique de Grenoble

UFR LLASIC
Département I3L

Mémoire de master 2 mention Sciences du Langage – 30 crédits

Parcours : Industrie de la langue (IDL)

Année universitaire 2020-2021

Remerciements

Je tiens d’abord à remercier mon directeur de mémoire, monsieur Fabien Ringeval, pour sa patience, sa disponibilité, ses guides éminents et ses conseils judicieux, qui ont contribué à m’aider à réussir ce stage et à alimenter ma réflexion.

Je remercie également Sina Alisamir pour les aides successives qu’il m’a apportées tout au long de ce stage, ainsi que pour sa patience et toute sa bienveillance.

Je remercie mon tuteur universitaire Solange Rossato pour avoir pris le temps de me suivre pendant mon stage, pour avoir accepté de répondre à mes questions durant ce stage et pour m’avoir conseillée.

DÉCLARATION ANTI-PLAGIAT

1. Ce travail est le fruit d'un travail personnel et constitue un document original.
2. Je sais que prétendre être l'auteur d'un travail écrit par une autre personne est une pratique sévèrement sanctionnée par la loi.
3. Personne d'autre que moi n'a le droit de faire valoir ce travail, en totalité ou en partie, comme le sien.
4. Les propos repris mot à mot à d'autres auteurs figurent entre guillemets (citations).
5. Les écrits sur lesquels je m'appuie dans ce mémoire sont systématiquement référencés selon un système de renvoi bibliographique clair et précis.

PRENOM : ……

NOM : ……

DATE : ……

Sommaire

Remerciements 3
Sommaire 6
Introduction 8
Partie 1 - Etat de l'art 10
Chapitre 1. Qu'est-ce que la modélisation auto-supervisée 11
1. Définition de la Modélisation 11
2. Approches générales de la modélisation de la parole 11
3. Modélisation auto-supervisée 12
Chapitre 2. Reconnaissance de la parole affective - deux challenges : AVEC et ComParE 14
1. Présentation de AVEC 14
2. Présentation de ComParE 14
3. Analyse : l'apprentissage non-supervisé et la fusion avec les informations linguistiques 15
4. Analyse : Tendance vers l'apprentissage non-supervisé 16
Chapitre 3. Apprentissage auto-supervisé pour les représentations de la parole 19
1. Wav2vec 19
2. Wav2vec 2.0 19
3. Application pour la reconnaissance des émotions 21
Chapitre 4. Apprentissage auto-supervisé pour les représentations textuelles 23
5. Modèles pré-entraînés 23
Partie 2 - Méthode 26
Chapitre 5. Extraction des représentations 27
1. Par l'apprentissage auto-supervisé 29
2. Par le lexique relatif aux émotions 33
3. Par l'approche Tf-idf 34
Chapitre 6. Alignement et Rééchantillonnage 35
1. Alignement 35
2. Rééchantillonnage 35
Partie 3 - Expérimentation 37
Chapitre 7. Introduction 38
1. Tâche 38
2. Vue d'ensemble 38
Chapitre 8. Corpus 40
1. Présentation générale 40
2. Répartition 40
3. Transcriptions 40
4. Annotation Gold standard 40

Chapitre 9. Modèles 41
1. Linear-Tanh 41
2. GRU 42
3. Optimiseur : Adam 44
Chapitre 10. Métrique d'évaluation 45
1. CCC 45
2. RMSE 45
Chapitre 11. Conclusion de l'expérimentation 46
Partie 4 - Résultats et Analyses 48
Chapitre 12. Résultats et Analyses 49
1. Résultats en CCC 49
3. Analyses 51
Chapitre 13. Comparaison et Analyses 54
1. Acoustique et textuelle 54
2. Auto-supervisée et autres approches 55
3. CCC et RMSE 56
4. Modèle Multilingue et Modèle en français 56
Partie 5 - Conclusion et Perspectives 58
Chapitre 14. Conclusion 59
Chapitre 15. Discussion 60
Chapitre 16. Perspectives 61
1. Mise en commun (Pooling) 61
2. Co-attention Fusion 61
3. Fusion superficielle pour des modèles de type BERT 62
Bibliographies 63
Sitographie 66
Glossaire 67
Sigles et abréviations utilisés 68
Table des Figures 69
Table des Tableaux 70
Table des annexes 71
Table des matières ..... 错误!未定义书签。

Introduction

La reconnaissance des émotions de la parole demeure un sujet important en TAL en raison de son rôle spécial et indispensable qui consiste à rendre la communication homme-machine plus naturelle et conviviale. Récemment, plusieurs approches d'apprentissage auto-supervisé ont été exploré pour le traitement automatique de la parole. Les recherches utilisant des représentations auto-supervisées dans les tâches de reconnaissance vocale ont réussi à améliorer les performances, même avec des ressources limitées (Baevski et al., 2020; Kawakami et al., 2020). Ces succès suggèrent que la modélisation auto-supervisée (MA) mérite d'être transféré dans des domaines similaires, tels que la reconnaissance des émotions de la parole (REP). En plus, vu que la plupart des études concernant la MA sont réalisées pour l'anglais, il nous semble donc intéressant d'étudier cette approche pour la langue française.

Dans cet esprit, nous nous intéressons à interroger l'utilisation de la MA dans la tâche REP en temps continu. Nous avons fait trois hypothèses : 1. Les représentations auto-supervisées de la parole (acoustiques) contribueraient à faire des bonnes prédictions dans la tâche REP en temps continu ; 2. Les représentations auto-supervisées des transcriptions de la parole (textuelles) pourraient faire des bonnes prédictions dans la tâche REP en temps continu ; 3. La combinaison des représentations acoustiques et des représentations textuelles pourrait améliorer la performance du système de REP.

La première hypothèse a été examiné par (Evain et al., 2021) dans la section « reconnaissance automatique des émotions ». Nous nous concentrons, dans ce mémoire, sur la deuxième hypothèse et discutons les perspectives et les approches intéressantes vers la troisième hypothèse.

Le problématique de ce mémoire est l'évaluation des diverses approches de modélisation auto-supervisée pour les transcriptions de la parole, sur la tâche de prédiction des émotions de la parole en temps continu.

Utiliser les transcriptions pour faire la prédiction des émotions de la parole en temps continu est une nouvelle approche, ce mémoire est à la fois une extension de la recherche d'apprentissage auto-supervisé des représentations acoustiques et une étape intermédiaire ouvrant le chemin d'exploration de la combinaison des représentations auto-supervisée acoustiques et textuelles dans le domaine REP.

Ce mémoire se compose de quatre parties, en premier lieu, l'état de l'art met l'accent sur la définition de modélisation auto-supervisée ainsi que les méthodes existantes d'apprentissage auto-supervisé des représentations. Par la suite, nous présentons notre méthode par expliquer notre accès à des représentations et l'alignement des transcriptions avec la parole. En troisième lieu, nous présentons en détails notre conception de l'expérimentation. En dernier lieu, nous effectuons une série d'analyses sur les résultats de l'expérience et discutons quelques perspectives possibles pour la combinaison de MA de la parole et de MA des transcriptions de la parole.

Partie 1

-

Etat de l’art

Chapitre 1. Qu’est-ce que la modélisation auto-supervisée

1. Définition de la Modélisation

Le terme « modélisation » s’applique à l’action de représenter d’un objet en taille réduite, à base de certaines règles, dans l’intention de comprendre le fonctionnement de cet objet. En TAL, la modélisation se rapporte au processus d’extraction des caractéristiques à partir des données, autrement dit, la modélisation décrit les scénarios dans lesquelles les représentations sont extraites, en suivant certains protocoles. Ces représentations incluant des informations de haut niveau sont ensuite utilisées dans des tâches en aval.

2. Approches générales de la modélisation de la parole

Les signaux vocaux sont généralement représentés de deux façons, soit par des techniques basées sur des connaissances d'experts, soit par des techniques de l'apprentissage automatique. Selon l'utilisation des données annotées ou brutes, une distinction est faite entre l'apprentissage supervisé et non supervisé.

2.1 Apprentissage supervisé et non supervisé

L'apprentissage supervisé basée sur des connaissances d'experts est souvent plus performante dans la tâche spécifique pour laquelle qu'il est entraîné, pourtant cette approche nécessite une grande quantité de données annotées et elle est donc très coûteuse. Face à ce problème, vu que d'innombrables quantités de données brutes sont disponibles gratuitement sur Internet, les approches d'apprentissage non-supervisés ont été proposé pour profiter de ces données non étiquetées. L'apprentissage non-supervisé consiste à apprendre un modèle de données sous-jacent à partir d'une grande quantité de données non étiquetées. L’apprentissage non-supervisé a généralement une structure auto-encodeur, ce qui consiste à apprendre des représentations par la reconstruction des données d'entrée.

2.2 Différences entre l’auto-encodeur et l’apprentissage auto-supervisé

L’apprentissage auto-supervisé est similaire à l’apprentissage non-supervisé à l’égard de l’utilisation des données brutes non-étiquetées. Cependant, ces deux approches sont très éloignées à propos des normes de la préservation de l’information et des modèles d’entraînement.

Concrètement, l'apprentissage non-supervisé cherche à reconstruire les données d'entrée et essaie d'enlever des informations communes des différents données, tandis que

l’apprentissage auto-supervisé vise à prédire les tokens masqués ou la phrase suivante et essaie d'apprendre le modèle commun à partir de différentes tâches. L'apprentissage auto-supervisé permet de préserver les informations communes qui distinguent les modèles, en parallèle, d'enlever des données communes apparaissant dans tous les cadres et des données contenant peu d’information intéressante en raison de sa large généralité.

Comparer avec les méthodes de l'apprentissage non supervisé, l'apprentissage auto-supervisé a montré sa capacité de produire le meilleur résultat dans plusieurs tâches à propos de la parole. L'avantage de l'apprentissage auto-supervisé est que, même s'il existe très peu de données d’échantillon, l’optimisation du système est possible à travers l'entraînement et l'affinement du modèle de l'apprentissage auto-supervisé. Par exemple, le modèle wav2vec (Schneider et al., 2019) utilisant 100 fois moins de données étiquetées pour affiner le modèle de transcription de la parole, alors qu'il arrive à produire de meilleurs résultats que le meilleure système d’apprentissage semi-supervisé dans une tâche de reconnaissance de la parole.

3. Modélisation auto-supervisée

Le processus l’appelle « la modélisation auto-supervisée » est d’extraire des représentations robustes et génériques à travers les modèles d’apprentissage auto-supervisée pré-entraîné.

3.1. Définition de l’apprentissage auto-supervisé

Selon (Chen et al., 2020), l’apprentissage auto-supervisé apprend des représentations en utilisant des fonctions objectives similaires à celles utilisées pour l'apprentissage supervisé, mais entraîne les réseaux neuronaux à effectuer des tâches prétextes (tâches auxiliaires) où les entrées et les étiquettes sont dérivées d'un ensemble de données non étiquetées.

3.2 Applications des représentations de l’apprentissage auto-supervisé

Dans le domaine du traitement des images, les études de (Bachman et al., 2019; Chen et al., 2020) témoignent de la supériorité des représentations extraites de l’apprentissage auto-supervisé à travers l’augmentation importante de l’exactitude sur ImageNet.

Dans le domaine du traitement automatique de la langue, BERT, FlauBERT, XLM etc, les modèles auto-supervisées pré-entrainés ont connu beaucoup de succès dans les tâches

telles que question-réponse et l'inférence du langage(Devlin et al., 2019 ; H. Le et al., 2020 ; Lample & Conneau, 2019).

Dans le domaine du traitement automatique de la parole, plusieurs études (Chung & Glass, 2018; Hjelm et al., 2019; Oord et al., 2019; Schneider et al., 2019) montrent que l'apprentissage auto-supervisé des représentations est capable d'améliorer la performance du système dans les tâches en aval, telle que la reconnaissance de la parole.

Chapitre 2. Reconnaissance de la parole affective - deux challenges : AVEC et ComParE

La reconnaissance automatique des émotions vise à détecter l'état émotionnel des humaines. On peut détecter à partir de la parole collectées grâce à des microphones. Les applications dans les domaines de la santé, de l'éducation, de l'art, demandent de plus en plus les systèmes REP performants, pour pouvoir à améliorer la communication homme-machine.

Pour connaître l'état de l'art de la reconnaissance de la parole affective, nous avons parcouru les deux challenges les plus connus dans le domaine de l'informatique affective et nous avons sélectionné les sujets relatifs à la détection des émotions présentes dans la parole, entre 2016 et 2020 et ensuite organisé des données recueillies. Cette section est dédiée aux analyses des approches utilisées dans le domaine l'informatique affective.

1. Présentation de AVEC

AVEC est l'abréviation de « The Audio/Visual Emotion Challenge », il s'agit un colloque qui est tenu annuellement avec différents sujets dans le domaine du traitement automatique des informations multimodales et le domaine de l'analyse des émotions. Les techniques en informatique linguistique et en apprentissage automatique sont ainsi étudiées. AVEC vise à fournir un ensemble de tests de référence communs pour le traitement de l'information multimodale et de réunir les communautés de la reconnaissance des affects audio, visuels et audio-visuels, afin de comparer les mérites relatifs des approches de l'analyse automatique de la santé et des émotions dans des conditions bien définies.

Une autre motivation repose sur la nécessité de faire progresser les systèmes de reconnaissance de la santé et des émotions pour qu'ils soient capables de traiter des comportements entièrement naturels, dans de grands volumes de paroles non segmentées, non typiques et non présélectionnées, car c'est exactement le type de données auxquelles les interfaces de communication multimédia et homme-machine/homme-robot doivent faire face dans le monde réel.

2. Présentation de ComParE

ComParE est une abréviation de « The Interspeech Computational Paralinguistics ChallengE ». ComParE est un challenge ouvert dans le domaine de la paralinguistique

informatique qui concerne les états et les traits des locuteurs manifestés dans les propriétés de leur signal vocal. Ce challenge est aussi tenu annuellement. Chaque année, de nouvelles tâches sont introduites en raison de la large quantité et de la diversité des phénomènes paralinguistique qui sont hautement pertinents mais non couverts. S'adressant aux communautés du traitement automatique la parole et du traitement du signal, du traitement du langage naturel, de l'intelligence artificielle, de l'apprentissage automatique, de l'informatique affective et comportementale, de l'interaction homme-machine/robot, de la santé mobile, de la psychologie et de la médecine, ainsi qu'à tout autre participant, ComParE est un évènement important dans le domaine de la reconnaissance des émotions de la parole.

3. Analyse : l'apprentissage non-supervisé et la fusion avec les informations linguistiques

Nous avons sélectionné les tâches relatives aux émotions entre 2016 et 2020 dans AVEC et ComParE. Nous avons ensuite organisé un tableau contenant les meilleurs résultats et les résultats de référence de chaque approche pour chaque tâche. Le tableau complet est dans l'annexe 1.

Comme l'approche apprentissage non-supervisé et l'approche fusion avec les informations linguistiques sont liées à notre sujet de recherche, nous voulons mettre l'accent sur ces deux sections. Tableau 1 présente le panorama de l'apprentissage non-supervisé et l'approche de la fusion avec les informations linguistiques des deux challenges dans la période 2016 - 2020. Ce tableau fait une partie de l'annexe 1 mais nous avons enlevé des lignes vides car elles signifient qu'il n'y a pas de participants utilisant ces deux approches pour ces tâches.

Nous observons que, à partir 2018, l'apprentissage non-supervisé est en plein essor, de même, l'approche utilisant la fusion avec les informations linguistiques développe rapidement depuis 2019, ce qui indique que notre idée d'utiliser l'apprentissage auto-supervisé des représentations acoustiques et de combiner les représentations textuelles pour augmenter la performance du système de REP sont très prometteuses.

Tableau 1 Les participants ayant obtenu les meilleurs résultats en utilisant l'approche non-supervisée ou la fusion avec les informations linguistiques dans les tâches affectives d'AVEC et de ComParE de 2016 à 2020

Challenge	Tâche	Evaluation	Référence	Non - Supervisé	Fusion avec les informations linguistiques
ComParE 2020	Elderly Emotion (Arousal)	UAR[1]	Baseline	44.3	44.0
			(Soğancıoğlu et al., 2020)		63.7
	Elderly Emotion (Valence)	UAR	Baseline	33.8	49.0
			(Soğancıoğlu et al., 2020)		57.5
AVEC 2019	Depression Detection with AI	CCC[2]	Baseline		
			(Rodrigues Makiuchi et al., 2019)		.403
ComParE 2019	Continuous Sleepiness	PC[3]	Baseline	.325	
	Baby Sound	UAR	Baseline	48.1	
			(Yeh et al., 2019)	62.4	
ComParE 2018	Atypical Affect	UAR	Baseline	35.6	
	Self-Assessed Affect	UAR	Baseline	57.3	
			(Montacié & Caraty, 2018)		68.4
	(Infant) Crying	UAR	Baseline	71.1	

4. Analyse : Tendance vers l'apprentissage non-supervisé

Différentes approches sont utilisées dans AVEC et ComParE, afin de connaître la tendance des méthodes préférées dans le domaine paralinguistique informatique et le domaine traitement automatique de la parole, nous avons faire une statistique sur les pourcentages de chaque approche dans les deux challenges entre 2016 et 2020, le résultat est présenté dans Figure 1.

A travers le Figure 1, nous constatons d'abord une baisse progressive du pourcentage de l'approche « hand-crafted ». L'approche hand-crafted est une approche traditionnelle utilisant LLDs (descripteurs de bas niveaux) pour représenter la parole. Le terme "hand-crafted" implique que les caractéristiques sont obtenues par des équations mathématiques

[1] Unweighted Average Recall
[2] Concordance Correlation Coefficent
[3] Pearson correlation coefficient

conçu par les humains. La baisse de pourcentage révèle le changement des méthodes de recherches : les méthodes dépendant des connaissances et du travail des experts sont progressivement abandonnées en raison de ses prix coûteux.

Enfin, nous remarquons une hausse significative au pourcentage des approches utilisant l'apprentissage automatique, surtout par l'apprentissage non-supervisé. A partir de 2018, l'apprentissage non-supervisé est en plein essor. En outre, l'approche fusion avec les informations linguistiques obtient de plus en plus d'attention, nous pouvons observer une croissance du pourcentage de cette approche depuis 2018.

En résume, une tendance vers l'approche d'apprentissage non-supervisé est clairement observée à partir des statistiques collectées dans de variées tâches de deux challenges AVEC et ComParE entre 2016 et 2020.

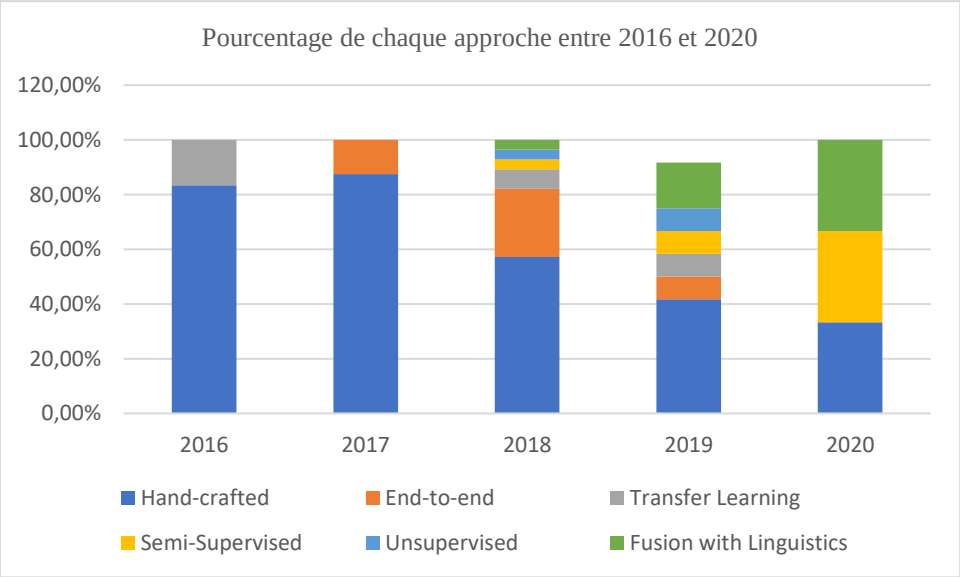


Figure 1 Pourcentage de chaque approche dans les deux challenges AVEC et ComParE entre 2016 et 2020

	Hand-crafted	End-to-end	Transfer Learning	Semi-Supervised	Unsupervised	Fusion with linguistics
2016	5		1			
2017	7	1				
2018	16	7	2	1	1	1
2019	5	1	1	1	1	2
2020	1				1	1

Tableau 2 Nombre d'études concernant l'informatique affective de chaque approche dans challenge AVEC et ComParE entre 2016 et 2020

Chapitre 3. Apprentissage auto-supervisé pour les représentations de la parole

Dans cette section, nous présentons l'état de l'art sur ce sujet : l'extraction de représentation à partir de la parole à travers les modèles pré-entraînés de l'apprentissage auto-supervisé. Wav2vec est le modèle auto-supervisé le plus connu dans le domaine du traitement de la parole, nous mettons donc l'emphase sur ce modèle et ses applications pour la tâche de reconnaissance des émotions.

1. Wav2vec

Wav2vec (Schneider et al., 2019) est un modèle pré-entrainé sur une grande quantité de données audio non-étiquetées. Il apprend des représentations à partir des audios bruts.

Le modèle est composé de deux réseaux de neurones convolutionnels simples : un réseau de l'encodeur et un réseau du contexte. Le réseau de l'encodeur $f : X \rightarrow Z$ prend des audios bruts $x_i \in X$ comme l'entrée et prend des représentations de basse fréquence ($z_1, z_2, \ldots, z_T$) comme la sortie. Ce réseau de l'encodeur encode environ 30 ms de 16 kHz toutes les 10 ms. Le réseau du contexte $g : Z \rightarrow C$ transforme les représentations de basse fréquence en représentations contextuelles de plus haut niveau $c_i = g(z_i \ldots z_{i-v})$ pour un champ réceptive *v*. Après être passé par les deux réseaux, le champ réceptif total est de 210 ms pour la version de base et 810 ms pour la version large.

Wav2vec est entrainée sur environs 1000 heures d'audio bruts en anglais avec la tâche de classification binaire contrastive du bruit où l'objectif est de distinguer les vrais échantillons futurs des distractions.

2. Wav2vec 2.0

Wav2vec 2.0 est le successeur de Wav2vec. Avec seulement 10 minutes de parole transcrite et 53K heures de parole non étiquetée, ce nouvel modèle pré-entrainé auto-supervisé permet aux systèmes de reconnaissance vocale d'atteindre un taux d'erreur sur les mots (WER) de 8,6 pour cent sur la parole bruyante et de 5,2 pour cent sur la parole propre, sur le benchemark standard de LibriSpeech.

Excellente performance de Wav2vec 2.0 en cas d'absence des données étiquetées ouvre la voie à la construction des systèmes de reconnaissance de la parole pour beaucoup

de langues qui sont contraintes à construire leurs modèles de reconnaissance de la parole à cause du nombre faible de données.

Le modèle Wav2vec2 est composé de trois parties : un encodeur convolutionnel multi-couches, un réseau du context (Transfomer) et un module de quantification. L'encodeur $f : X \rightarrow Z$ prend l'audio brut X comme l'entrée et prend des représentations latentes de la parole $z_1, \ldots, z_T$ pour T pas-de-temps comme la sortie. Ensuite, ces représentations sont discrétisées q_t avec un module de quantification $Z \rightarrow Q$ pour représenter les cibles dans l'objectif auto-supervisé. Enfin, ces représentations entrent dans Transformer $g : Z \rightarrow C$ pour construire des représentations $c_1, \ldots, c_T$ qui capturent les informations de la séquence entière. Wav2vec2 construit des représentations contextuelles de la parole en temps continu et son mécanisme auto-attention permet de capture des dépendances entre des représentations latentes de la séquence entière.

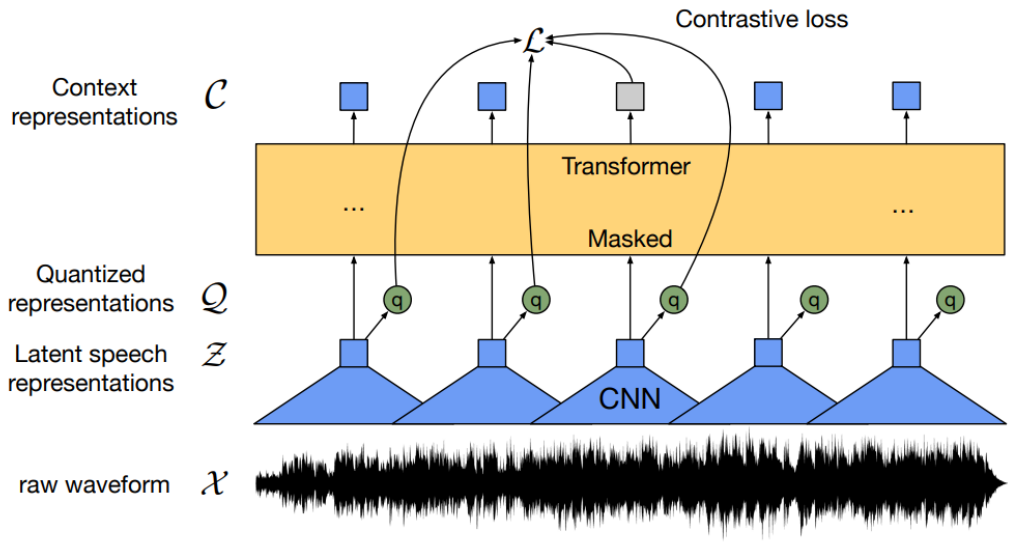


Figure 2 Illustration de l'architecture de Wav2vec 2.0

2.1 Wav2vec2 pour le français

LeBenchemark (Evain et al., 2021) publie quatre modèles wav2vec2 pré-entraînés sur différents données françaises contenant de la parole spontanée, lue et diffusée. *wav2vec2-FR-M-Large* et *wav2vec2-FR-S-Large* sont entraînés sur environ 3k heures de la parole en

français et *wav2vec2-FR-M-Base* et *wav2vec2-FR-S-Base* sont entraînés sur 1k heures de parole en français. Les modèles sont disponibles sur HuggingFace[4].

3. Application pour la reconnaissance des émotions

L'évaluation sur les représentations obtenues par wav2vec dans la tâche de REP a été réalisée par (Evain et al., 2021). Les représentations *W2V2-Fr-M-base* (modèle français) sont comparées avec les représentations MFB (Mel filterbank) et *XLSR-53-large* (modèle multilingue).

Les expérimentations sont effectuées sur le corpus RECOLA contenant 3.8 h d'enregistrements sans bruit, au sujet d'interactions spontanées entre des francophones résolvant une tâche collaborative à distanciel et AlloSat contenant 37 h de conversations réelles en français dans un centre d'appels. RECOLA possède des annotations de deux dimensions : la valence et l'arousal tandis que Allosat a une seule dimension d'annotation : la satisfaction. Concernant les modèles uitilisés, le modèle Linear-Tanh est un modèle simple basé sur une couche linéaire reliant les caractéristiques à une dimension émotionnelle, suivie d'une fonction hyperbolique tangente. Le modèle GRU (Gated recurrent unit) est GRU de 1 couche avec les couches cachées de 32 et de 64, suivi par la fonction de Linear-Tanh. Une optimiseur ADAM est utilisé et la patience a été fixée à 15 époques. Le coefficient de corrélation de concordance a été utilisé comme fonction de perte pour entraîner les modèles(Evain et al., 2021).

Les résultats (Evain et al., 2021) (Tableau 3) de l'évaluation montrent que les représentations Wav2vec2 surpassent les autres dans la tâche de prédiction des émotions, ce qui révèle l'avantage du modèle Wav2vec2 auto-supervisé dans les tâches ayant des sources limitées. Cette étude démontre l'état de l'art du modèle Wav2vec français pour la tâche REP.

[4]https://huggingface.co/LeBenchmark/wav2vec2-FR-M-large#wav2vec2-fr-m- -model-and-data-descriptions

Tableau 3 Résultats mesurés en coefficient corrélation concordance de la reconnaissance automatique des émotions sur le test set de RECOLA et AlloSat

Corpus		**RECOLA**		**AlloSat**
Model	**Feature**	**Arousal**	**Valence**	**Satisfaction**
Linear-Tanh	MFB	0.192	0.075	0.065
Linear-Tanh	W2V2-Fr-M-*base*	**0.385**	**0.090**	**0.193**
Linear-Tanh	XLSR-53-*large*	0.155	0.024	0.093
GRU-32	MFB	0.654	0.252	0.437
GRU-32	W2V2-Fr-M-*base*	**0.767**	**0.376**	**0.507**
GRU-32	XLSR-53-*large*	0.605	0.320	0.446
GRU-64	MFB	0.712	0.307	0.400
GRU-64	W2V2-Fr-M-*base*	**0.760**	**0.352**	**0.507**
GRU-64	XLSR-53-*large*	0.585	0.280	0.434

Chapitre 4. Apprentissage auto-supervisé pour les représentations textuelles

5. Modèles pré-entraînés

Nous avons cherché dans la littérature pour connaître des modèles d'apprentissage auto-supervisé les plus fréquents à propos du traitement automatique de la langue écrite. Etant donné que le nombre de citations reflète les influences qu'un article a exercé sur le monde académique, nous prenons le nombre de citations sur le site Google Scholar comme un indicateur qui mesure les influences d'un modèle. Le résultat de recherche sont résumés dans le Tableau 4.

Tableau 4 Résumé des modèles apprentissage auto-supervisé pour des représentations de textes

Modèle	**Référence**	**Descriptions**	**Nombre de citations sur Google scholar**
BERT	(Devlin et al., 2019)	En se basant sur le modèle Transformer, BERT est un modèle auto-supervisé qui initialement proposer de prendre en compte les contextes à gauche et à droite lors de l'encodage des représentations. BERT a connu un grand succès, il a fait avancer l'état de l'art de 11 tâches en TAL.	23925
DistilBERT	(Sanh et al., 2019)	En se basant sur le modèle BERT, DistilBERT réduit la taille de BERT par 40%, tout en conservant 97 % de ses capacités de compréhension du langage et en étant 60 % plus rapide, à travers les techniques de la distillation des connaissances adoptées pendant la phrase de pré-entraînement.	904
RoBERTa	(Liu et al., 2019)	RoBERTa a répété le processus d'entraînement de BERT, mais RoBERTa a été entraîné plus longtemps, sur les séquences plus longues, avec la taille de lot plus grand sur un plus grand nombre de données. En excluant la tâche de la prédiction de la phrase suivante, RoBERTa modifie dynamiquement le modèle de masquage appliqué aux données d'enraînement.	1524

XLNet	(Yang et al., 2019)	XLNet utilise un modèle autorégressif généralisé où le token suivant est dépendant de tous les tokens précédents, ce qui permet d'éviter les entrées corrompues par le masquage des mots, provoqué par BERT. XLNet produit des résultats qui surpassent ceux de BERT dans 20 tâches différentes, y compris l'analyse des sentiments.	3078
ELMo	(Peters et al., 2018)	ELMo est une représentation profond contextuelle des mots qui modélise à la fois les caractéristiques complexes de l'utilisation des mots (par exemple, la syntaxe et la sémantique), et la façon dont ces utilisations varient selon les contextes linguistiques (c'est-à-dire, pour modéliser la polysémie). Les représentations sont apris à travers un modèle de langage bidirectionnel profond (biLM).	7509
GloVe	(Pennington et al., 2014)	GloVe adopte une méthodologie améliorée pour l'encodage des mots basée sur une approche mathématique solide. Le modèle combine deux classes de méthodes pour la représentation distribuée des mots : la factorisation matricielle globale et les Skip-grams. Il souligne l'importance de considérer les probabilités de co-occurrence entre les mots plutôt que les probabilités d'occurrence d'un seul mot.	22550
Word2Vec	(Mikolov et al., 2013)	Deux architectures de modèles ont été proposées pour calculer les représentations vectorielles continues des mots en utilisant l'approche non supervisée : le modèle Bag-of-Words continu (CBOW) et le modèle de Skip-gram continu.	23401
FastText	(Bojanowski et al., 2017)	FastText s'attaque au problème général dans le domaine de traitement automatique de la langue écrite : des mots inconnus. FastText se distingue des modèles précédents par sa capacité à construire des embeddings de mots à un niveau plus profond en exploitant les sous-mots et les caractères.	6375
BART	(Lewis et al., 2019)	L'encodeur et le décodeur de BART sont reliés par une attention croisée. Chaque couche du décodeur effectue une attention sur l'état caché final de la sortie de l'encodeur. Ce mécanisme permet au modèle de générer une sortie qui est étroitement liée à l'entrée originale.	1023

XLM	(Lample & Conneau, 2019)	XLM utilise une technique de prétraitement connue (BPE) et trois approches CLM, MLM et TLM pour le pré-entraînement. Le modèle surpasse les autres modèles dans une tâche de classification interlinguistique (inférence de phrases dans 15 langues) et améliore significativement la performance du système de la traduction automatique lorsqu'un modèle pré-entraîné est utilisé pour l'initialisation du modèle de traduction.	680
Doc2vec	(Q. Le & Mikolov, 2014)	Doc2Vec est un encodeur pour la phrase. Doc2Vec propose une approche permettant de représenter des fragments de textes de longueur variable (phrases, paragraphes et documents) sous la forme de vecteurs denses de taille fixe, appelés vecteurs de paragraphes.	8054
FlauBERT	(H. Le et al., 2020)	FlauBERT était le premier modèle auto-supervisé pour le français. FlauBERT est basé sur l'architecture Transformer et il est entrainé par la tâche du masquage des mots.	2

Selon Tableau 4, on peut observer que BERT, GloVe et Word2vec sont les modèles les plus populaires. Nous avons choisi des modèles pour nos expériences parmi les modèles mentionné dans le tableau en considérant sa disponibilité en français et sa popularité dans le domaine du traitement automatique de la langue écrite.

Partie 2

-

Méthode

Chapitre 5. Extraction des représentations

En nous basant sur l'état de l'art à l'égard de la modélisation auto-supervisée et de la reconnaissance des émotions, nous présentons dans cette section les méthodes que nous avons adoptées pour extraire des représentations à partir des transcriptions de la parole. En vue de comparer les performances de différents types de représentations, à part les modèles d'apprentissage auto-supervisé, nous avons choisi d'utiliser le lexique émotionnel qui est une méthode conventionnelle attribuant une valeur fixe aux mots affectifs, ainsi que la méthode vectorielle conventionnelle : Tf-idf. En total, six types de représentations sont extraites de manière variés, dont quatre sont des représentations de l'apprentissage auto-supervisé. Les modèles de l'apprentissage auto-supervisé sont choisis en considérant à la fois la popularité des modèles et le multilinguisme des modèles, car nous voulons effectuer les évaluations en français et garder la possibilité que ces évaluations seront explorées pour d'autres langues dans le futur.

La programmation du projet de ce mémoire est réalisée en python. En profitant la bibliothèque *transformers*, nous avons accès aux modèles pré-entraînés auto-supervisés tels que FlauBERT, BERT multilingue, XLM. Le modèle FastText est disponible à partir de la bibliothèque *fasttext.* Concernant la partie de Tf-idf, dans ce mémoire la bibliothèque *scikit-learn* est utilisée. Le lexique émotionnel MEmoLon est téléchargeable sous forme du ficher .tsv.

Avant de passer les transcriptions par les modèles de l'extraction des caractéristiques, une étape de normalisation est nécessaire pour que les transcriptions soitent plus propre et les résultats d'évaluations soient valides. D'abord, nous avons traité les ponctuations. Nous avons enlevé toutes les ponctuations dans les transcriptions, y compris les apostrophes à travers une fonction *removePunc.* Nous avons aussi passé toutes les phrases en minuscule, car certains modèles sont sensibles à la différence entre les majuscules et les minuscules et nous avons choisi les versions minuscules de tous les modèles utilisés dans ce mémoire.

Ensuite, à propos de la tokenisation, nous avons adopté trois stratégies. Notre première stratégie pour la tokenisation est d'utiliser le tokénizer qui est fourni avec le modèle. Les modèles sur la bibilothèque transformer HuggingFace sont foournis avec leurs propres tokénizers. Dans l'intention de respecter la démarche proposée par le créateur du modèle, ainsi que de maximiser les capacités du modèle, nous décidons d'utiliser ces tokénizers fournis.

Notre deuxième stratégie pour la tokenisation repose sur les expressions régulières. Nous avons construit un simple tokénizer à la base de la fonction *build_tokenizer()* de la bibliothèque scikit-learn. Nous avons utilisé ce tokénizer pour l'approche Tf-idf et l'approche MemoLon. Les expressions régulières que nous utilisons est « \b\w+\b » dont « \b » définit la frontière des mots et « \w » correspond à tout caractère de mot ([a-z A-Z 0-9]). Cette expression régulière nous permet de tokenizer des phrases par l'espace, ce qui est convenable pour les textes sans ponctuations.

Notre troisième stratégie se rapporte au tokénizer de NLTK. La bibliothèque NLTK dispose d'un tokénizer *word_tokenize* pour la langue française, nous avons l'utilisé pour le modèle FastText. La raison pour laquelle nous n'avons pas utilisé le tokénizer Europarl utilisé dans le pré-entraînement de FastText est que le tokénizer Europarl est écrit en langage perl et que nous avons rencontré des difficultés à impliquer cette fonction perl dans notre pipeline. Par conséquent, nous avons choisi le tokénizer NLTK car il fournit un tokénizer pour le français.

Les dimensions des vecteurs varient en fonction des modèles, nous présentons cette différence en détails dans les sections suivantes, mais tous les modèles génèrent les représentations par token, autrement dire, nous obtenons des représentations dont chaque token est représenté par une taille de vecteurs correspondants. Tableau 5 montre le nombre de représentations obtenues par token de chaque approche.

Tableau 5 Résumé des approches et la taille des vecteurs de représentations obtenues

Approches	**Taille des vecteurs de représentations obtenues par token**
flaubert_base_cased	3072
xlm-mlm-17-1280	5120
fastText-cc.fr.300.bin	300
Memolon	1
Tf-idf	1
bert-base-multilingual-cased	3072

1. Par l'apprentissage auto-supervisé

Le choix des modèles repose sur plusieurs critères : 1- Le modèle doit être disponible et accessible gratuitement. 2- Le modèle doit être multilingue ou au moins avoir une version pour le français. 3- Pour que le résultat d'expérimentation soit représentatif, le modèle est sélectionné à partir de l'état de l'art des techniques de l'apprentissage auto-supervisé dans le domaine du traitement automatique de la langue écrite.

Il existe plusieurs façons d'extraire des représentations à partir de modèles pré-entraînés. Selon les résultats des évaluations sur le modèle BERT utilisant des représentations de différentes couches dans la tâche CoNLL-2003 Reconnaissance d'entités nommées (Devlin et al., 2019), nous constatons que la concaténation des quatre dernières couches surpassent les autres, même si les différences entre les approches sont mineurs. (cf. Figure 3 qui est un extrait du résultat publié par (Devlin et al., 2019)).

Feature-based approach ($BERT_{BASE}$)		
Embeddings	91.0	-
Second-to-Last Hidden	95.6	-
Last Hidden	94.9	-
Weighted Sum Last Four Hidden	95.9	-
Concat Last Four Hidden	96.1	-
Weighted Sum All 12 Layers	95.5	-

Figure 3 Extrait des résultats du modèle BERT utilisant de différentes approches pour extraire des représentations dans la tâche CoNLL-2003 Reconnaissance d'entités nommées (Devlin et al., 2019)

Nous avons donc suivi l'approche démontrée ci-dessus et choisi de concaténer les quatre dernières couches cachées pour extraire des représentations à partir des modèles pré-entraîné auto-supervisés.

1.1. FlauBERT

FlauBERT(H. Le et al., 2020) est un ensemble de modèles auto-supervisés spécialement entraîné sur une très grande quantité de données françaises de divers corpus sur des genres variés. En se basant sur l'architecture de BERT et Transformer, FlauBERT est pré-entraîné sur la tâche MLM (Masqued language modeling) qui consiste à apprendre à prédire des tokens masqués de façon aléatoire dans une phrase. Ici, vu que MLM (Modélisation du langage masqué) est une méthode d'entraînement assez fréquente pour des

modèles de l'apprentissage auto-supervisé, nous voulons la présenter en détails. Le mécanisme derrière MLM est simple : En prenant une phrase comme entrée, le modèle doit masquer 15% des mots et puis fait passer la phrase masquée entière dans le modèle, finalement, il doit prédire les mots masqués. Cette méthode est différente des réseaux neuronaux récurrents traditionnels (RNN) qui traitent généralement les mots les uns après les autres, ou des modèles autorégressifs comme le GPT qui masquent en interne les futurs tokens.

FlauBERT est utile pour extraire des représentations vectorielles ou de fine-tuning pour d'autres tâches en ajoutant des couches spécifiques. Etant l'un des premiers modèles auto-supervisé pour le français, FlauBERT non seulement arrive à surpasser CamenBERT, un autre modèle auto-supervisé français, dans la tâche de classification des textes, sur FLUE, mais aussi à surpasser le modèle BERT multilingue sur certaines tâches de FLUE. Nous avons choisi ce modèle en raison de ses excellentes performances en matière de TALN en français.

FlauBERT publie deux modèle différents, la version BASE [L=12, H=768, A=12] et la version LARGE [L=24, H=1024, A=16]. L est le nombre des modules Transformer, H est la taille cachée, A est le nombre de têtes auto-attention. Dans ce mémoire, nous utilisons *flaubert_base_cased* à travers la bibliothèque *transformers* de HuggingFace, en considérant la limite des capacités de notre appareil.

En pratique, après avoir chargé le modèle flaubert_base_case, nous traitons d'abord chaque phrase des transcriptions par FlauberTokenizer, ensuite nous utilisons la bibliothèque *torch* pour traiter les tensors et faire la concaténation des tenseurs des quatre dernières couches. Enfin, nous utilisons la bibliothèque *numpy* pour transformer les tenseurs en vecteurs et nous enregistrons ces derniers dans les fichiers .csv acompagnés des tokens correspondants.

2. Finalement, la taille de vecteurs qu'on a obtenue pour chaque token est 3072.

2.1. BERT Multilingue

Se basant sur BERT, BERT Multilingue est un modèle de langue unique pré-entraîné à partir de corpus monolingues en 104 langues. Comme FlauBERT, ce modèle est aussi sensible à la différence entre les majuscules et les minuscules, et ce modèle est aussi entraîné sur la tâche MLM.

BERT Multilingue a une architecture identique à BERT, il est formé de l'encodeur Transformer bidirectionnel multicouche. Les représentations extraites à partir de BERT Multilingue sont contextualisées par le contexte gauche et droite, au niveau de la phrase. BERT Multilingue utilise WordPiece pour faire la tokenisation.

Récemment, BERT Multilingue a démontré ses excellentes performances dans l'aspect de l'apprentissage par le transfert « zero-shot », dans lequel les annotations spécifiques à une tâche dans une langue sont utilisées pour affiner le modèle pour l'évaluation effectuée dans une autre langue(Pires et al., 2019).

Nous choisissons ce modèle parce que BERT obtient de résultats impressionnants en TALN et nous voudrons inclure une version de BERT qui peut traiter le français pour que nous puisse comparer les performances des modèles multilingues et des modèles spécialisés pré-entraîné en français, dans la tâche REP.

Dans ce mémoire, nous utilisons le modèle *bert-base-multilingual-cased*, car nous voudrons rester cohérent avec d'autre modèle en termes de la minuscule des caractères. En pratique, le processus de l'extraction est largement similaire à celui de FlauBERT, comme nous utilisons ces deux modèles à partir de la même bibliothèque. Finalement, la taille de vecteurs qu'on a obtenue pour chaque token est 3072.

2.2. XLM

XLM (Lample & Conneau, 2019) est un modèle auto-supervisé à la base de Transformer. Trois approches sont proposées pour le pré-entraînement : TLM (Modélisation du langage de la traduction), CLM (Modélisation du langage causal) et MLM (Modélisation du langage masqué). L'auteur a évalué les différentes combinaisons de ces trois approches et ses performances sur les tâches de la classification multilingue et de la traduction automatique. Les résultats des expériences indiquent que CLM et MLM sont des approches les plus performantes pour obtenir des représentations solides multilingues.

Les modèles XLM sont entraînés sur les données de Wikipédia en différentes langues. Les phrases brutes sont d'abord extraites par WikiExtractor, ensuite tokénisé par Moses et finalement divisés en unité de sous-mots en utilisant BPE (Byte Pair Encoding). BPE permet de créer un vocabulaire partagé entre des langues ayant des mêmes alphabets.

Plusieurs versions de modèle pré-entraîné sont fournies, nous utilisons la version *xlm-mlm-17-1280*. Ce modèle est pré-entraîné sur l'approche MLM, dans 17 langues, y

compris le français. Une autre version pour 100 langues est aussi disponible, mais en considérant que le français est déjà inclus dans la version plus petite, il nous semble inutile d'utiliser le modèle plus grand.

En pratique, nous utilisons *xlm-mlm-17-1280* par la bibliothèque *transfomers* de HuggingFace, comme les modèles précédents, cependant, XLM nous demande de préciser la langue cible à travers un *language_id.* Après la concaténation des représentations des quatre dernières couches, nous obtenons un vecteur de 5120 valeurs par token, ce qui est beaucoup plus volumineux que les modèles tels que FlauBERT et BERT Multilingue.

2.3. FastText

À la différence des autres modèles auto-supervisés, FastText (Grave et al., 2018) n'est pas un modèle similaire à BERT. FastText dispose de deux types de modèles, l'un de skipgram et l'autre de CBOW. Dans ce mémoire, nous concentrons sur les modèles CBOW. Pour comprendre le mécanisme de FastText, nous devons d'abord présenter CBOW (continuous bag-of-words). Le modèle CBOW prédit le mot cible en fonction de son contexte qui est représenté comme un sac des mots contenus dans une fenêtre de taille fixe autour du mot cible. Par exemple, si la phrase était « Je suis un étudiant à l'UGA » et le mot cible était « étudiant », le modèle CBOW prendrait tous les mots dans une fenêtre autour e.g. {suis, un, à, UGA, l'} et puis prédirait le mot cible en utilisant la somme des vecteurs de ces mots.

Les modèle FastText que nous avons utilisé a été entraîné en utilisant CBOW avec des poids de position, en dimension 300, avec des n-grams de caractères de longueur 5, une fenêtre de taille 5 et 10 négatifs. FastText est disponible pour un total 157 langues, y compris le français. Les corpus d'entraînement sont des données non-étiquetées qui parviennent de Wikipédia et Common Crawl qui contient une ressource de grande quantité sur différents sujets. Pour le français, FastText utilise le tokénizer de Europarl.

Nous utilisons ce modèle à travers la bibliothèque fasttext, les instructions d'utilisation sont clairement présentées dans leur site officiel, alors que le chargement du modèle prend beaucoup de temps selon notre expérience.

3. Par le lexique relatif aux émotions

Nous voudrons comparer les performances des représentations de l'apprentissage auto-supervisé et des représentations plus traditionnelles, par exemple, à la base du lexique des émotions.

3.1. Introduction

Les lexiques des émotions est une approche qui est largement utilisée pour l'analyse des émotions. Les lexiques d'émotions sont des dictionnaires de mots qui sont étiquetés ou annotés en fonction d'une ou plusieurs catégories d'émotions.

3.2. MEmoLon

Nous avons choisi le lexique MEmoLon comme base lexicale. Parmi tous les lexiques des émotions, MEmoLon nous intéresse le plus parce que MEmoLon (Buechel et al., 2020) est un ensemble des lexiques émotionnels en 91 langues, qui sont générés automatiquement par un modèle de traduction bilingue et un modèle embedding de la langue cible à partir des ressources langagières relatives aux émotions.

Le processus de génération se déroule comme suit : tout d'abord, un lexique des émotions de la langue de source est divisé en trois parties, « entraînement », « validation » et « évaluation ». Ensuite, en utilisant un modèle de traduction, le lexique de source est traduit dans la langue cible pour construire un premier lexique des émotions de la langue cible. Les labels des émotions du lexique cible sont copiés à partir de la langue de source, donc reste inchangés. Enfin, ce lexique cible est utilisé pour entrainer un modèle embeddings (langue de cible) à prédire les labels. Une fois l'entraînement terminé et les paramètres optimisés par la partie « validation », ce modèle est utilisé pour prédire des nouveaux labels pour des mots dans la langue cible.

MEmoLon fournit un total de 8 labels émotionnels, chaque mot inclus dans le lexique dispose d'une seule valeur pour chaque catégorie. L'utilisation de MEmoLon est assez simple, nous parcourons le lexique par trouver la valeur correspondante pour chaque token. Le lexique MEmoLon est téléchargeable sur le site https://github.com/JULIELab/MEmoLon.

4. Par l'approche Tf-idf

Tf-idf (terme frequency-inverse document frequency) est une approche représentant des mots par un score calculé à partir la fréquence des mots et la fréquence inverse de document. Ignorant les contextes, cette approche statistique permet d'évaluer l'importance d'un terme donné dans un document, car Tf évalue l'occurrence d'un terme et idf évalue la rareté d'un terme dans un document, et par la combinaison de ces deux indicateurs, tf-idf reflète correctement la pertinence d'un mot dans un document. La formule Tf-idf s'écrit comme suit :

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

$W_{x,y}$ désigne le score d'importance d'un terme x quelconque dans un document y ; $tf_{x,y}$ désigne la fréquence du terme x dans le document y ; df_x désigne le nombre de documents contenant x et N est le nombre total de documents.

En pratique, nous utilisons le tfidfvectorizer de la bibliothèque scikit-learn pour calculer tf-idf. Différent des autres modèles utilisés, une étape de pré-traitement est mise en œuvre pour le calcul de Tf-idf, car la nature de tf-idf implique que la mesure soit réalisée au niveau du document, alors que notre système transforme les mots en vecteurs au niveau de la phrase. Nous avons donc d'abord calculé tf-idf à partir de chaque document de transcriptions et enregistré les scores accompagnés avec le mot correspondant dans un fichier json dont la structure est la suivante :

Document_1

| -- phrase_1

| -- mot_1 : score tf-idf

| -- mot_2 : score tf-idf

Chapitre 6. Alignement et Rééchantillonnage

La parole possède naturellement une caractéristique de continuité. Si nous voulons prédire les émotions dans la parole continue à travers des transcriptions, un problème crucial est l'alignement des représentations textuelles avec le temps.

1. Alignement

Comment associer les tokens avec le schéma chronologique de la parole ? Nous commençons par utiliser les frontières des morceaux de parole. On définit un morceau de parole comme entre une section acoustique entre deux pauses. Normalement, le temps du début et le temps de la fin de chaque morceau sont fournis avec les transcriptions. Pour chaque morceau de parole, nous calculons la distance du temps entre les frontières, ensuite, nous divisons ce temps par le nombre de tokens dans la transcription. Nous obtenons le schéma chronologique de chaque token.

Connaître le schéma chronologique de chaque token est insuffisant pour faire des prédictions sur la parole continue, car les intervalles de temps sont inégaux alors que nous avons besoin d'un intervalle régulier pour faire la prédiction. Nous devons trouver l'intervalle le plus adapté pour le corpus. Notre méthode pour faire cela est de trouver le plus petit intervalle dans toutes les transcriptions. Dans notre expérience sur RECOLA, le plus petit intervalle de temps entre les tokens est 0.1026s, environ 100ms.

2. Rééchantillonnage

Après avoir aligné les tokens dans le schéma chronologique et avoir trouver le plus petit intervalle de temps entre les tokens, nous devons rééchantillonner les transcriptions ainsi que les annotations à cet intervalle.

Un processus de rééchantillonnage est nécessaire pour les transcriptions, ainsi que pour les annotations, cependant, les deux rééchantillonnages sont effectués de façon différente. Pour les transcriptions, par un pas de 100ms, on répète les tokens accompagnés avec les représentations vectorielles lorsque le point du temps est entre des frontières chronologiques des tokens, ou, on ajoute des zéros de la même dimension que celle des représentations lorsqu'il n'y a pas de parole. La Figure 4 montre un organigramme qui présente l'extrait du processus de rééchantillonnage des transcriptions.

Pour les annotations, nous utilisons la fonction *numpy.interp* de la bibliothèque Numpy pour faire le rééchantillonnage. Cette fonction *interp* sert à faire l'interpolation linéaire unidimensionnelle pour des points d'échantillonnage à croissance monotone, c'est-à-dire, insérer des valeurs cohérentes dans l'ancien échantillonnage pour former les valeurs dans la nouvelle fréquence.

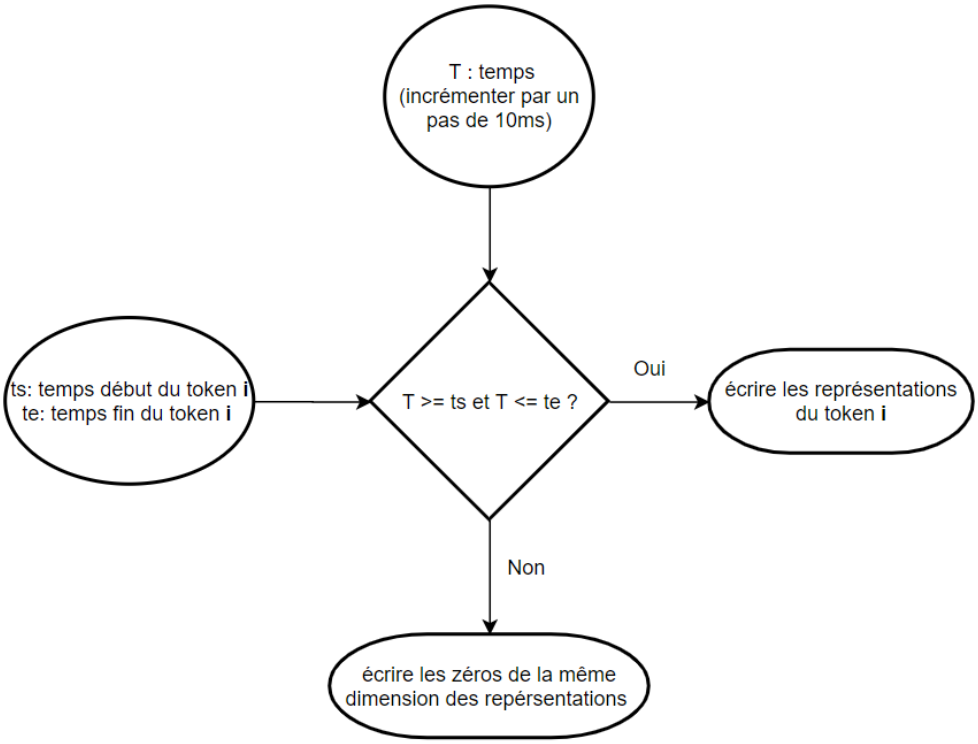


Figure 4 Démonstration du processus d'rééchantillonnage des tokens

Partie 3

-

Expérimentation

Chapitre 7. Introduction

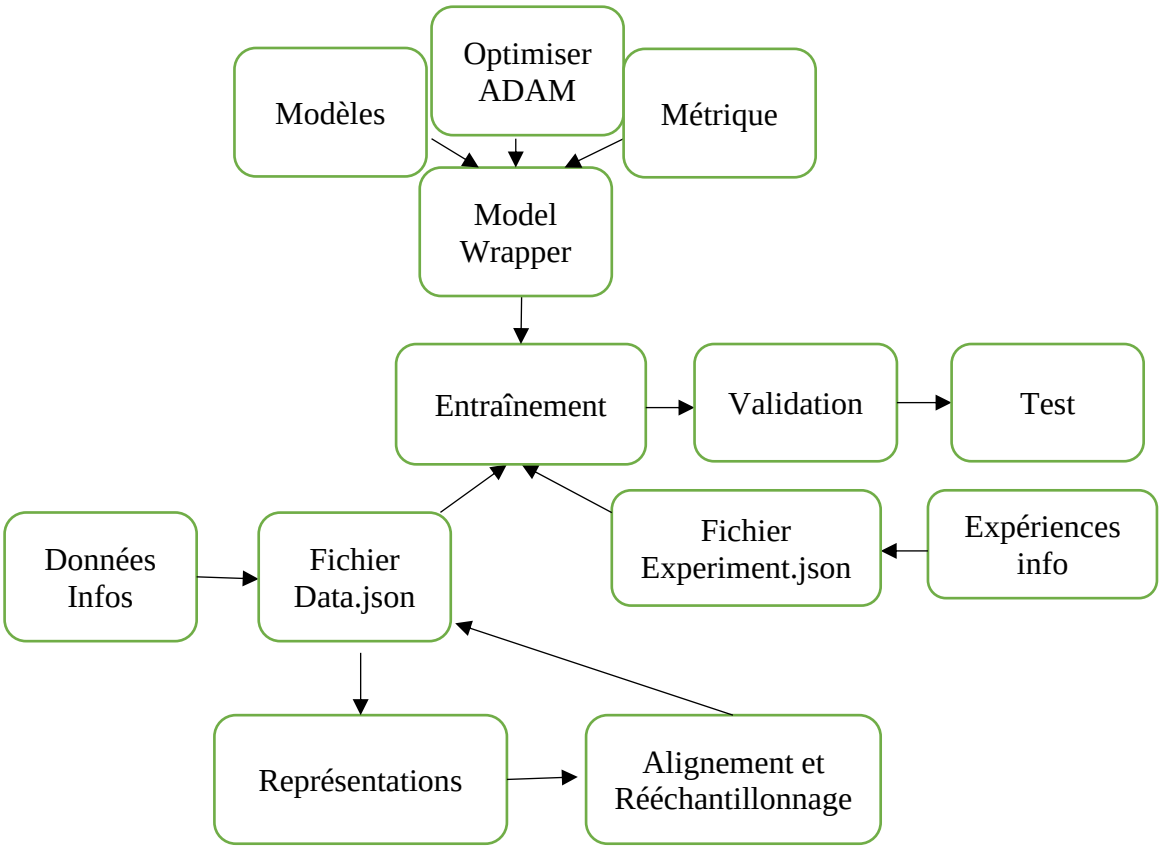
Nous avons réalisé une série d'expérimentations pour évaluer les performances des représentations que nous avons obtenues. L'idée principale est d'évaluer les représentations obtenues par les différentes approches, et, parmi les approches auto-supervisées, de comparer les résultats de différents modèles, sur la tâche de la prédiction en temps continu des émotions, tout en se basant sur le corpus RECOLA contenant les discours affectifs spontanés en français. Dans cette partie, nous décrivons l'organisation de nos expériences, ainsi que la mise en place des expériences.

1. Tâche

La tâche d'évaluation est la prédiction des émotions en temps continu. La prédiction en temps continue est une tâche utilisée fréquemment dans le domaine du traitement automatique de la parole en raison la continuité des signaux vocaux en termes du temps.

2. Vue d'ensemble

Figure 5 Vue d'ensemble de l'expérimentation



La Figure 5 présente la vue d'ensemble de notre processus d'expérimentation. Un point à remarquer est que nous avons utilisé un fichier json pour organiser toutes les informations des données et les représentations, ainsi qu'un autre fichier json pour sauvegarder les informations et les résultats de chaque expérience.

Selon Figure 5, on peut voir que les informations des données sont d'abord enregistrées dans un fichier json « data.json », ensuite ces informations sont utilisés dans le processus de l'extraction des représentations, de l'alignement et du rééchantillonnage. Les informations des représentations sont ajoutées dans data.json. Au final, la structure du fichier « data.json » est comme la suivante :

Corpus

| -- ID1

| -- partition

| -- chemin du fichier wav

| -- chemin de transcription

| -- annotation

| -- chemin

| -- représentations

| -- représentation1

| -- chemin

| -- taille

| -- représentation2

| -- chemin

| -- taille

Les modèles et l'optimiseur, ainsi que les codes concernant les métriques d'évaluations sont enveloppés dans une enveloppe de modèle « model wrapper » pour faciliter le processus de l'utilisation. Nous définissons les paramètres de l'expériences et nous les enregistrons dans un fichier json « experiment.json ». Ce fichier organise nos expériences et sauvegarde les résultats de chaque expérience.

Chapitre 8. Corpus

Nous avons utilisé RECOLA comme le corpus de nos expérimentations.

1. Présentation générale

RECOLA (Ringeval et al., 2013) est un corpus multimodal qui contient les interactions collaboratives, affectives, spontanées de 46 participants francophones. Les 46 participants ont été invité à discuter une tâche de survie en binôme, par l'application visioconférence Skype. En total, le corpus RECOLA contient 3.8h d'enregistrements audio.

2. Répartition

Le corpus est divisé en trois partie, la partie d'entraînement, la partie de validation et la partie de test, chacune occupe un tiers des données.

3. Transcriptions

Les transcriptions sont d'abord générées automatiquement par Google, ensuite, elles sont corrigées manuellement par deux différents annotateurs, nous avons aussi contribué à cette vérification. En utilisant Audacity, nous arrivons à assurer les frontières des transcriptions conforment aux frontières de la parole, ce qui rend le résultat de l'alignement plus fiable et solide.

4. Annotation Gold standard

Les annotations de Gold standard sont fournies par RECOLA. Les annotations Gold standard sont les annotations qui sont manuellement annotées et vérifiés par des experts. RECOLA est annoté par 6 assistants francophones, sur deux dimensions en temps continu : l'arousal et la valence. L'annotation Gold standard est obtenue en faisant la moyenne de ces six annotations. L'arousal mesure les émotions de la passivité à l'activité, tandis que la valence mesure les émotions de la négativité ou de la positivité(Ringeval et al., 2013).

Chapitre 9. Modèles

Nous avons utilisé deux modèles dans nos expériences. Le premier est un modèle simple à la base d'une couche linaire, suivi par la fonction hyperbolique tangente. Le deuxième est GRU de 1-couche avec les couches cachées D = [32, 64, 128], suivi par la fonction hyperbolique tangente. Nous avons utilisé l'optimiseur ADAM et la patience a été fixé à 15 époques. Les modèles sont construits à travers la bibliothèque torch.

1. Linear-Tanh

1.1 Réseaux neuronaux du type feed-forward

Les réseaux neuronaux du type feed-forward sont l'un des modèles les plus simples et les plus classiques. Le principe de ce type de réseaux neuronaux est de faire passer les entrées par les nœuds cachés dans les couches cachées, couche par couche, jusqu'à la couche de sortie, dans ce processus, les informations sont transférées dans une seule direction, sans feedback. Suivi par une fonction d'activation, ce processus est itéré plusieurs fois, et chaque fois les sorties sont produites, la perte est calculée et une correction est envoyée au modèle (processus de rétropropagation). De cette façon, le modèle apprend des poids et des biais permettant de faire des prédictions correctes.

Dans notre cas, notre modèle Linear-Tanh dispose d'une seule couche, donc les entrées sont passées directement à la couche de sortie qui est une couche linaire, nous utilisons la fonction hyperbolique tangente comme fonction d'activation, et le coefficient de corrélation de concordance comme fonction de perte. Des entrées sont passées par le modèle linaire et nous obtenons des sorties qui sont ensuite appliqués à la fonction Tanh. Ensuite, les erreurs entre la cible prédite et la cible de référence sont calculées en utilisant le coefficient corrélation de concordance. Enfin, ces informations des erreurs sont renvoyées aux modèles pour apprendre et ajuster les poids et les biais.

1.2 Fonction hyperbolique tangente

La fonction hyperbolique tangente est souvent utilisée comme fonction d'activation dans les réseaux neuronaux. Les fonctions d'activation servent transformer de manière non-linéaire les données. Les fonctions d'activation sont appliquées à chaque couche car sa non-linéarité permet de changer la représentation des données, d'avoir une nouvelle approche sur ces données. La formule ci-dessous est la formule hyperbolique tangente. Tanh est capable de transformer l'entrées en une sortie de [-1, 1].

$$f(x) = \tanh(x) = \frac{2}{1 + e^{-2x}} - 1$$

1.3 Fonction de perte : CCC

Les fonctions de perte sont utilisées pour calculer des erreurs entre la cible prédite et la cible de référence. Nous avons utilisé CCC le coefficient corrélation de concordance comme la fonction de perte pour les deux modèles. La raison pour laquelle nous avons choisir cette fonction est que CCC combine la corrélation et MSE (mean squared error) en une seule fonction différentiable, ce qui permettre CCC de calculer à la fois la corrélation entre les séquences régressives des sorties et les contours d'émotions à valeur continue, et mesurer l'écart entre la cible et la référence. CCC est donc une façon très efficace pour évaluer les performances de REP en temps continu.

2. GRU

En plus du modèle linaire, nous avons aussi utilisé un modèle GRU (Gated Reccurent Units) de 1-couche, avec des couches cachées de 32, 64 et 128.

2.1 Réseaux neuronaux récurrents RNN

RNN est différent des modèles feed-forward car il prend en compte non seulement l'entrée actuelle mais aussi l'entrée précédente par un mécanisme de boucle (feedback loop). Les RNN peuvent « mémoriser » les informations des entrées précédentes. Les informations précédentes sont appelées l'état caché. Les RNN fonctionnent de manière séquentielle, donc ils prennent un mot à la fois comme l'entrée. Chaque mot est encodé en utilisant les informations du mot actuel et l'état caché, de cette manière, lorsqu'on arrive au dernier mot des séquences, RNN encoderait les informations de tous les mots dans les étapes précédentes. En plus, contrairement aux modèles feed-forward où les poids et les biais de chaque connexion sont différents, les connexions RNN ont le même poids et le même biais pour toutes les couches, ce qui réduit la complexité de paramètres croissants. RNN utilise aussi le mécanisme de rétropropagation pour entraîner le modèle.

L'avantage de cette architecture est que RNN mémorise toutes les informations précédentes, ce qui rend le modèle utile pour faire la prédiction sur les tâches qui nécessite de mémoriser les informations précédentes pour faire la prédiction, par exemple, la reconnaissance de la parole et la traduction automatique.

Les défauts de cette architecture sont liés à la nature de RNN. Premièrement, RNN a du mal à retenir les informations des étapes précédentes lorsque le nombre des étapes s'accroît. Parce que RNN utilise la combinaison de l'information actuelle et l'information de toutes les étapes précédentes à chaque tour de la boucle, au long de l'augmentation du nombre des étapes, les informations de premier mot occupent de moins en moins d'importance. Ce phénomène est connu sous le nom de « mémoire de court terme ».

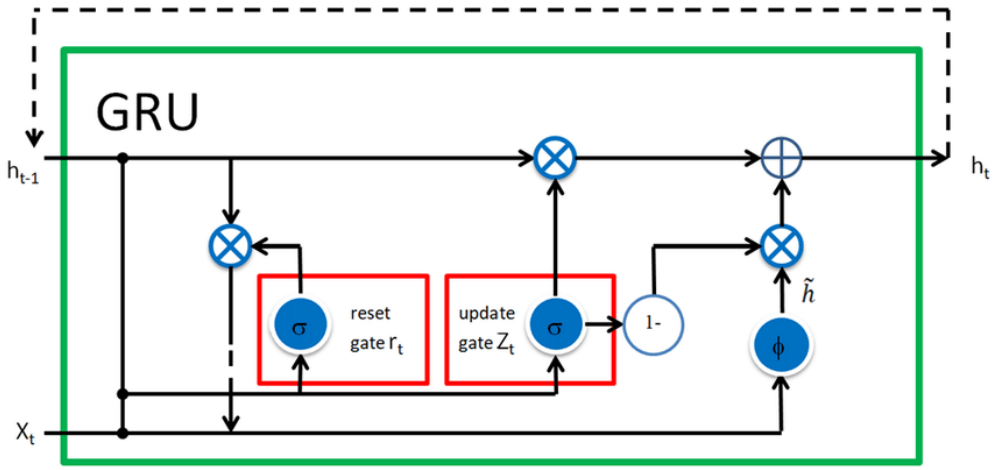
Deuxièmement, le problème de gradient disparu est dû à la nature de la rétropropagation. Le gradient est la valeur utilisée pour ajuster les poids internes du réseau, permettant au réseau d'apprendre. Plus le gradient est grand, plus les ajustements sont importants et vice versa. Lors de la rétropropagation, chaque nœud d'une couche calcule son gradient en fonction des effets des gradients de la couche précédente. Ainsi, si les ajustements des couches précédentes sont faibles, les ajustements de la couche actuelle seront encore plus faibles. Cela entraîne une réduction exponentielle des gradients au fur et à mesure qu'ils se propagent vers le bas. Les couches précédentes ne parviennent pas à apprendre car les poids internes sont à peine ajustés en raison de gradients extrêmement faibles. C'est cela le problème de gradient disparu.

2.2 GRU (Gated Recurrent Units)

Les unités récurrentes avec portes (GRU) fournissent une solution contre le problème de gradient disparu, ainsi que le problème de mémoire de court terme, en incluant « la porte de mise à jour » et « la porte de réinitialisation ». Ces deux portes sont des vecteurs décidant quelles informations doivent être transmis à la sortie. La Figure 6 venant de (Su & Kuo, 2019) montre la structure d'une unité GRU. Dans la figure, h_{t-1} désigne l'état précédent, x_t désigne l'entrée, r_t désigne la porte de réinitialisation et z_t désigne la porte de mise à jour, $\tilde{h}$ désigne l'état actuel.

La porte de réinitialisation est calculée à partir h_{t-1} et x_t, suivi par une fonction Tanh. La porte de mise à jour est calculée de la même manière. L'état actuel $\tilde{h}$ est calculé par le produit de Hadamard de h_{t-1} et r_t , en combinant x_t. Finalement, l'état à conserver h_t est calculé par $h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h}$.

Figure 6 Le schéma d'une cellule de GRU



2.3 Fonction d'activation et fonction de perte

Comme le modèle linear-Tanh, notre GRU modèle utilise la fonction hyperbolique tangente comme la fonction d'activation et calculer la perte en utilisant le coefficient de corrélation de concordance.

3. Optimiseur : Adam

Adam (Kingma & Ba, 2014) est une extension de la descente de gradient stochastique qui vise à modifier les poids du réseau de manière itérative en fonction des données d'entraînement. Adam est l'un des optimiseurs le plus largement utilisé dans le domaine de l'apprentissage profond en raison de son efficacité, son excellente performance, sa simplicité en termes de l'implémentation etc. A la différence de la descente de gradient stochastique qui utilise un seul taux d'apprentissage pendant le processus d'entraînement, Adam calcule les taux d'apprentissage adaptatifs individuels pour différents paramètres à partir des estimations des premier et second moments des gradients. En bref, Adam est un remplacement de la descente de gradient stochastique qui sert à entraîner le modèle en réduisant la perte.

Nous avons utilisé Adam à travers la bibliothèque *torch*, et l'avons configuré avec les paramètres de configuration par défaut. La patience a été fixée à 15 époques.

Chapitre 10. Métrique d'évaluation

1. CCC

CCC désigne le coefficient corrélation concordance. Cette mesure d'évaluation est spécialement adapté pour évaluer les tâches de prédiction en temps continu (Weninger et al., 2016). Nous avons choisi cette métrique d'évaluation pour nos expérimentations car CCC non seulement mesure la corrélation entre la cible prédite et la référence, mais aussi prend en compte l'écart moyen entre la cible prédite et la référence. Pour des tâches REP en temps continu, CCC est une des métriques d'évaluation les plus adaptées. Comme CCC mesure la corrélation et la concordance, une valeur proche de 1 signifie la concordance est forte.

2. RMSE

RMSE est l'abréviation de root-mean-square error. RMSE sert à mesurer l'écart des erreurs entre la valeur de référence et la valeur prédite. RMSE accorde un poids relativement élevé aux erreurs importantes. Cela signifie que la RMSE est plus utile lorsque des erreurs importantes sont particulièrement indésirables. Un RMSE faible signifie que les résultats comportent moins d'erreurs.

Chapitre 11. Conclusion de l'expérimentation

Nous avons effectué en total 48 expériences, sur deux modèles, six diverses représentations en deux dimensions (arousal et valence). Nous visualisons les évolutions de la perte sur la partition de validation du modèle GRU-128 et du modèle Linear-tanh dans le Tableau 6 et le Tableau 7, en prenant la dimension arousal comme un exemple. Il est évident que les pertes de Linear-tanh sont plus élevées que celles de GRU-128, ce qui reflète les performances de ces deux modèles.

Nous observons que Linear-tanh a besoins plus de époques que GRU-128 pour atteindre l'optimisation, surtout lors des représentations fastText et Tf-idf. Nous pouvons donc déduire que GRU-128 est plus efficace. Comme GRU-128 dispose d'une structure plus complexe, cette différence sur l'efficacité est raisonnable.

A travers Tableau 6, nous constatons que les représentations auto-supervisées telles que BERT-M, FlauBERT, XLM, FastText, arrivent à minimiser la perte en utilisant moins de époques que d'autre représentations, ce qui peut être indique que ces représentations auto-supervisée contiennent plus d'informations.

Tableau 6 Evolution de la perte sur la partition validation du modèle GRU-128 sur la dimension arousal

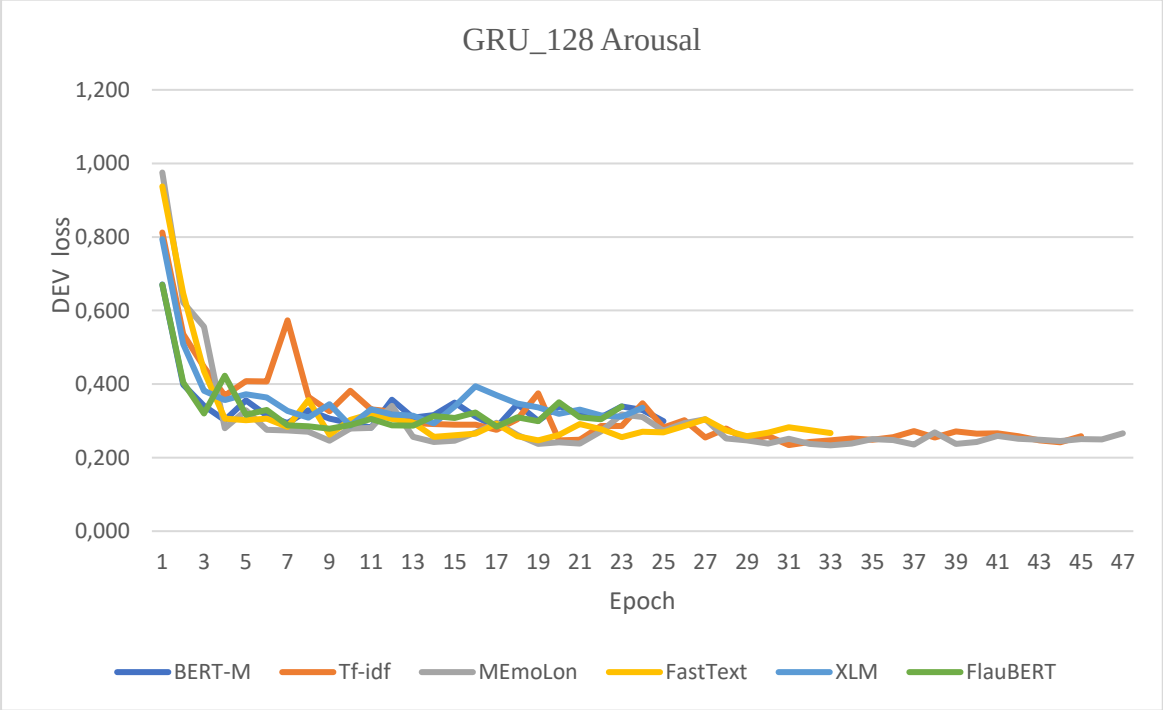
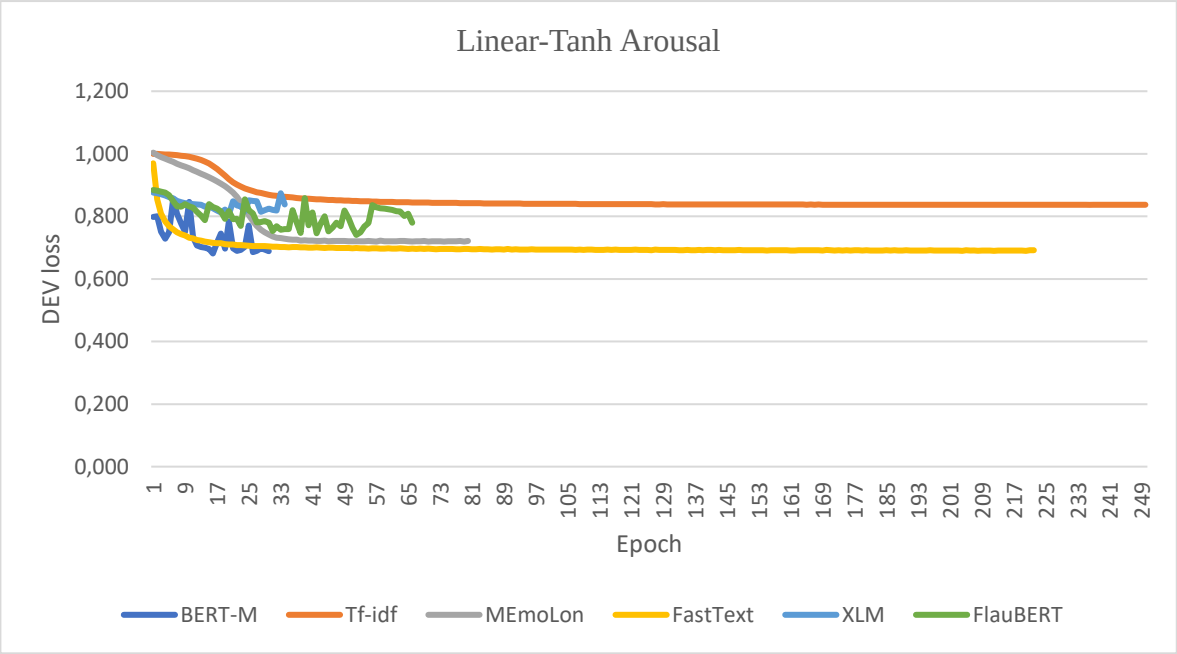


Tableau 7 Evolution de la perte sur la partition validation du modèle Linear-tanh sur la dimension arousal



Partie 4

-

Résultats et Analyses

Chapitre 12. Résultats et Analyses

Nous présentons dans cette section les résultats de l'expérimentations et nous abordons les analyses pertinentes en nous basant sur ces résultats. Tout d'abord, dans Tableau 8 et Tableau 9, les résultats mesurés par CCC (coefficient corrélation concordance) et RMSE (erreur quadratique moyenne) sont présentés en fonction des modèles. Ensuite, afin de visualiser les résultats et de montrer directement les corrélations entre la cible et la référence, nous choisissons une des meilleures représentations et faisons un graphique pour comparer les sorties et les références. Enfin, nous comparons nos résultats avec ceux obtenus avec des représentations acoustiques.

1. Résultats en CCC

Tableau 8 présente les résultats mesurés par coefficient corrélation concordance sur le test set de RECOLA.

Tableau 8 Résultats de Reconnaissance des émotions de la parole en temps continu (Coefficient Corrélation Concordance) sur test set de RECOLA

Corpus		RECOLA	
Modèle	Représentations	Arousal	Valence
Linear-Tanh	flaubert_base_cased	0.222	0.015
Linear-Tanh	xlm-mlm-17-1280	0.148	0.025
Linear-Tanh	fastText-cc.fr.300.bin	0.278	**0.05**
Linear-Tanh	Memolon	0.251	0.041
Linear-Tanh	Tf-idf	0.154	0.043
Linear-Tanh	bert-base-multilingual-cased	**0.287**	0.018
GRU-32	flaubert_base_cased	0.687	0.163
GRU-32	xlm-mlm-17-1280	0.659	0.139
GRU-32	fastText-cc.fr.300.bin	0.763	0.255
GRU-32	Memolon	**0.772**	0.233
GRU-32	Tf-idf	0.768	**0.279**
GRU-32	bert-base-multilingual-cased	0.729	0.215
GRU-64	flaubert_base_cased	0.728	0.215
GRU-64	xlm-mlm-17-1280	0.704	0.176
GRU-64	fastText-cc.fr.300.bin	0.761	**0.29**
GRU-64	Memolon	0.77	0.254
GRU-64	Tf-idf	**0.774**	0.276
GRU-64	bert-base-multilingual-cased	0.698	0.205
GRU-128	flaubert_base_cased	0.736	0.196
GRU-128	xlm-mlm-17-1280	0.728	0.193

GRU-128	fastText-cc.fr.300.bin	0.765	**0.287**
GRU-128	Memolon	**0.77**	0.263
GRU-128	Tf-idf	0.764	0.277
GRU-128	bert-base-multilingual-cased	0.723	0.243

2. Résultats en RMSE

Tableau 9 présente les résultats mesurés par l'erreur quadratique moyenne sur le test set de RECOLA.

Tableau 9 Résultats de Reconnaissance des émotions de la parole en temps continu (erreur quadratique moyenne) sur test set de RECOLA

Corpus		RECOLA	
Modèle	Représentations	Arousal	Valence
Linear-Tanh	flaubert_base_cased	0.0087	0.051
Linear-Tanh	xlm-mlm-17-1280	0.041	0.032
Linear-Tanh	fastText-cc.fr.300.bin	**0.0021**	**0.0006**
Linear-Tanh	Memolon	0.0024	0.0009
Linear-Tanh	Tf-idf	0.0085	0.0012
Linear-Tanh	bert-base-multilingual-cased	0.0028	0.0026
GRU-32	flaubert_base_cased	0.0003	0.0008
GRU-32	xlm-mlm-17-1280	0.0004	0.0005
GRU-32	fastText-cc.fr.300.bin	**0.0001**	0.0005
GRU-32	Memolon	**0.0001**	0.0004
GRU-32	Tf-idf	**0.0001**	**0.0003**
GRU-32	bert-base-multilingual-cased	0.0002	**0.0003**
GRU-64	flaubert_base_cased	0.0002	0.0005
GRU-64	xlm-mlm-17-1280	0.0003	0.0005
GRU-64	fastText-cc.fr.300.bin	**0.0001**	**0.0003**
GRU-64	Memolon	**0.0001**	0.0004
GRU-64	Tf-idf	**0.0001**	0.0004
GRU-64	bert-base-multilingual-cased	0.0003	0.0004
GRU-128	flaubert_base_cased	**0.0001**	0.0004
GRU-128	xlm-mlm-17-1280	**0.0001**	**0.0003**
GRU-128	fastText-cc.fr.300.bin	**0.0001**	0.0004
GRU-128	Memolon	**0.0001**	0.0004
GRU-128	Tf-idf	0.0002	0.277
GRU-128	bert-base-multilingual-cased	0.0002	0.0004

3. Analyses

3.1 En général

Selon Tableau 8 et Tableau 9, nous constatons que, en général, les résultats de GRUs surpassent ceux de Linear-Tanh. En considérant que GRU utilisant des réseaux neuronaux récurrents combinés avec des portes de mise à jour et des portes de réinitialisation, est plus performant que Linear-Tanh étant le simple réseau feed-forward, on peut déduire que ces résultats attendus.

Globalement, les prédictions sur la dimension arousal sont meilleures que celles sur la dimension valence. Cela s'explique peut-être par le fait que les émotions passives et actives sont plus faciles à distinguer que les émotions négatives et positives. Nous avons observé cette différence dans d'autres articles, donc cela s'agit un problème général qui devra être étudié dans le domaine de la reconnaissance des émotions.

Toutes les représentations arrivent à faire des bonnes prédictions, ce qui prouve qu'utiliser les transcriptions pour faire des prédictions sur des émotions de la parole en temps continu est une approche intéressante. Les performances varient notamment en fonction des modèles, et des représentations différentes du même modèle produisent des résultats proches. Parmi des représentations auto-supervisées, les représentations de fastText sont légèrement supérieures que les autres en termes de précision des prédictions.

3.2 Sur les résultats Linear-Tanh

Selon Tableau 8, les représentations de BERT Multilingue produisent des meilleurs résultats dans la dimension arousal parmi tous les résultats de Linear-Tanh, tandis que les représentations FastText surpassent les autres représentations dans la dimension valence. Sur la dimension arousal, les représentations les moins performantes sont celles de XLM et Tf-idf. Les représentations les moins performantes sont celles de FlauBERT, sur la dimension valence. Une comparaison des tableaux 8 et 9 montre que les résultats sont légèrement différents en raison des différentes métriques d'évaluation, mais les deux résultats restent cohérents.

3.3 Sur les résultats GRU

Nous constatons que le nombre des couches cachées de GRU exerce une influence importante sur les résultats. Plus le nombre de couches cachées est élevé, meilleures sont les performances. Du côté des prédictions sur la dimension arousal, selon Tableau 8, les

représentations Memolon légèrement surpassent les autres représentations en GRU-32 et GRU-128. Du côté des prédictions sur la dimension valence, les représentations fastText ont obtenu les meilleurs résultats en GRU-64 et GRU-128.

3.4 Sur les résultats des représentations auto-supervisées

A travers le Tableau 8 et 9, nous constatons que les représentations FastText ont les meilleures performances parmi toutes les représentations auto-supervisées, même si la différence est mineure et cette différence diminue lorsque la complexité du modèle augmente. En gros, toutes les représentations auto-supervisées peuvent faire des plutôt bonnes prédictions dans la tâche de reconnaissance des émotions de la parole en temps continu.

3.5 Les meilleures résultats

Parmi tous nos expériences, le meilleur résultat pour la dimension arousal est obtenu par les représentations Tf-idf utilisant le modèle GRU-64. Le meilleur résultat pour la dimension valence est produit par les représentations FastText utilisant le modèle GRU-64.

Nous avons visualisé les corrélations entre les annotations et les sorties de nos meilleures combinaisons (modèle + représentation) pour la dimension arousal et valence dans le Tableau 10 et 11.

Tableau 10 Comparaison des annotations et des sorties du modèle GRU-64 dans la dimension arousal en utilisant les représentations Tf-idf sur les données de test_01

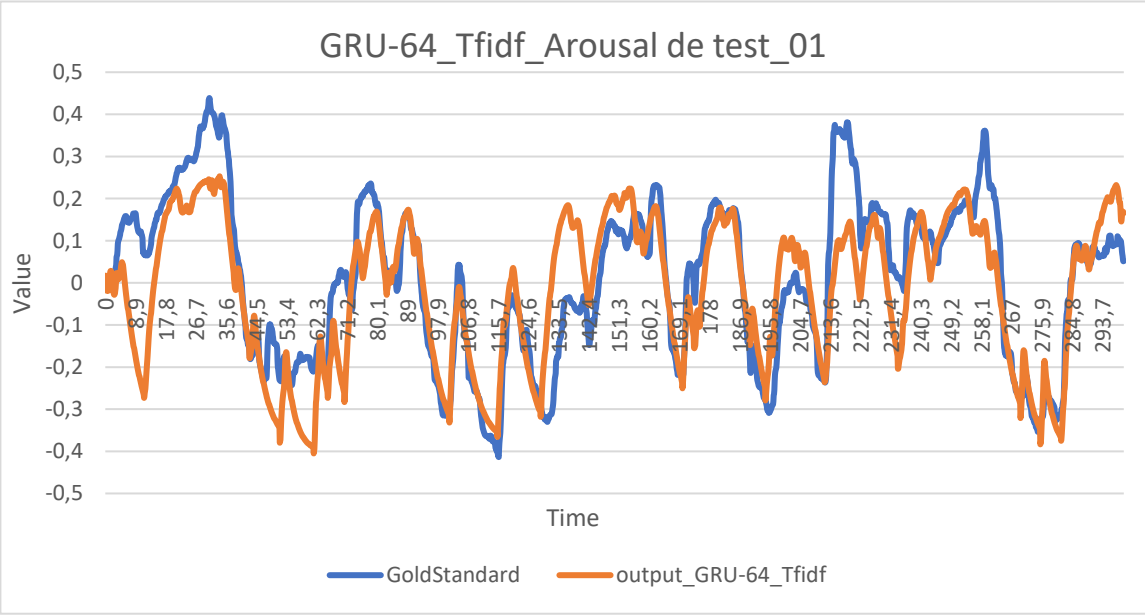
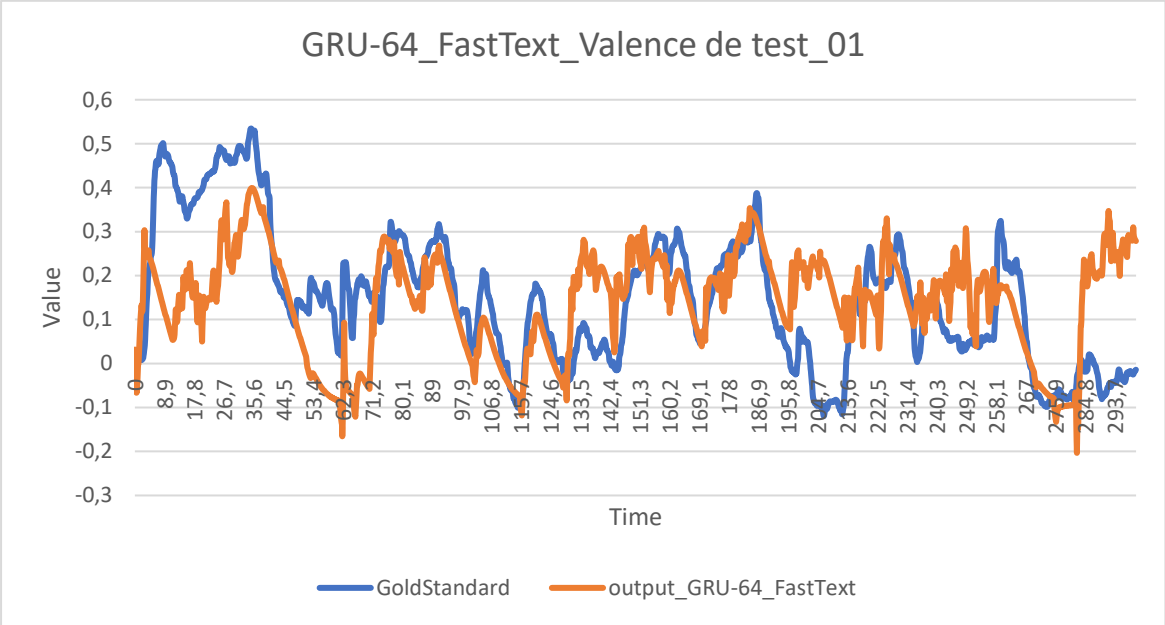


Tableau 11 Comparaison des annotations et des sorties du modèle GRU-64 dans la dimension valence en utilisant les représentations FastText sur les données de test_01



Chapitre 13. Comparaison et Analyses

Les comparaisons permettent d'avoir un regard critique sur les résultats. Nous proposons trois comparaisons sur les résultats pour détailler nos analyses.

1. Acoustique et textuelle

Ce mémoire est une extension de la section « reconnaissance automatique des émotions » de (Evain et al., 2021). Les deux expérimentations partagent partiellement les même modèles, corpus et métrique d'évaluation, il nous semble donc intéressant de comparer les performances des représentations auto-supervisées acoustiques et textuelles dans la tâche REP en temps continu. Le Tableau 12 présente les comparaisons des deux types de représentations en fonction des modèles. Nous constatons que les résultats des représentations textuelles sont proches de ceux des représentations acoustiques, ce qui indique que les transcriptions sont utiles dans la tâche de REP en temps continu.

Tableau 12 Comparaison des résultats de CCC des représentations acoustiques et textuelles

Corpus			RECOLA	
Modèle	Représentations		Arousal	Valence
Linear-Tanh	Textuelles	flaubert_base_cased	0.222	0.015
		xlm-mlm-17-1280	0.148	0.025
		fastText-cc.fr.300.bin	0.278	**0.05**
		Memolon	0.251	0.041
		Tf-idf	0.154	0.043
		bert-base-multilingual-cased	**0.287**	0.018
	Acoustiques	MFB	0.192	0.075
		W2V2-Fr-M-base	**0.385**	**0.090**
		XLSR-53-large	0.155	0.024
GRU-32	Textuelles	flaubert_base_cased	0.687	0.163
		xlm-mlm-17-1280	0.659	0.139
		fastText-cc.fr.300.bin	0.763	0.255
		Memolon	**0.772**	0.233
		Tf-idf	0.768	**0.279**
		bert-base-multilingual-cased	0.729	0.215
	Acoustiques	MFB	0.664	0.252
		W2V2-Fr-M-base	**0.767**	**0.376**
		XLSR-53-large	0.605	0.320
GRU-64	Textuelles	flaubert_base_cased	0.728	0.215

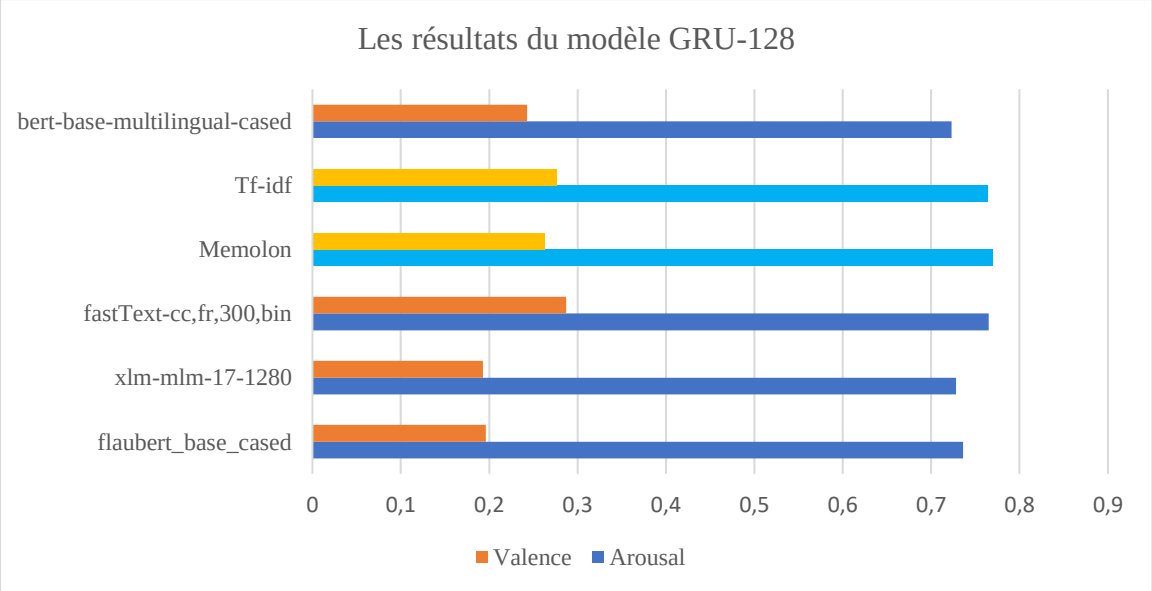
		xlm-mlm-17-1280	0.704	0.176
		fastText-cc.fr.300.bin	0.761	**0.29**
		Memolon	0.77	0.254
		Tf-idf	**0.774**	0.276
		bert-base-multilingual-cased	0.698	0.205
	Acoustiques	MFB	0.712	0.307
		W2V2-Fr-M-base	**0.760**	**0.352**
		XLSR-53-large	0.585	0.280

2. Auto-supervisée et autres approches

Selon les résultats de nos expérimentations montrés dans les Tableau 8 et 9, nous constatons que les représentations auto-supervisées n'atteignent pas les meilleurs résultats, au contraire, le lexique relatif aux émotions Memolon a obtenu les meilleurs résultats dans les modèles GRU-32 et GRU-128.

En prenant les résultats du modèle GRU-128 comme un exemple, Tableau 13 montre les résultats du modèle du modèle GRU-128 (les résultats de Tf-idf et Memolon sur la dimension arousal sont marqué en bleu clair). Nous pouvons observer que les écarts entre les représentations auto-supervisées et les autres représentations telles que Memolon et Tf-idf, ne sont pas importants. Pour ce faire, on peut conclure que dans la tâche REP, les représentations auto-supervisées textuelles ne disposent pas d'avantages, en comparant avec les représentations obtenues par Tf-idf ou le lexique relatif aux émotions.

Tableau 13 Les résultats du modèle GRU-128



Le succès des représentations Memolon prouve que le lexique relatif aux émotions est une approche solide pour la reconnaissance des émotions. Toutes les informations d'une séquence donnée n'ont pas la même importance pour la reconnaissance des émotions. Comme Tf-idf mesure l'importance des mots d'une séquence en considérant sa fréquence, il nous semble raisonnable que Tf-idf obtienne de bons résultats dans la reconnaissance des émotions.

3. CCC et RMSE

En observant les Tableau 8 et 9, on remarque les différences entre les résultats en CCC et les résultats en RMSE. Comme CCC et RMSE évalue les sorties en prenant compte de différents critères, il nous semble raisonnable que la combinaison (modèle + représentation) obtenant les meilleurs résultats en CCC ne correspond pas celle de RMSE. Nous trouvons que RMSE est moins précis en termes de l'évaluation des performances, car RMSE calcule les sorties de plusieurs approches mais donne les résultats identiques. Par exemple, sur les résultats des modèles GRU, quand les écarts entre les cibles et les références sont mineurs, RMSE n'arrivent pas à montrer les différences entre les différentes représentations.

Malgré des différences, CCC et RMSE sont tous efficaces pour évaluer les sorties et les résultats de deux évaluations restent généralement cohérents.

4. Modèle Multilingue et Modèle en français

Nous avons utilisé différents modèles auto-supervisés pré-entraînés pour extraire des représentations. XLM et BERT Multilingue sont les modèles multilingues ayant entrainés sur des données de plusieurs langues, tandis que FlauBERT et FastText-cc.fr.300.bin sont entrainés sur les données monolingues en français. Tableau 14 et 15 montrent les comparaisons des résultats des modèles multilingues et des modèles spécialisés en français. Les résultats des modèles multilingues sont marqués en gris.

Nous pouvons constater que, dans toutes les dimensions, les modèles du français légèrement surpassent les modèles multilingues. Tous les deux modèles en français parviennent à produire meilleurs résultats que les modèles multilingues. Ce fait révèle les avantages des modèles spécialement entraînés sur la langue cible lors des traitements automatiques de cette langue.

Tableau 14 Comparaison des résultats CCC dans la dimension arousal pour les modèles multilingues et les modèles en français

	xlm-mlm-17-1280	bert-base-multilingual-cased	flaubert_base_cased	fastText-cc.fr.300.bin
Linear-Tanh	0.148	0.287	0.222	0.278
GRU-32	0.659	0.729	0.687	0.763
GRU-64	0.704	0.698	0.728	0.761
GRU-128	0.728	0.723	0.736	0.765

Tableau 15 Comparaison des résultats CCC dans la dimension Valence pour les modèles multilingues et les modèles en français

	xlm-mlm-17-1280	bert-base-multilingual-cased	flaubert_base_cased	fastText-cc.fr.300.bin
Linear-Tanh	0.025	0.018	0.015	0.05
GRU-32	0.139	0.215	0.163	0.255
GRU-64	0.176	0.205	0.215	0.29
GRU-128	0.243	0.243	0.196	0.287

Partie 5

-

Conclusion et Perspectives

Chapitre 14. Conclusion

Pour conclure, dans cette étude nous avons abordé différentes approches de la modélisation auto-supervisée sur les transcriptions en français pour la tâche de reconnaissance des émotions spontanées de la parole en temps continu. Les résultats de nos expérimentations montrent que les représentations textuelles des transcriptions obtiennent les résultats comparables à ceux de représentations acoustiques de la parole brute, ce qui prouve que les transcriptions sont utiles et efficaces pour la reconnaissance des émotions de la parole. D'après les résultats obtenus, les représentations auto-supervisées ont atteint les performances de haut niveau et elles ont aboutis à produire les bons résultats même si la taille du corpus est limitée et les modèles utilisés disposent d'une architecture NN simple. En plus, les modèles auto-supervisés ayant spécialement entrainés pour le français surpassent les modèles multilingues en termes de performances dans la tâche mentionné.

Les contributions de ce mémoire sont : 1. Examiner une méthode de l'alignement pour accorder les représentations textuelles avec le temps. 2. Explorer les différentes approches de modélisation auto-supervisées pour les transcriptions de la parole. 3. Comparer les performances de différentes représentations auto-supervisées dans la tâche de la reconnaissance des émotions de la parole en temps continu.

Les résultats de ce mémoire prouvent que les représentations auto-supervisées des transcriptions sont favorables pour l'amélioration des performances des systèmes de reconnaissance des émotions de la parole. Nous pouvons donc déduire que la combinaison des représentations auto-supervisées venant de la parole et des textes obtiendrait probablement d'excellents résultats dans la tâche REP.

Chapitre 15. Discussion

En comparant avec les performances des représentations tf-idf ou Memolon (le lexique relatif aux émotions), les représentations auto-supervisées n'ont pas pu apporter de meilleurs résultats. Cela est peut être lié au fait que les modèles auto-supervisés qu'on a utilisé pour extraire les représentations ne sont pas ajustés pour cette tâche spécifique. Ajouter une étape de fine-tuning permettrait peut-être d'améliorer les performances de ces représentations auto-supervisées.

Les méthodes de tokenisation influencent peut-être les résultats de prédiction comme les représentations sont centrées sur des tokens. Nous n'avons pas pu unifier les méthodes de tokenisation, si nous affinons les processus de la normalisation et unifier les méthodes de tokenisation, nous obtiendrions peut-être de meilleurs résultats.

Pendant l'expérimentation, nous avons accidentellement utiliser BERT original qui est entraîné sur les données de l'anglais. Cette erreur nous permet de comparer les performances de BERT Multilingue et BERT original. Tableau 16 et 17 montrent ces comparaisons. Nous sommes étonnés que les deux modèles apportent des résultats similaires et BERT surpassent BERT Multilingue dans certaines expériences. Cette similarité est peut-être due au fait que l'anglais et le français ont beaucoup de point de communs et que l'anglais est une langue germanique avec une influence française et latine. Cette observation aussi explique les différences mineures entre les modèles multilingues et les modèles du français.

Tableau 16 Résultats CCC des modèles BERT sur Arousal

	bert-base-multilingual-cased	bert-base-cased
Linear-Tanh	0.287	0.279
GRU-32	0.729	0.731
GRU-64	0.698	0.732
GRU-128	0.723	0.747

Tableau 17 Résultats CCC des modèles BERT sur Valence

	bert-base-multilingual-cased	bert-base-cased
Linear-Tanh	0.018	-0.0004
GRU-32	0.215	0.252
GRU-64	0.205	0.213
GRU-128	0.243	0.257

Chapitre 16. Perspectives

La combinaison des représentations auto-supervisées textuelles et des représentations auto-supervisées acoustiques nous intéresse beaucoup. Nous voulons explorer les méthodes pour combiner ces deux représentations et évaluer ses performances dans la tâche de la reconnaissance des émotions de la parole. Dans cette section, nous discutons quelques approches pour réaliser cette combinaison pour la tâche de reconnaissance des émotions.

1. Mise en commun (Pooling)

Les couches « mise en commun » sont souvent utilisées pour combiner les représentations de différentes modalités, les méthodes traditionnelles incluent la concaténation, l'addition par éléments (element-wise addition) et la multiplication par éléments (element-wise multiplication). L'étude de (Macary et al., 2021) montre que la concaténation des représentations de différentes modalités n'arrive pas à produire des bons résultats dans la tâche de reconnaissance des émotions. Pour faire l'addition ou la multiplication par élément, il faut que les représentations de deux modalités aient les mêmes dimensions mais il est difficile d'avoir deux représentations de la même dimension.

(Aldeneh et al., 2017) a comparé les méthodes traditionnelles avec la méthode « outer-product » et leur résultat prouve que « outer-product pooling » est l'opération la plus efficace pour combiner les représentations linguistiques et les représentations acoustiques sur la prédiction de valence. « outer-product pooling » n'exige pas que les deux représentations combinées disposent de la même dimension. Il nous semble intéressant d'explorer cette approche pour la prédiction des émotions dans la dimension arousal.

2. Co-attention Fusion

La co-attention fusion est une nouvelle approche basée sur le mécanisme d'attention et sur le modèle transformer multimodal, initialement proposé par (Tsai et al., 2019) et exploré par (Khare et al., 2021). Le principe de cette approche est d'échanger la paire clé-valeur entre des modalités pour que le modèle apprenne les interactions entre des différents modalités.

3. Fusion superficielle pour des modèles de type BERT

Cette nouvelle stratégie de fusion est proposée par (Siriwardhana et al., 2020) qui a exploré l'utilisation des modèles pré-entrainés du type BERT pour obtenir des représentations acoustiques et textuelles pour la reconnaissance des émotions de la parole. Les modèles du type BERT disposent des tokens spéciaux [CLS] qui peut être utilisé comme une représentation de la phrase entière. [CLS] signifie la classification et il est le première token de chaque séquence d'entrée. Cette approche « fusion superficielle » s'agit de concaténer les représentations des tokens spéciaux [CLS] de deux modalités. (Siriwardhana et al., 2020) prouve que cette fusion superficielle est efficace et il peut produire des résultats de l'état de l'art dans le domaine de reconnaissance des émotions de la parole.

Bibliographies

Aldeneh, Z., Khorram, S., Dimitriadis, D., & Provost, E. M. (2017). Pooling acoustic and lexical features for the prediction of valence. *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 68-72.

Bachman, P., Hjelm, R. D., & Buchwalter, W. (2019). Learning Representations by Maximizing Mutual Information Across Views. *arXiv:1906.00910 [cs, stat]*. http://arxiv.org/abs/1906.00910

Baevski, A., Auli, M., & Mohamed, A. (2020). Effectiveness of self-supervised pre-training for speech recognition. *arXiv:1911.03912 [cs]*. http://arxiv.org/abs/1911.03912

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135-146.

Buechel, S., Rücker, S., & Hahn, U. (2020). Learning and evaluating emotion lexicons for 91 languages. *arXiv preprint arXiv:2005.05672*.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A Simple Framework for Contrastive Learning of Visual Representations. *International Conference on Machine Learning*, 1597-1607. http://proceedings.mlr.press/v119/chen20j.html

Chung, Y.-A., & Glass, J. (2018). Speech2vec : A sequence-to-sequence framework for learning word embeddings from speech. *arXiv preprint arXiv:1803.08976*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. http://arxiv.org/abs/1810.04805

Evain, S., Nguyen, H., Le, H., Boito, M. Z., Mdhaffar, S., Alisamir, S., Tong, Z., Tomashenko, N., Dinarelli, M., Parcollet, T., Allauzen, A., Esteve, Y., Lecouteux, B., Portet, F., Rossato, S., Ringeval, F., Schwab, D., & Besacier, L. (2021). LeBenchmark : A Reproducible Framework for Assessing Self-Supervised Representation Learning from Speech. *arXiv:2104.11462 [cs, eess]*. http://arxiv.org/abs/2104.11462

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning Word Vectors for 157 Languages. *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., & Bengio, Y. (2019). Learning deep representations by mutual information estimation and maximization. *arXiv:1808.06670 [cs, stat]*. http://arxiv.org/abs/1808.06670

Kawakami, K., Wang, L., Dyer, C., Blunsom, P., & Oord, A. van den. (2020). Learning Robust and Multilingual Speech Representations. *arXiv:2001.11128 [cs, eess]*. http://arxiv.org/abs/2001.11128

Khare, A., Parthasarathy, S., & Sundaram, S. (2021). Self-Supervised learning with cross-modal transformers for emotion recognition. *2021 IEEE Spoken Language Technology Workshop (SLT)*, 381-388.

Kingma, D. P., & Ba, J. (2014). Adam : A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lample, G., & Conneau, A. (2019). Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., & Schwab, D. (2020). FlauBERT: des modèles de langue contextualisés pré-entraînés pour le français. *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2: Traitement Automatique des Langues Naturelles*, 268-278.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. *International conference on machine learning*, 1188-1196.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). Bart : Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

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 preprint arXiv:1907.11692*.

Macary, M., Tahon, M., Estève, Y., & Rousseau, A. (2021). On the use of Self-supervised Pre-trained Acoustic and Linguistic Features for Continuous Speech Emotion Recognition. *2021 IEEE Spoken Language Technology Workshop (SLT)*, 373-380.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Montacié, C., & Caraty, M.-J. (2018). Vocalic, Lexical and Prosodic Cues for the INTERSPEECH 2018 Self-Assessed Affect Challenge. *INTERSPEECH*, 541-545.

Oord, A. van den, Li, Y., & Vinyals, O. (2019). Representation Learning with Contrastive Predictive Coding. *arXiv:1807.03748 [cs, stat]*. http://arxiv.org/abs/1807.03748

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove : Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532-1543.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is Multilingual BERT? *arXiv:1906.01502 [cs]*. http://arxiv.org/abs/1906.01502

Ringeval, F., Sonderegger, A., Sauer, J., & Lalanne, D. (2013). Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 1-8. https://doi.org/10.1109/FG.2013.6553805

Rodrigues Makiuchi, M., Warnita, T., Uto, K., & Shinoda, K. (2019). Multimodal fusion of bert-cnn and gated cnn representations for depression detection. *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 55-63.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). wav2vec : Unsupervised Pre-training for Speech Recognition. *arXiv:1904.05862 [cs]*. http://arxiv.org/abs/1904.05862

Siriwardhana, S., Reis, A., Weerasekera, R., & Nanayakkara, S. (2020). Jointly Fine-Tuning" BERT-like" Self Supervised Models to Improve Multimodal Speech Emotion Recognition. *arXiv preprint arXiv:2008.06682*.

Soğancıoğlu, G., Verkholyak, O., Kaya, H., Fedotov, D., Cadée, T., Salah, A. A., & Karpov, A. (2020). Is Everything Fine, Grandma ? Acoustic and Linguistic Modeling for Robust Elderly Speech Emotion Recognition. *arXiv preprint arXiv:2009.03432*.

Su, Y., & Kuo, C.-C. J. (2019). On extended long short-term memory and dependent bidirectional recurrent neural network. *Neurocomputing, 356*, 151-161.

Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., & Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. *Proceedings of the conference. Association for Computational Linguistics. Meeting, 2019*, 6558.

Weninger, F., Ringeval, F., Marchi, E., & Schuller, B. W. (2016). Discriminatively Trained Recurrent Neural Networks for Continuous Dimensional Emotion Recognition from Audio. *IJCAI, 2016*, 2196-2202.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet : Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems, 32*.

Yeh, S.-L., Chao, G.-Y., Su, B.-H., Huang, Y.-L., Lin, M.-H., Tsai, Y.-C., Tai, Y.-W., Lu, Z.-C., Chen, C.-Y., Tai, T.-M., & others. (2019). Using Attention Networks and Adversarial Augmentation for Styrian Dialect Continuous Sleepiness and Baby Sound Recognition. *Interspeech*, 2398-2402.

Sitographie

Tous les documents et toutes les informations accessibles via les adresses URL regroupées sur cette page ont été consulté pour la dernière fois le 26/08/2021.

http://arxiv.org/abs/1906.00910
http://arxiv.org/abs/1911.03912
http://arxiv.org/abs/1810.04805
https://towardsdatascience.com/deep-learning-feedforward-neural-network-26a6705dbdc7
https://fr.wikipedia.org/wiki/Fonction_d%27activation
https://towardsdatascience.com/illustrated-guide-to-recurrent-neural-networks-79e5eb8049c9

Glossaire

Représentation : les informations extraites des données ayant des caractéristiques représentatives d'un phénomène, utilisées par les modèles de l'apprentissage automatique.

Token : une unité sémantique utile pour le traitement automatique de la langue.

Modélisation : le processus de l'extraction des représentations à partir des données

Apprentissage auto-supervisé : une approche de l'apprentissage automatique qui entraîne les réseaux neuronaux à effectuer des tâches auxiliaires où les entrées et les étiquettes sont dérivées d'un ensemble de données non étiquetées

Sigles et abréviations utilisés

TAL :	traitement automatique de la langue
MA :	modélisation auto-supervisée
REP :	reconnaissance des émotions de la parole
NN :	réseau neuronal
MLM :	modélisation de langage masqué

Table des Figures

Figure 1 Pourcentage de chaque approche dans les deux challenges AVEC et ComParE entre 2016 et 2020 17
Figure 2 Illustration de l'architecture de Wav2vec 2.0 20
Figure 3 Extrait des résultats du modèle BERT utilisant de différentes approches pour extraire des représentations dans la tâche CoNLL-2003 Reconnaissance d'entités nommées (Devlin et al., 2019) 29
Figure 4 Démonstration du processus d'rééchantillonnage des tokens 36
Figure 5 Vue d'ensemble de l'expérimentation 38
Figure 6 Le schéma d'une cellule de GRU 44

Table des Tableaux

Tableau 1 Les participants ayant obtenu les meilleurs résultats en utilisant l'approche non-supervisée ou la fusion avec les informations linguistiques dans les tâches affectives d'AVEC et de ComParE de 2016 à 2020 16
Tableau 2 Nombre d'études concernant l'informatique affective de chaque approche dans challenge AVEC et ComParE entre 2016 et 2020 18
Tableau 3 Résultats mesurés en coefficient corrélation concordance de la reconnaissance automatique des émotions sur le test set de RECOLA et AlloSat 22
Tableau 4 Résumé des modèles apprentissage auto-supervisé pour des représentations de textes .. 23
Tableau 5 Résumé des approches et la taille des vecteurs de représentations obtenues 28
Tableau 6 Evolution de la perte sur la partition validation du modèle GRU-128 sur la dimension arousal 46
Tableau 7 Evolution de la perte sur la partition validation du modèle Linear-tanh sur la dimension arousal 47
Tableau 8 Résultats de Reconnaissance des émotions de la parole en temps continu (Coefficient Corrélation Concordance) sur test set de RECOLA 49
Tableau 9 Résultats de Reconnaissance des émotions de la parole en temps continu (erreur quadratique moyenne) sur test set de RECOLA 50
Tableau 10 Comparaison des annotations et des sorties du modèle GRU-64 dans la dimension arousal en utilisant les représentations Tf-idf sur les données de test_01 52
Tableau 11 Comparaison des annotations et des sorties du modèle GRU-64 dans la dimension valence en utilisant les représentations FastText sur les données de test_01 53
Tableau 12 Comparaison des résultats de CCC des représentations acoustiques et textuelles 54
Tableau 13 Les résultats du modèle GRU-128 55
Tableau 14 Comparaison des résultats CCC dans la dimension arousal pour les modèles multilingues et les modèles en français 57
Tableau 15 Comparaison des résultats CCC dans la dimension Valence pour les modèles multilingues et les modèles en français 57
Tableau 16 Résultats CCC des modèles BERT sur Arousal 60
Tableau 17 Résultats CCC des modèles BERT sur Valence 60

Table des annexes

Annexe 1 Tableau de la comparaison des approches utilisées entre 2016- 2020 dans challenges ComParE et AVEC 72

Annexe 1
Tableau de la comparaison des approches utilisées entre 2016- 2020 dans challenges ComParE et AVEC

Challenge	Task	Metric	Reference	Hand-crafted	End-to-end	Transfer Learning	Semi-Supervised	Unsupervised	Fusion with Linguistics
ComParE 2020	Elderly Emotion (Arousal)	UAR	Baseline	49.1		50.4		44.3	44.0
			Participants	54.3					63.7
	Elderly Emotion (Valence)	UAR	Baseline	41.7		40.3		33.8	49.0
			Participants	59.0					57.5
AVEC 2019	Depression Detection with AI	CCC	Baseline	.045		.108			
			Participants				.430		.403
ComParE 2019	Continuous Sleepiness	PC	Baseline	.314				.325	
			Participants	.383	.335				
	Baby Sound	UAR	Baseline	57.7				48.1	
			Participants	59.5				62.4	
AVEC 2018	Cross-cultural Emotion Recognition (Arousal)	CCC	Baseline	.236					
			Participants		.377				
	Cross-cultural Emotion Recognition (Valence)	CCC	Baseline	.217					
			Participants		.389				
	Gold-standard Emotion (Arousal)	CCC	Baseline	.651		.495			
			Participants						

	Gold-standard Emotion (Valence)	CCC	Baseline	.346		.158			
			Participants						
ComParE 2018	Atypical Affect	UAR	Baseline	43.1	28.0			35.6	
			Participants	41.1	47.8				
	Self-Assessed Affect	UAR	Baseline	65.2	46.6			57.3	
			Participants	67.0	48.3		48.9		68.4
	(Infant) Crying	UAR	Baseline	73.2	63.5			71.1	
			Participants	70.1					
ComParE 2017	Cold	UAR	Baseline	70.2	60.0		64.8		
			Participants	72.0	71.2				
ComParE 2016	Deception	UAR	Baseline	68.3					
			Participants	72.1		50.7	72.2		

MOTS-CLÉS : Reconnaissance des émotions de la parole, Apprentissage auto-supervisé, Représentation auto-supervisée, apprentissage des représentations

RÉSUMÉ

Les modèles auto-supervisés pré-entraîné utilisant des données non étiquetées pour extraire des représentations ont été largement exploré dans le domaine du traitement automatique de la parole. Ce mémoire explore une nouvelle approche consistant à extraire des représentations linguistiques des transcriptions à partir des modèles auto-supervisés pré-entraîné pour la reconnaissance des émotions de la parole spontanée en temps continu. Nous avons examiné une méthode d'alignement pour accorder les représentations linguistiques avec du temps. Les résultats des expérimentations montrent que les représentations auto-supervisées linguistiques peuvent prédire les émotions en dimension arousal et valence aussi bien que les représentations auto-supervisées acoustiques.

KEYWORDS : Speech Emotion Recognition, Self-Supervised Representation Learning

ABSTRACT

Self-supervised pre-trained models using unlabeled data to extract representations have been widely explored in the field of automatic speech processing. This paper explores a new approach of extracting linguistic representations from transcriptions using pre-trained self-supervised models for time-continuous emotion recognition. We examined an alignment method to fit linguistic representations with time. Experimental results show that linguistic self-supervised representations can predict emotions in arousal and valence dimensions and achieve excellent results as acoustic self-supervised representations.