LeBenchmark, un référentiel d'évaluation pour le français oral *

Hang Le, Sina Alisamir, Marco Dinarelli, Fabien Ringeval, Solène Evain, Ha Nguyen, Marcely Zanon Boito, Salima Mdhaffar, Ziyi Tong, Natalia Tomashenko, et al.

LeBenchmark, un référentiel d'évaluation pour le français oral *

Hang Le[1] Sina Alisamir[1,2] Marco Dinarelli[1] Fabien Ringeval[1] Solène Evain[1] Ha Nguyen[1, 4]
Marcely Zanon Boito[4] Salima Mdhaffar[4] Ziyi Tong[1] Natalia Tomashenko[4] Titouan Parcollet[4]
Alexandre Allauzen[5] Yannick Estève[4] Benjamin Lecouteux[1] François Portet[1] Solange Rossato[1]
Didier Schwab[1] Laurent Besacier[3]

(1) Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000 Grenoble, France
(2) Atos, Échirolles, France (3) Naver Labs Europe, France
(4) LIA, Avignon Université (5) ESPCI, CNRS LAMSADE, PSL Research University, France
http://www.lebenchmark.com ; mailto:lebenchmark@univ-grenoble-alpes.fr

RÉSUMÉ

L'apprentissage autosupervisé a apporté des améliorations remarquables dans de nombreux domaines tels que la vision par ordinateur ou le traitement de la langue et de la parole, en exploitant de grandes quantités de données non étiquetées. Dans le contexte spécifique de la parole, cependant, et malgré des résultats prometteurs, il existe un manque évident de normalisation dans les processus d'évaluation permettant des comparaisons précises de ces modèles, en particulier pour les autres langues que l'anglais. Nous présentons ici à la communauté francophone *LeBenchmark*, un cadre de référence en sources ouvertes et reproductible pour évaluer des modèles autosupervisés à partir de corpus de parole en français. Il est composé de quatre tâches : reconnaissance automatique de la parole, compréhension du langage parlé, traduction automatique de la parole et reconnaissance automatique d'émotions. Nous encourageons la communauté francophone à utiliser ce référentiel dans ses futures expérimentations, notamment pour l'évaluation de modèles autosupervisés.

ABSTRACT

***LeBenchmark*, an Evaluation Framework for French Speech**

Self-supervised learning (SSL) has made remarkable improvements in many areas, including computer vision or language and speech processing, by exploiting large amounts of unlabelled data. In the specific context of speech, however, and despite promising results, there is a clear lack of standardisation in the evaluation process for full comparisons of these models especially for languages other than English. We present here to the French-speaking community *LeBenchmark*, an open-source and reproducible framework for evaluating SSLs from French speech data. It is composed of four tasks : automatic speech recognition, spoken language understanding, automatic speech translation and automatic emotion recognition. We encourage the French-speaking community to use this benchmark in future experiments.

MOTS-CLÉS : Référentiel d'évaluation, Modèles Autosupervisés, Reconnaissance Automatique de la Parole, Compréhension Automatique de la Parole, Traduction Automatique de la Parole, Reconnaissance Automatique d'Émotions.

KEYWORDS: Evaluation Framework, Automatic Speech Recognition, Self-Supervised Learning, Speech Language Understanding, Speech Translation, Automatic Emotion Recognition.

∗. note aux relecteurs : cette soumission est une sous-partie réadaptée d'un article publié à Neurips fin 2021. Nous avons tenu à faire cet effort afin de présenter ces travaux concernant un référentiel d'évaluation pour le traitement du français à la communauté francophone. Il ne s'agit pas d'une simple traduction : certains détails et informations sont précisés ici qui étaient absents dans l'article original.

1 Introduction

L'apprentissage autosupervisé permet d'exploiter d'énormes quantités de données non étiquetées, et a été exploré avec succès pour le traitement des images et du langage naturel (Bachman *et al.*, 2019; Chen *et al.*, 2020; Devlin *et al.*, 2018; Raffel *et al.*, 2019). Les performances obtenues dans des tâches classiques du traitement automatique de la parole ont également été améliorées par les approches autosupervisées (Baevski *et al.*, 2019; Kawakami *et al.*, 2020). Cependant, le développement accéléré des modèles autosupervisés de la parole est bien souvent accompagné d'évaluations réalisées sur différents corpus de données, dont la plupart en langue anglaise. Afin de quantifier et comparer l'impact réel de ces approches autosupervisées, il est important de réaliser leur évaluation dans un cadre rigoureux, en s'appuyant sur des référentiels d'évaluation communs et facilement utilisables. Alors que ces référentiels sont aujourd'hui largement adoptés pour l'écrit (Ruder, 2021), y compris en français (Le *et al.*, 2020), ils sont nettement moins courants dans le domaine de la parole, domaine qui a pourtant une longue tradition d'évaluation comme le prouvent les tâches partagées du NIST et du DARPA pour la reconnaissance de la parole. Ainsi, nous n'avons connaissance que de 2 initiatives similaires pour l'évaluation des modèles autosupervisés de la parole : le Speech processing Universal PERformance Benchmark (SUPERB) (Yang *et al.*, 2021) ainsi que le récent SLUE (Shon *et al.*, 2021) (Spoken Language Understanding Evaluation) qui ne visent toutefois que la langue anglaise.

Contribution. Cet article présente notre référentiel d'évaluation pour les modèles autosupervisés de parole en français [1]. Ce référentiel est disponible et est accompagné d'un classement de modèles sur le site Web suivant : `http://www.lebenchmark.com`. Nous proposons un ensemble destiné à évoluer, aujourd'hui composé de quatre tâches principales et dix sous-tâches utilisant la parole en langue française : la reconnaissance automatique de la parole (RAP), la compréhension de la langue parlée (CLP), la traduction de la parole (TAP) et la reconnaissance d'émotions (RAE).

Nous expliquons les principes que nous avons suivi pour mettre en place ces tâches d'évaluation et les illustrons grâce à quelques modèles.

2 Évaluation de Modèles Autosupervisés dans le Benchmark

2.1 Principe général

Nous évaluons les modèles sur quatre tâches : 1) la reconnaissance automatique de la parole (RAP - parole vers texte), tâche qui consiste à transcrire automatiquement de l'oral ; 2) la compréhension du langage parlé (CLP), tâche qui consiste à identifier certains éléments du discours dans des domaines précis (ex : action à effectuer, destination, type de restaurant…) ; 3) la reconnaissance automatique d'émotions (RAE) qui consiste à identifier dans notre cas des dimensions affectives telles que l'éveil et le plaisir intrinsèque (valence) ; 4) la traduction automatique multilingue (parole vers texte – TAP), tâche qui consiste à traduire automatiquement de l'oral vers le texte correspondant dans plusieurs langues.

Ces tâches ont été choisies en fonction des critères suivants : (a) diversité des problèmes : régression (RAP), étiquetage de séquences (CLP) et génération conditionnelle de langage naturel (RAP, TAP),

1. Les modèles eux-mêmes sont présentés dans un article compagnon également soumis aux JEPs 2022.

(b) diversité des informations extraites : transcription (RAP), sémantique (CLP), traduction (TAP) et paralinguistique (RAE), et (c) diversité des ressources annotées disponibles pour la tâche : grande (RAP), moyenne (CLP, TAP), petite (RAE).

Notre objectif est d'évaluer l'impact de l'apprentissage autosupervisé pour les meilleures systèmes de base de chaque tâche abordée. Nous avons ainsi une architecture différente pour chaque tâche, architecture qui correspond à la meilleure performance que nous obtenons en utilisant les caractéristiques MFCC/FBANK. De même, nous évaluons les différents modèles autosupervisés comme extracteurs de caractéristiques pour chacune de ces tâches. Ces « extracteurs autosupervisés » sont cependant soit « agnostiques », soit « spécifiques » (modèles autosupervisés affinés sur les données de la tâche), comme nous l'expliquons pour chacune des tâches.

À des fins d'illustration, nous présentons les résultats pour chacune des tâches sur les références de base (MFCC, MFB, spectogrammes) ainsi que sur les modèles suivants :

— XLSR-53-large qui est un modèle multilingue de 53 langues incluant du français [2].
— Fr-3K-large est un modèle de type Wav2Vec 2.0 appris sur 2 933 h de parole en français avec une architecture de 24 couches transformeur, une dimension du modèle de 1 024, une dimension interne de 4 096, 16 têtes d'attention et 500 000 mises à jour.
— Fr-7K-large est un modèle d'architecture et du nombre de mises à jour identiques au précédant mais appris sur 7 739 h de français.

2.2 Reconnaissance automatique de la parole (RAP)

La RAP est évaluée en utilisant à la fois des approches hybrides DNN-HMM et des approches de bout en bout. En plus du code source utilisé pour réaliser ces expériences (entraînement + décodage), *LeBenchmark* fournit un script de normalisation pour le texte de sortie dérivé de celui appliqué lors des campagnes d'évaluation officielles françaises ESTER et ETAPE (Gravier *et al.*, 2012) ainsi qu'un script unique pour calculer le taux d'erreur sur les mots (WER) à partir de la sortie RAP.

Datasets Les tâches RAP ciblent deux types de corpus différents : Common Voice (Ardila *et al.*, 2020) et ETAPE (Gravier *et al.*, 2012). Common Voice est un très grand corpus (477 h) de parole lue en français incluant également des transcriptions – train : 428 h ; dev : 24 h ; test : 25 h – tandis qu'ETAPE est un corpus plus petit (36 h), mais plus ardu, composé de divers programmes télévisés français – train : 22 h ; dev : 7 h ; test : 7 h.

DNN-HMM hybride Les modèles acoustiques (AM) sont entraînés sur des caractéristiques MFCC ou autosupervisés haute résolution (*hires*) de 40 dimensions à l'aide de la boîte à outils Kaldi (Povey *et al.*, 2011) avec un réseau neuronal factorisé à retardement (TDNN-F) (Povey *et al.*, 2018; Peddinti *et al.*, 2015) sur le corpus d'entraînement ETAPE (Gravier *et al.*, 2012) uniquement. Deux modèles de langue trigrammes ont été utilisés pour l'évaluation : (1) entraîné sur les données d'entraînement ESTER-1.2 et EPAC – avec un vocabulaire de 82k – et (2) entraîné sur les données d'entraînement ETAPE uniquement – avec un vocabulaire plus petit de 17,5k – voir (Evain *et al.*, 2021).

Bout en bout Nos systèmes de bout en bout (e2e) sont implémentés avec la boîte à outils SpeechBrain (Ravanelli *et al.*, 2021). Le système e2e de base est alimenté par des caractéristiques de type banc de filtres (MFB) à 80 dimensions et basé sur une architecture d'encodeur/décodeur avec attention. Lorsqu'il est utilisé avec un modèle autosupervisé préentraîné Wav2Vec2.0, le système e2e ajoute

2. https://github.com/pytorch/fairseq/tree/main/examples/wav2vec

Modèles de langue	ETAPE		ESTER-1.2 + EPAC	
Traits	**Dev**	**Test**	**Dev**	**Test**
hires MFCC	36.89±0.66	38.50±0.71	29.56±0.70	31.93±0.75
(a) prétraitement agnostique				
XLSR-53-*large*	34.28±0.69	36.03±0.72	27.01±0.68	29.64±0.77
Fr-3K-*large*	31.85±0.64	33.46±0.69	26.54±0.65	28.56±0.72
Fr-7K-*large*	**28.75**±0.62	**30.30**±0.68	**23.62**±0.63	**25.64**±0.70
(c) prétraitement spécifique à la tâche (ajustement fin pour la RAP sur ETAPE)				
Fr-3K-*large*	**28.82**±0.62	**30.19**±0.67	23.67±0.62	**25.22**±0.70
Fr-7K-*large*	28.84±0.61	30.29±0.66	**23.44**±0.62	25.36±0.70

TABLE 1 – Résultats RAP (WER,%) sur le corpus ETAPE pour les modèles accoustiques DNN-HMM avec une topologie TDNN-F. Les nombres en gris indiquent une intervalle de confiance à 95%.

Corpus	CommonVoice		ETAPE	
Traits	**Dev**	**Test**	**Dev**	**Test**
MFB	17.67±0.37	20.59±0.41	54.03±1.33	54.36±1.32
XLSR-53-*large*	16.41±0.27	19.40±0.29	58.55±0.65	61.03±0.70
Fr-3K-*large*	**8.34**±0.18	**9.75**±0.20	**23.51**±0.68	**26.14**±0.77
Fr-7K-*large*	8.55±0.18	9.94±0.21	24.14±0.70	27.25±0.78

TABLE 2 – Résultats de la RAP de bout en bout (WER%) sur les corpus *Common Voice* et ETAPE, avec des modèles wav2vec2.0 préentraînés puis affinés sur des données transcrites

simplement une couche cachée supplémentaire et une couche de sortie au-dessus de l'architecture Wav2Vec2.0. Les détails sont donnés dans (Evain *et al.*, 2021).

Résultats Les résultats du WER sur les ensembles de données de développement et de test ETAPE pour les modèles hybrides DNN-HMM sont présentés dans le tableau **??** et ceux des modèles de bout en bout sont donnés dans le tableau **??**. Parmi les modèles formés sur les caractéristiques autosupervisées, tous les modèles autosupervisés présentés constituent une amélioration par rapport au modèle basé sur les MFCC ou bancs de filtres (MFB) que ce soit avec ou sans affinage. Les modèles français (Fr-3K et Fr-7K) sont par ailleurs plus performants que le modèle multilingue ce qui montre l'importance de préentraîner des modèles spécifiques à la langue cible.

2.3 Compréhension du langage parlé (CLP)

Dataset. La CLP vise à extraire une représentation sémantique d'un signal vocal dans des applications d'interaction humain-machine (Mori, 1997). Puisqu'une application de CLP générique est difficile à créer, de nombreux travaux se concentrent sur des domaines de spécialité. Nous nous concentrons sur le domaine de l'information et de la réservation d'hôtel fourni par le corpus français MEDIA (Bonneau-Maynard *et al.*, 2006). Ce corpus est constitué de 1 250 dialogues humain-machine acquis avec une approche magicien d'Oz, où 250 utilisateurs ont suivi 5 scénarios de réservation. Les données parlées ont été transcrites manuellement et annotées avec les concepts prédéfinis dans l'ontologie du domaine. Le corpus officiel est réparti en 12 908 énoncés (41,5 h) pour l'entraînement, 1 259 énoncés (3,5 h) pour le développement et 3 005 énoncés (11,3 h) pour le test. Nous notons que, alors que tous les tours de parole ont été transcrits manuellement et peuvent être utilisés pour entraîner les modèles RAP, seuls les tours des utilisateurs ont été annotés avec des concepts et peuvent être utilisés pour entraîner les modèles CLP. Il en résulte seulement $16,8$ heures de données d'entraînement oral pour

Traits en entrée	Dev	Test
spectrogramme	**29.07**±1.31	**31.10**±0.83
(a) modèles agnostiques		
Fr-3K-*large*	**15.96**±1.02	**15.95**±0.62
Fr-7K-*large*	17.25±1.02	16.35±0.66
XLSR-53-*large*	18.45±1.15	18.78±0.66
(b) modèles affinés (auto-supervisés sur MEDIA)		
Fr-3K-*large*	15.93±1.01	**14.94**±0.60
Fr-7K-*large*	**15.42**±1.03	15.17±0.60
XLSR-53-*large*	16.77±1.09	15.56±0.61
(c) modèles affinés (supervisés RAP sur MEDIA)		
Fr-3K-*large*	**14.49**±1.06	13.97±0.59
Fr-7K-*large*	14.58±1.01	**13.78**±0.58
XLSR-53-*large*	16.05±1.05	15.46±0.60

TABLE 3 – Résultats de la tâche SLU (CER) sur le corpus MEDIA avec modèles de bout-en-bout.

les modèles CLP.

Expériences Tous nos modèles sont basés sur une architecture *seq2seq* avec LSTM et mécanisme d'attention (Hochreiter & Schmidhuber, 1997; Bahdanau *et al.*, 2015). Comme dans (Chan *et al.*, 2016), nous utilisons 3 couches LSTM bidirectionnelles empilées de manière pyramidale dans l'encodeur et 2 couches LSTM unidirectionnelles dans le décodeur. Toutes les couches ont une taille de 256. En plus d'utiliser, comme caractéristiques en entrée des modèles, les spectrogrammes et les caractéristiques extraites avec des modèles autosupervisés agnostiques, nous utilisons également les caractéristiques extraites avec des modèles autosupervisés affinés sur la tâche CLP (MEDIA). Deux types d'affinage sont effectués : *autosupervisé*, qui consiste à reprendre l'entraînement du modèle autosupervisé en utilisant les données d'entraînement MEDIA et en minimisant la fonction de coût *Wav2Vec* (*'(b) autosupervisé sur MEDIA'* dans le tableau), également appelé préentrainement adapté à la tâche dans (Gururangan *et al.*, 2020)) ; et *supervisé RAP* (*(c) supervisé RAP sur MEDIA* dans le tableau) qui consiste à affiner le modèle autosupervisé complet pour une tâche supervisée en aval avec comme objectif la minimisation de la fonction de coût CTC (Graves *et al.*, 2006). Dans ce travail, nous avons choisi d'affiner les modèles par rapport à la tâche RAP sur MEDIA (au lieu de la tâche CLP) pour voir comment cela se compare à l'affinement autosupervisé.

Les résultats sur la tâche CLP obtenus avec différentes caractéristiques en entrée sont présentés dans le tableau 3. Les résultats sont donnés en termes de taux d'erreur au niveau des concepts (CER). Ceci est calculé de la même manière que le taux d'erreur des mots (WER) mais sur des séquences de concepts. Les CER sont accompagnés d'écarts types (en gris), calculés avec la méthode bootstrap (Bisani & Ney, 2004). L'utilisation des caractéristiques des modèles autosupervisés en entrée permet une baisse importante du CER. Au mieux, parmi les modèles agnostiques, nous obtenons un CER de 15,95 sur les données de test avec les caractéristiques Fr-3K-*large*. Étonnamment, en utilisant les caractéristiques du modèle entraîné avec 7 000 heures de données (Fr-7K-*large*), les résultats sont inférieurs, à la fois sur les données de test et de développement. Le meilleur modèle affiné est le modèle 7k affiné pour la RAP sur MEDIA, bien que les résultats soient proches de ceux obtenus avec les caractéristiques du modèle 3k (13.97 vs. 13.78). Cela montre que les modèles autosupervisés

peuvent être spécialisés en utilisant un préentraînement spécifique à la tâche avec soit un apprentissage autosupervisé sur la parole brute (bloc (b) dans le tableau), soit un réglage fin sur la parole brute et les transcriptions associées (bloc (c) dans le tableau), ce dernier étant meilleur que le premier.

2.4 Traduction automatique Multilingue parole vers texte (TAP)

La traduction automatique de la parole vers le texte (TAP) consiste à traduire un énoncé vocal dans une langue source vers un texte dans une langue cible. Dans ce travail, nous nous intéressons à la traduction directe de la parole en français vers un texte dans une autre langue (bout en bout).

Dataset Nous avons sélectionné des sous-ensembles ayant le français comme source dans le jeu de données multilingue TEDx (Salesky *et al.*, 2021). Notre benchmark couvre les traduction depuis le français vers trois langues cibles : L'anglais (en), l'espagnol (es) et le portugais (pt), avec les tailles d'apprentissage suivantes : 50 h (en), 38 h (es) et 25 h (pt).

Expériences Les modèles de base sont des modèles utilisant des caractéristiques MFB à 80 dimensions. Pour les représentations apprises dérivées des modèles autosupervisés, nous nous sommes concentrés sur l'approche d'extraction de caractéristiques où ces dernières sont extraites à partir d'un préentraînement agnostique ou spécifique à la tâche. Le préentraînement agnostique se réfère à l'utilisation directe des modèles autosupervisés comme extracteurs de caractéristiques, tandis que la méthode spécifique à la tâche consiste en une phase supplémentaire où les modèles autosupervisés sont entraînés sur les données de la tâche dans le domaine, avec étiquettes (affinement fin supervisé) ou sans (affinement fin autosupervisé). Nous avons effectué un réglage fin supervisé avec des transcriptions vocales comme étiquettes et nous laissons le réglage fin supervisé avec des données de traduction de parole pour des travaux futurs. Dans le scénario spécifique à la tâche, nous avons considéré trois modèles autosupervisés : deux modèles pour le français (Fr-3K-*large* et Fr-7K-*large*) et un modèle autosupervisé incluant d'autres langues (XLSR-53-*large*). Comme la parole française se chevauche entre les paires de langues, nous avons sélectionné la paire ayant le plus de données vocales (fr-en) pour effectuer un préentraînement spécifique à la tâche et nous avons utilisé les modèles obtenus pour extraire les caractéristiques des autres paires (fr-es et fr-pt). Pour une comparaison équitable, nous n'avons pas utilisé de technique supplémentaire d'augmentation des données ni de préentrainement de l'encodeur RAP dans les expériences.

Résultats Le tableau 4 présente quelques exemples de résultats. On peut observer que les caractéristiques autosupervisées, qu'elles soient agnostiques ou spécifiques à la tâche et qu'elles soient préentraînées sur des données françaises ou multilingues, surpassent largement les modèles de base exploitant les caractéristiques MFB (à l'exception du modèle multilingue agnostique XLSR-53 sur les deux paires fr-es et fr-pt, qui se trouvent dans des environnements à très faibles ressources). Parmi les trois groupes utilisant les caractéristiques autosupervisées (préentrainement agnostique, autosupervisé spécifique à la tâche et réglage fin spécifique à la tâche de RAP), l'approche de réglage fin pour la RAP (c) donne les meilleurs résultats. Nous observons des améliorations importantes en passant de l'autosupervisé spécifique à la tâche (b) au réglage fin spécifique à la tâche (c) ($+6.19$, $+8.50$, $+8.53$ en moyenne pour en, es, et pt, respectivement) alors que les bénéfices de l'utilisation du réglage fin autosupervisé par rapport au préentraînement agnostique de la tâche sont seulement marginaux ou même légèrement négatifs. Les gains substantiels obtenus lors de l'utilisation de l'approche de réglage fin supervisé (même avec un signal quelque peu indirect, à savoir les transcriptions de la tâche RAP en aval) montrent qu'il est utile de donner plus d'information des données spécifiques à la tâche aux modèles autosupervisés. En particulier, dans le cas du réglage fin autosupervisé spécifique

Traits	Validation			Test		
	en	es	pt	en	es	pt
MFB	1.15±0.17	0.67±0.15	0.61±0.13	1.10±0.14	0.87±0.12	0.32±0.03
(a) préentraînement agnostique						
Fr-3K-*large*	17.94±0.51	16.40±0.49	8.64±0.34	18.00±0.51	18.12±0.48	9.55±0.36
Fr-7K-*large*	19.23±0.54	17.59±0.49	9.68±0.37	19.04±0.53	18.24±0.49	10.98±0.41
XLSR-53-*large*	7.81±0.33	0.49±0.13	0.43±0.07	6.75±0.29	0.52±0.08	0.36±0.05
(b) Préentraînement spécifique à la tâche (autoapprentissage sur mTEDx)						
Fr-3K-*large*	18.54±0.53	16.40±0.48	8.81±0.36	18.38±0.52	17.84±0.48	10.57±0.41
Fr-7K-*large*	19.65±0.55	17.53±0.47	9.35±0.36	19.36±0.54	18.95±0.53	10.94±0.38
XLSR-53-*large*	6.83±0.33	0.54±0.14	0.34±0.03	6.75±0.32	0.34±0.03	0.29±0.03
(c) préentraînement spécifique à la tâche (Affinement RAP sur mTEDx)						
Fr-3K-*large*	21.09±0.53	19.28±0.53	14.40±0.47	21.34±0.58	21.18±0.52	16.66±0.49
Fr-7K-*large*	**21.41**±0.51	20.32±0.49	**15.14**±0.48	**21.69**±0.58	**21.57**±0.52	**17.43**±0.52
XLSR-53-*large*	21.09±0.54	**20.38**±0.56	14.56±0.45	20.68±0.53	21.14±0.55	17.21±0.54

TABLE 4 – BLEU sur les ensembles de validation et de test de TEDx multilingue (mTEDx). La valeur la plus élevée de chaque groupe (agnostique, spécifique à la tâche et affinage fin supervisé) est soulignée tandis que la meilleure valeur de chaque colonne est en **gras**. Les chiffres gris indiquent l'écart type calculé à l'aide du rééchantillonnage d'amorçage de (Koehn, 2004).

à la tâche (b), nous avons entraîné les modèles autosupervisés pour plus d'étapes sur les données brutes spécifiques à la tâche, tandis que dans le scénario de réglage fin RAP (c), nous avons utilisé les données brutes ainsi que les transcriptions pour guider les modèles autosupervisés.

2.5 Reconnaissance automatique d'émotions (RAE)

L'objectif principal des systèmes de RAE consiste à détecter un ensemble de catégories émotionnelles à partir de segments de parole, ou à prédire en temps continu des dimensions affectives telles que l'éveil ou le plaisir intrinsèque (valence émotionnelle). Ici, nous utilisons des modèles séquence-à-séquence pour prédire des dimensions continues de l'émotion.

Datasets Nous utilisons les jeux de données RECOLA (Ringeval *et al.*, 2013) et AlloSat (Macary *et al.*, 2020) comme dans (Evain *et al.*, 2021). RECOLA est un corpus bien connu pour l'évaluation des systèmes de RAE ; il contient des enregistrements d'interactions spontanées entre des sujets francophones acquis dans des conditions de laboratoire. AlloSat est un jeu de données plus récent qui inclut des conversations réelles en français et issues de centres d'appels téléphoniques. Les deux jeux de données sont annotés en temps continu par plusieurs annotateurs. Les différentes annotations sont moyennées pour chaque dimension affective afin de définir une cible (*gold-standard*) pour l'apprentissage des modèles. Ces dimensions sont l'éveil (de passif à actif) et la valence (de négatif à positif) échantillonnées à 25 Hz pour le corpus RECOLA, et la satisfaction pour le corpus AlloSat, échantilloné à 4 Hz.

Expériences En plus d'utiliser les caractéristiques autosupervisées, nous avons extrait des caractéristiques de type MFB à 40 dimensions qui ont été centrées et réduites selon les mesures effectuées sur la partition d'apprentissage. Nous avons utilisé deux types de modèle de RAE : un modèle simple (LinTh) avec une seule couche linéaire suivie d'une fonction tangente hyperbolique, et un modèle permettant de mémoriser le contexte avec un modèle GRU à une couche dont la dimension D varie

Traits	Corpus - Tâches								
	RECOLA - Eveil			RECOLA - Valence			AlloSat - Satisfaction		
	Modèle								
	LinTh	GRU-32	GRU-64	LinTh	GRU-32	GRU-64	LinTh	GRU-32	GRU-64
MFB	.139	.655	.649	.107	.373	.421	.121	.611	.612
XLSR-53-*large*	.237	.661	.669	.005	.322	.200	.242	.578	.582
Fr-3K-*large*	.378	.267	.349	.130	.202	.033	.009	.468	.473
Fr-7K-*large*	.310	.203	.078	.020	.214	.068	.007	.510	.474

TABLE 5 – Coefficient de corrélation par concordance des prédictions émotionnelles obtenues sur les partitions de test des corpus RECOLA et AlloSat.

$D = [32, 64]$, suivie de la couche LinTh. La métrique d'évaluation est le coefficient de corrélation par concordance (Lawrence & Lin, 1989) entre les prédictions obtenues par le modèle de RAE et les cibles obtenues par les annotations humaines (Weninger *et al.*, 2016).

Résultats Un résultat notable parmi ceux présentés dans le tableau **??** est que les caractéristiques autosupervisées obtiennent de bien meilleurs résultats que les caractéristiques MFB avec un modèle de RAE simple (LinTanh). Lorsque les modèles deviennent plus complexes (GRU-32 et GRU-64), les gains de performance obtenus par les caractéristiques autosupervisées par rapport aux caractéristiques MFB est moins notable sur les jeux de caractéristiques testées. Cela montre l'efficacité des représentations de plus haut niveau (autosupervisées) pour la RAE uniquement lorsqu'un modèle de faible complexité (LinTanh) est utilisé. Ainsi, même si les modèles autosupervisés sont capables d'atteindre des informations de plus haut niveau que les MFB, ils peinent à extraire des informations liées à l'affect. Nous devons cependant souligner le fait que le préentraînement des modèles 3k impliquait moins de 1% de données émotionnelles. Nous observons également de grandes variations de performance d'un modèle autosupervisé à l'autre, probablement parce que la RAE est une tâche à très faible ressource et qu'il est donc difficile de conclure sur l'efficacité de nos modèles autosupervisés entraînés sur des données françaises par rapport à ceux entraînés sur des données multilingues ou anglaises. Enfin, des tentatives de préentraînement spécifiques à la tâche (non rapportées ici) ont également été faites sur RECOLA avec des modèles Fr-3k, mais dans les scénarios de réglage fin autosupervisé et basé sur le RBA, les modèles n'ont pas convergé. Des recherches supplémentaires sont nécessaires afin de mieux comprendre ce comportement. Notons toutefois que des gains de performance peuvent être obtenus avec les modèles autosupervisés par rapport aux modèles MFB en choisissant des corpus spécifiques tels que ceux utilisés dans le modèle 2.7k (Evain *et al.*, 2021).

3 Conclusions & perspectives

Nous avons mis en place un site Web (`http://lebenchmark.com`) dans le but de : (a) d'établir un lien vers les modèles préentraînés et les scripts permettant de reproduire les expériences présentées dans cet article, (b) de garder une trace, par le biais d'un *classement*, des futurs articles et résultats qui utiliseraient notre cadre d'évaluation, et (c) de soutenir les contributions pour d'autres langues afin de faire croître *LeBenchmark* de manière dynamique.

Nous encourageons la communauté à participer par la création de nouveaux corpus d'évaluation libres et facilement accessibles pour stimuler la recherche sur la langue française.

Références

ARDILA R., BRANSON M., DAVIS K., HENRETTY M., KOHLER M., MEYER J., MORAIS R., SAUNDERS L., TYERS F. M. & WEBER G. (2020). Common voice : A massively-multilingual speech corpus. In *LREC*.

BACHMAN P., HJELM R. D. & BUCHWALTER W. (2019). Learning representations by maximizing mutual information across views. *arXiv preprint arXiv :1906.00910*.

BAEVSKI A., AULI M. & MOHAMED A. (2019). Effectiveness of self-supervised pre-training for speech recognition. *CoRR*, **abs/1911.03912**.

BAHDANAU D., CHO K. & BENGIO Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. In *Proc. of ICLR*.

BISANI M. & NEY H. (2004). Bootstrap estimates for confidence intervals in ASR performance evaluation. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, p. I–409 : IEEE.

BONNEAU-MAYNARD H., AYACHE C., BECHET F., DENIS A., KUHN A., LEFÈVRE F., MOSTEFA D., QUIGNARD M., ROSSET S., SERVAN C. *et al.* (2006). Results of the french evalda-media evaluation campaign for literal understanding. In *The fifth international conference on Language Resources and Evaluation (LREC 2006)*.

CHAN W., JAITLY N., LE Q. V. & VINYALS O. (2016). Listen, Attend and Spell : A Neural Network for Large Vocabulary Conversational Speech Recognition. In *ICASSP 2016* : IEEE.

CHEN T., KORNBLITH S., NOROUZI M. & HINTON G. (2020). A simple framework for contrastive learning of visual representations. In *PMLR*.

DEVLIN J., CHANG M., LEE K. & TOUTANOVA K. (2018). BERT : pre-training of deep bidirectional transformers for language understanding. *CoRR*, **abs/1810.04805**.

EVAIN S., NGUYEN M. H., LE H., ZANON BOITO M., MDHAFFAR S., ALISAMIR S., TONG Z., TOMASHENKO N., DINARELLI M., PARCOLLET T., ALLAUZEN A., ESTÈVE Y., LECOUTEUX B., PORTET F., ROSSATO S., RINGEVAL F., SCHWAB D. & BESACIER L. (2021). Task Agnostic and Task Specific Self-Supervised Learning from Speech with LeBenchmark. In *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS 2021)*, NeurIPS 2021 Datasets and Benchmarks Track, on-line, United States.

GRAVES A., FERNÁNDEZ S., GOMEZ F. J. & SCHMIDHUBER J. (2006). Connectionist temporal classification : labelling unsegmented sequence data with recurrent neural networks. In *ICML*, volume 148 of *ACM International Conference Proceeding Series*, p. 369–376 : ACM.

GRAVIER G., ADDA G., PAULSON N., CARRÉ M., GIRAUDEL A. & GALIBERT O. (2012). The etape corpus for the evaluation of speech-based tv content processing in the french language. In *LREC*.

GURURANGAN S., MARASOVIĆ A., SWAYAMDIPTA S., LO K., BELTAGY I., DOWNEY D. & SMITH N. A. (2020). Don't stop pretraining : Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 8342–8360, Online : Association for Computational Linguistics.

HOCHREITER S. & SCHMIDHUBER J. (1997). Long short-term memory. *Neural Comput.*, **9**(8).

KAWAKAMI K., WANG L., DYER C., BLUNSOM P. & VAN DEN OORD A. (2020). Learning robust and multilingual speech representations. In *EMNLP*.

KOEHN P. (2004). Statistical significance tests for machine translation evaluation. In *EMNLP*, p. 388–395 : ACL.

LAWRENCE I. & LIN K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, p. 255–268.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBE B., BESACIER L. & SCHWAB D. (2020). FlauBERT : Unsupervised Language Model Pre-training for French. In *LREC*, Marseille, France.

MACARY M., TAHON M., ESTÈVE Y. & ROUSSEAU A. (2020). Allosat : A new call center french corpus for satisfaction and frustration analysis. In *Proceedings of the 12th Language Resources and Evaluation Conference*, p. 1590–1597.

MORI R. D. (1997). *Spoken Dialogues with Computers*. Orlando, FL, USA : Academic Press, Inc.

PEDDINTI V., POVEY D. & KHUDANPUR S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. In *Interspeech*, p. 3214–3218.

POVEY D., CHENG G., WANG Y., LI K., XU H., YARMOHAMMADI M. *et al.* (2018). Semi-orthogonal low-rank matrix factorization for deep neural networks. In *Interspeech*, p. 3743–3747.

POVEY D., GHOSHAL A., BOULIANNE G., BURGET L., GLEMBEK O., GOEL N., HANNEMANN M., MOTLICEK P., QIAN Y., SCHWARZ P. *et al.* (2011). The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*.

RAFFEL C., SHAZEER N., ROBERTS A., LEE K., NARANG S., MATENA M., ZHOU Y., LI W. & LIU P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv :1910.10683*.

RAVANELLI M., PARCOLLET T., ROUHE A., PLANTINGA P., RASTORGUEVA E., LUGOSCH L., DAWALATABAD N., JU-CHIEH C., HEBA A., GRONDIN F., ARIS W., LIAO C.-F., CORNELL S., YEH S.-L., NA H., GAO Y., FU S.-W., SUBAKAN C., DE MORI R. & BENGIO Y. (2021). Speechbrain. `https://github.com/speechbrain/speechbrain`.

RINGEVAL F., SONDEREGGER A., SAUER J. & LALANNE D. (2013). Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, p. 1–8 : IEEE.

RUDER S. (2021). Challenges and Opportunities in NLP Benchmarking. `http://ruder.io/nlp-benchmarking`.

SALESKY E., WIESNER M., BREMERMAN J., CATTONI R., NEGRI M., TURCHI M., OARD D. W. & POST M. (2021). The multilingual tedx corpus for speech recognition and translation. *arXiv preprint arXiv :2102.01757*.

SHON S., PASAD A., WU F., BRUSCO P., ARTZI Y., LIVESCU K. & HAN K. J. (2021). Slue : New benchmark tasks for spoken language understanding evaluation on natural speech.

WENINGER F., RINGEVAL F., MARCHI E. & SCHULLER B. (2016). Discriminatively Trained Recurrent Neural Networks for Continuous Dimensional Emotion Recognition from Audio. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI 2016)*, p. 2196–2202, New York City (NY), USA : IJCAI/AAAI.

YANG S., CHI P., CHUANG Y., LAI C. J., LAKHOTIA K., LIN Y. Y., LIU A. T., SHI J., CHANG X., LIN G., HUANG T., TSENG W., LEE K., LIU D., HUANG Z., DONG S., LI S., WATANABE S., MOHAMED A. & LEE H. (2021). SUPERB : speech processing universal performance benchmark. *CoRR*, **abs/2105.01051**.