Modèles neuronaux pré-appris par auto-supervision sur des enregistrements de parole en français

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

Modèles neuronaux pré-appris par auto-supervision sur des enregistrements de parole en français *

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

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

RÉSUMÉ

L'apprentissage auto-supervisé a ouvert des perspectives prometteuses dans de nombreux domaines comme la vision par ordinateur, le traitement automatique de la langue ou celui de la parole. Les modèles pré-appris sur de grandes quantités de données non étiquetées peuvent être ajustés sur de petits ensembles de données transcrites manuellement. Ceux de type wav2vec2.0 ont montré des performances remarquables pour la reconnaissance automatique de la parole. Les premiers modèles partagés à la communauté ayant été appris sur des données en anglais ou multilingues, nous proposons dans cet article sept modèles de type wav2vec2.0, appris sur 1 000, 3 000 et 7 000 heures de parole en français. Leur Apprentissage nécessitant des capacités de calcul très importantes, et dans un esprit de science ouverte, ceux-ci sont librement accessibles. Des résultats expérimentaux sur la reconnaissance automatique de la parole sont également présentés et confirment le bénéfice de l'utilisation de tels modèles.

ABSTRACT

Self-supervised pre-trained neural networks with French speech recordings

Self-Supervised Learning (SSL) has yielded remarkable improvements in many different domains including computer vision, natural language processing and speech processing. Models pre-trained on a large amount of unlabeled data can be further fine-tuned with very small manually-labeled datasets. Wav2Vec2.0 models showed very remarkable results for automatic speech recognition. The very first models shared to the community had been trained on English or multiple languages. We offer, in this paper, seven wav2vec2.0 models, trained on 1 000, 3 000 and 7 000 hours of French speech. Considering that the training of such models requires large computing capacities, and in a spirit of open science, our French models are freely available. Experimental results on ASR are also displayed, confirming the benefit of using such pre-trained models.

MOTS-CLÉS : Modèles Autosupervisés, Reconnaissance Automatique de la Parole.

∗. note aux relecteurs : cette soumission est une sous-partie ré-adaptée d'un article publié à Neurips fin 2021. Nous avons tenu à faire cet effort afin de présenter ces travaux sur des modèles 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 de l'article original.

KEYWORDS: Self-Supervised Representation Learning, Automatic Speech Recognition.

1 Introduction

L'apprentissage auto-supervisé a ouvert des perspectives prometteuses dans de nombreux domaines, comme la vision par ordinateur et le traitement automatique de la langue écrite (Bachman *et al.*, 2019; Chen *et al.*, 2020; Devlin *et al.*, 2018; Raffel *et al.*, 2019). Des études récentes ont également montré l'intérêt de ces approches pour construire des représentations de la parole particulièrement pertinentes pour la reconnaissance automatique de la parole (RAP) (Baevski *et al.*, 2019; Kawakami *et al.*, 2020). Les modèles utilisés dans ces travaux et leurs extensions ont principalement eu pour objet la langue anglaise ou bien le multilinguisme. L'apprentissage de ces modèles nécessite une quantité importante d'enregistrements audio et requiert une puissance de calcul considérable (en termes de nombre d'heures et d'infrastructure). Néanmoins ces modèles sont distribués et accessibles pour la communauté scientifique, voire industrielle, du domaine. Dans cet article, nous présentons les premiers modèles neuronaux auto-supervisés pré-appris pour la langue française, les données utilisées pour l'apprentissage, les processus d'apprentissage dédiés, ainsi que quelques résultats obtenus pour une tâche de RAP. Cet ensemble de modèles neuronaux pré-appris par auto-supervision sur de grandes quantités de parole en français (1 000, 3 000, et 7 000 heures) est accessible à la communauté, en particulier via la plateforme *HuggingFace* [1].

2 Apprentissage auto-supervisé

L'apprentissage auto-supervisé à partir du signal de parole consiste à construire un modèle (ici un réseau de neurones profonds) capable de résoudre des *pseudo-tâches* qui ne nécessitent pas d'annotation humaine. Par exemple, le codage prédictif autorégressif (APC) tente de capturer la structure séquentielle de la parole par la prédiction d'informations sur la trame future (Chung & Glass, 2020b). Le codage prédictif contrastif (CPC), quant à lui, définit une tâche de classification binaire : prédire si une trame de parole future est bien la bonne (positive) ou si elle correspond à une trame tirée aléatoirement dans le reste du signal (négative) (Baevski *et al.*, 2019; Schneider *et al.*, 2019). L'objectif de ces modèles est d'apprendre à représenter le signal de parole : une partie du modèle dédiée à la représentation du signal est réutilisée pour construire un système spécialisé sur une vraie tâche. Comme le montre (Chung & Glass, 2020a), ces représentations peuvent améliorer les performances de systèmes pour plusieurs tâches nécessitant un traitement automatique de la parole. De plus, ces représentations peuvent, dans certains cas, se transférer à d'autres langues (Riviere *et al.*, 2020).

En 2020, une méthode performante est publiée pour l'apprentissage par auto-supervision de modèles neuronaux dédiés au traitement de la parole. Il s'agit du modèle wav2vec 2.0 (Baevski *et al.*, 2020b) qui repose sur l'idée du CPC (Baevski *et al.*, 2019; Schneider *et al.*, 2019) combinée à la construction de représentations latentes discrétisées de trames de parole. Le modèle wav2vec 2.0 est alimenté directement par une séquence de valeurs numériques représentant la forme d'onde d'un enregistrement audio. Ces valeurs sont normalisées : la séquence est centrée réduite. La séquence de valeurs normalisées est alors traitée par 7 blocs convolutifs – un bloc est constitué d'une convolution

1. https://huggingface.co/LeBenchmark

1D, d'une couche de normalisation et de l'application d'une fonction d'activation GELU : *Gaussian Error Linear Units* (Hendrycks & Gimpel, 2016). Ces 7 blocs convolutifs produisent une séquence de vecteurs de caractéristiques latentes. Un module de quantification, qui s'appuie sur une distribution Gumbel-Softmax (Jang *et al.*, 2017), est appliqué à chacun de ces vecteurs dans le but de projeter ces vecteurs vers un nombre fini de vecteurs (ici 102 400 possibilités). En parallèle, la séquence de vecteurs de caractéristiques latentes non quantifiée sert d'entrée d'un modèle neuronal constitué d'une pile de blocs d'encodeur de type Transformer. De manière aléatoire, certains vecteurs de cette séquence sont masqués : environ la moitié des vecteurs sont remplacés par un vecteur de masquage durant l'apprentissage. En sortie de la couche d'encodeurs Transformer, pour chacun des vecteurs associés à un vecteur d'entrée masqué, il s'agit de présenter 101 vecteurs discrets : celui qui correspond à la version quantifiée du vecteur d'entrée avant masquage, et 100 vecteurs *distracteurs* tirés parmi les 102 399 autres vecteurs possibles. Une fonction de coût nommée *contrastive loss* est alors appliquée, qui tend à rapprocher, du point de vue de la similarité cosinus, le vecteur prédit du vecteur quantifié d'origine, tout en l'éloignant des vecteurs distracteurs. Pour s'assurer que les 102 400 possibilités de vecteurs discrets sont utilisés, un terme de diversité nommé *diversity loss* est également pris en compte. Pour plus de détails, le lecteur peut se référer à l'article (Baevski *et al.*, 2020a).

3 Collecte d'une grande quantité de données orales hétérogènes en français

De grands corpus multilingues comprenant du français ont récemment été mis à disposition : MLS (Pratap *et al.*, 2020) (1,096 h), ou encore voxpopuli (Wang *et al.*, 2021) (+4,500 h). Ces corpus se restreignent à la lecture ou à la parole préparée énoncée par des professionnels et ne permettent pas ou trop peu l'accès à des données orales diverses comme de la parole spontanée, expressive, ou avec un accent. Afin de compléter ces corpus, nous avons rassemblé une grande quantité de données orales en français comprenant différents accents (MLS, African Accented Speech, CaFE), des émotions actées (GEMEP, CaFE, Att-Hack), des conversations téléphoniques (PORTMEDIA), de la lecture (MLS, African Accented French, MaSS), de la parole spontanée (CFPP2000, ESLO2, MPF, TCOF, NCCFr), de la parole journalistique (EPAC) et de la parole professionnelle (Voxpopuli). En comparaison à MLS et Voxpopuli utilisés seuls, notre ensemble de données est plus divers et les représentations des tours de parole plus réalistes. Le tableau 1 rassemble la durée moyenne des segments ainsi que quelques métadonnées telles que le genre, le type de parole, et le nombre de locuteurs.

Pré-traitement des enregistrements avant apprentissage auto-supervisé. Les enregistrements ont été segmentés en fonction des temps de début et de fin renseignés dans leurs annotations d'origine. Nous avons ensuite filtré les segments inférieurs à 1 s et supérieurs à 30 s, comme présenté dans (Baevski *et al.*, 2020b). Quand cela était possible, les segments avec de la parole superposée ont également été retirés. L'ensemble des enregistrements a été converti en mono, PCM, 16 bits, 16 000 Hz. Nous avons conservé, quand cela était renseigné, l'identifiant de chaque locuteur ainsi que son genre.

Corpus 1k (≈ 1 100 heures). Cette première collection n'est composée que du corpus MLS, ceci à des fins de comparaison avec l'étude wav2vec2.0 (Baevski *et al.*, 2020b) où seule de la parole lue en anglais est présente. Elle est équilibrée en termes de genre.

Corpus 3k (≈ 2 900 heures). Il comprend les 1 096 heures du corpus 1k, ainsi que 115 heures

Corpus$_{Licence}$	**# Segments**	**Durée**	**# Locuteurs**	**Durée moyenne seg.**	**Type de parole**
		Corpus 1k			
MLS français$_{CCBY4.0}$ (Pratap *et al.*, 2020)	**263 055** 124 590 / 138 465 / –	**1 096 :43** 520 :13 / 576 :29 / –	**178** 80 / 98 / –	**15 s** 15 s / 15 s / –	Lecture
		Corpus 3k			
African Accented French$_{Apache2.0}$ (SLR57, 2003)	**16 402** 373 / 102 / 15 927	**18 :56** – / – / 18 :56	**232** 48 / 36 / 148	**4 s** – / – / –	Lecture
Att-Hack$_{CCBYNCND}$ (Le Moine & Obin, 2020)	**36 339** 16 564 / 19 775 / –	**27 :02** 12 :07 / 14 :54 / –	**20** 9 / 11 / –	**2,7 s** 2,6 s / 2,7 s / –	Expressive actée
CaFE$_{CCNC}$ (Gournay *et al.*, 2018)	**936** 468 / 468 / –	**1 :09** 0 :32 / 0 :36 / –	**12** 6 / 6 / –	**4,4 s** 4,2 s / 4,7 s / –	Expressive actée
CFPP2000$_{CCBYNCSA}$* (Branca-Rosoff *et al.*, 2012)	**9853** 166 / 1 184 / 8 503	**16 :26** 0 :14 / 1 :56 / 14 :16	**49** 2 / 4 / 43	**6 s** 5 s / 5 s / 6 s	Spontanée
ESLO2$_{NC}$ (Eshkol-Taravella *et al.*, 2011)	**62 918** 30 440 / 32 147 / 331	**34 :12** 17 :06 / 16 :57 / 0 :09	**190** 68 / 120 / 2	**1,9 s** 2 s / 1,9 s / 1,7 s	Spontanée
EPAC**$_{NC}$ (Estève *et al.*, 2010)	**623 250** 465 859 / 157 391 / –	**1 626 :02** 1 240 :10 / 385 :52 / –	**Unk** – / – / –	**9 s** – / – / –	Journalistique
GEMEP$_{NC}$ (Bänziger *et al.*, 2012)	**1 236** 616 / 620 / –	**0 :50** 0 :24 / 0 :26 / –	**10** 5 / 5 / –	**2,5 s** 2,4 s / 2,5 s / –	Expressive actée
MPF (Françoise, 2017), (MPF, 2019)	**19 527** 5 326 / 4 649 / 9 552	**19 :06** 5 :26 / 4 :36 / 9 :03	**114** 36 / 29 / 49	**3,5 s** 3,7 s / 3,6 s / 3,4 s	Spontanée
PORTMEDIA$_{NC}$ (French) (Lefèvre *et al.*, 2012)	**19 627** 9 294 / 10 333 / –	**38 :59** 19 :08 / 19 :50 / –	**193** 84 / 109 / –	**7,1 s** 7,4 s / 6,9 s / –	Conversations téléphoniques actées
TCOF (Adults) (ATILF, 2020)	**58 722** 10 377 / 14 763 / 33 582	**53 :59** 9 :33 / 12 :39 / 31 :46	**749** 119 / 162 / 468	**3,3 s** 3,3 s / 3,1 s / 3,4 s	Spontanée
Total corpus 3k	**1 111 865** 664 073 / 379 897 / 67 895	**2 933 :24** 1 824 :53 / 1 034 :15 / 74 :10	-	-	-
		Corpus 7k			
MaSS (Boito *et al.*, 2020)	**8 219** 8 219 / – / –	**19 :40** 19 :40 / – / –	**Unk** – / – / –	**8,6 s** 8,6 s / – / –	Lecture
NCCFr$_{NC}$ (Torreira *et al.*, 2010)	**29 421** 14 570 / 13 922 / 929	**26 :35** 12 :44 / 12 :59 / 00 :50	**46** 24 / 21 / 1	**3 s** 3 s / 3 s / 3 s	Spontanée
Voxpopuli$_{CC0}$ (Wang *et al.*, 2021) *Unlabeled*	**568 338** – / – / –	**4 532 :17** – / – / 4 532 :17	**Unk** – / – / –	**29 s** – / – / –	Parole professionnelle
Voxpopuli$_{CC0}$ (Wang *et al.*, 2021) *transcribed*	**76 281** – / – / –	**211 :57** – / – / 211 :57	**327** – / – / –	**10 s** – / – / –	Parole professionnelle
Total corpus 7k***	**1 814 242** 682 322 / 388 217 / 99 084	**7 739 :22** 1 853 :02 / 1 041 :07 / 4 845 :07	-	-	-

TABLE 1 – Statistiques des corpus audio utilisés pour apprendre les modèles auto-supervisés en fonction du genre des locuteurs (homme / femme / non renseigné). Le corpus 1k ne comprend que le corpus MLS. Chacun des corpus est construit sur la base du corpus précédent auquel est ajouté des données ; durée : heure(s) :minute(s).

*Composé d'enregistrements non inclus dans le corpus CEFC v2.1 du 02/2021 ; **les locuteurs n'ont pas un identifiant unique ; ***Les statistiques relatives aux corpus CFPP2000, MPF et TCOF ont changé légèrement après amélioration du script d'extraction des données ; Licence : CC=Creative Commons ; NC= Non commercial ; BY= Attribution ; SA= Partage dans les mêmes conditions ; ND = Pas de modification ; CC0 = Transfert dans le domaine public

supplémentaires de lecture, 1 626 heures de parole journalistique, 123 heures de parole spontanée, 38 heures de conversations téléphoniques actées et 29 heures de parole expressive actée, pour un total de 2 933 heures de parole. En ce qui concerne le genre, le corpus contient 1 824 heures de parole d'hommes, 1 034 heures de parole de femmes et 74 heures de parole pour lesquelles le genre est non renseigné.

Corpus 7k (≈ 7 700 heures). Il comprend quatre corpus de plus : MaSS, NCCFr et Voxpopuli (non transcrit et transcrit). Cela représente au total 7 739 heures de parole, dont 1 135 heures de lecture, 1 626 heures de parole journalistique, 165 heures de parole spontanée, 38 heures de conversations téléphoniques actées, 29 heures de parole expressive actée et 4 744 heures de parole professionnelle. Sur les quatre corpus ajoutés, seul le corpus NCCFr contient des informations sur le genre des locuteurs.

4 Apprentissage et présentation des modèles

Dans le cadre de cette étude, nous avons appris sept modèles wav2vec 2.0 (Baevski *et al.*, 2020b) sur les données françaises décrites dans la section 3. Deux architectures wav2vec 2.0 différentes ont été considérées – *base* et *large* – pour apprentissage à partir des corpus *1k*, *3k* et *7k* pour former notre ensemble de modèles wav2vec 2.0 : Fr-1K-*base*, Fr-1K-*large*, Fr-3K-*base*, Fr-3K-*large*, Fr-7K-*base*, Fr-7K-*large*. Nous fournissons également un modèle spécifique (Fr-2.7K-*base*) pré-appris sur un sous-ensemble de notre corpus *3k* contenant uniquement MLS et EPAC (2,7K heures d'audio) afin de permettre une étude plus approfondie de l'impact de la parole spontanée sur l'apprentissage des représentations.

Les hyperparamètres et les architectures *base* et *large* sont identiques à ceux introduits dans (Baevski *et al.*, 2020b). Les modèles Fr-1K, Fr-3K, Fr-2.7K et Fr-7K sont appris avec respectivement 200K, 500K, 500K et 500K mises à jour des poids sur 4, 32, 32 et 64 Nvidia Tesla V100 (32GB). Un résumé des différents hyperparamètres utilisés pour apprendre nos modèles est disponible dans le Tableau 2. En pratique, l'apprentissage est stoppé au bout d'un nombre rond de mises à jour dès que la fonction de coût observée sur l'ensemble de développement du corpus MLS atteint un point stable.

L'ensemble des modèles wav2vec 2.0 pré-appris sont partagés avec la communauté via HuggingFace pour permettre l'intégration avec des bibliothèques populaires comme SpeechBrain (Ravanelli *et al.*, 2021), Fairseq (Ott *et al.*, 2019) ou Kaldi (Povey *et al.*, 2011).

Modèle	**Taille des données d'entraînement**	**Couches Transformer (encodeurs)**	**Dimension *embedding* externe**	**Dimension *embedding* interne**	**Nombre de têtes d'attention**	**Nombre de mises à jour**
Fr-1K-*base*	1 096 h	12	768	3 072	8	200K
Fr-1K-*large*	1 096 h	24	1 024	4 096	16	200K
Fr-2.7K-*base*	2 773 h	12	768	3 072	8	500K
Fr-3K-*base*	2 933 h	12	768	3 072	8	500K
Fr-3K-*large*	2 933 h	24	1 024	4 096	16	500K
Fr-7K-*base*	7 739 h	12	768	3 072	8	500K
Fr-7K-*large*	7 739 h	24	1 024	4 096	16	500K

TABLE 2 – Hyperparamètres de chacun de nos modèles pré-appris sur le français

5 Reconnaissance automatique de la parole

Afin de mesurer l'apport des modèles wav2vec 2.0 décrits dans les sections précédentes, il est possible de mesurer leur impact sur les performances d'un système de RAP dit de bout en bout : un modèle neuronal unique qui transcrit un signal de parole sans module intermédiaire. Des modèles wav2vec 2.0 sont distribués par Facebook [2], pré-appris sur des données de langue anglaise ou sur des corpus multilingues. Dans les expériences présentées dans cet article, qui s'inspirent du protocole

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

présenté dans Baevski *et al.* (2020b), nous comparons nos modèles au modèle wav2vec 2.0 LS960-*large*, pré-appris à partir des données LibriSpeech sur 960 heures de parole anglaise lue, et au modèle XLSR-53-*large* pré-appris sur 60 000 heures de parole de 53 langues différentes (Conneau *et al.*, 2020), y compris le français.

Nous comparons également ces modèles à un modèle neuronal alimenté par des paramètres acoustiques classiques de type banc de filtres répartis sur une échelle Mel. Ces paramètres sont des vecteurs de dimension 80.

Dans cet article, nous avons choisi de présenter les résultats d'expériences réalisées sur les données distribuées dans le cadre de la campagne d'évaluation ETAPE (Gravier *et al.*, 2012; Galibert *et al.*, 2014). Pour des résultats obtenus sur un plus grand nombre de tâches ou de données, le lecteur est invité à lire l'article compagnon de celui-ci, également soumis aux JEP et décrivant le référentiel *LeBenchmark* disponible et accompagné d'un classement de modèles sur le site Web suivant : `http://www.lebenchmark.com`.

Données expérimentales L'expérience menée ici relève d'un scénario dans lequel peu de données annotées sont disponibles, puisque le corpus d'apprentissage se limite à 22h de paroles transcrites manuellement [3]. Les données ETAPE sont composées d'enregistrements d'émissions radiophoniques et télévisées. Ces émissions sont de différents types : journaux d'actualité, débats, divertissement, provenant de BFM TV, LCP, TV8 Mont Blanc, et France Inter (Galibert *et al.*, 2014). La répartition des données, en termes d'heures de parole transcrite, est présentée dans le tableau 3.

corpus d'apprentissage	**corpus de développement**	**corpus de test**
22 heures	7 heures	7 heures

TABLE 3 – Répartition des données distribuées lors de la campagne d'évaluation ETAPE.

Architectures neuronales Nos systèmes de reconnaissance automatique de la parole sont construits à partir de la boîte à outils SpeechBrain (Ravanelli *et al.*, 2021). Ces systèmes sont chacun composés d'un modèle neuronal dit de bout en bout. Le modèle alimenté par des paramètres acoustiques plus classiques s'appuie sur une architecture encodeur/décodeur avec mécanisme d'attention. L'encodeur est composé de 3 couches convolutives, de 5 couches biGRU et 2 couches denses de 500 dimensions chacune. Le décodeur est composé d'une couche GRU et le mécanisme d'attention est de type *location-aware*. En sortie, le modèle émet des sous-unités lexicales de type *bype pair encoding* (bpe) parmi un vocabulaire de taille 500. Ce modèle est celui qui obtenait le meilleur résultat à partir de paramètres acoustiques de 80 dimensions et de type logarithmes de banc de filtres répartis sur une échelle de Mél.

Pour les modèles s'appuyant sur des modèles wav2vec 2.0, nous avons simplement ajouté une couche cachée au sommet du modèle wav2vec 2.0, ainsi qu'une couche de sortie émettant des caractères, tels que proposé dans Baevski *et al.* (2020b). Ces modèles sont appris à l'aide d'une fonction de coût de type *Connectionist Temporal Classification* (CTC) pendant 80 *epochs*. Durant l'apprentissage sur les données du corpus ETAPE, les poids des modèles wav2vec 2.0 sont mis à jour.

3. seules les données effectivement distribuées durant la campagne ETAPE sont utilisées ici ; lors de la campagne, les données ESTER, ESTER 2 et EPAC étaient autorisées : elles ne sont pas utilisées dans ces expériences

Corpus	**WER sur ETAPE**	
Paramètres ou modèle	**Dev**	**Test**
Banc de filtres Mél	54.03±1.33	54.36±1.32
En-LS960-*large*	42.14±0.72	44.82±0.74
XLSR-53-*large*	58.55±0.65	61.03±0.70
Fr-2.7K-*base*	26.23±0.78	29.08±0.80
Fr-3K-*base*	26.14±0.70	28.86±0.79
Fr-3K-*large*	**23.51**±0.68	**26.14**±0.77
Fr-7K-*base*	25.13±0.68	28.16±0.79
Fr-7K-*large*	24.14±0.70	27.25±0.78

TABLE 4 – Taux d'erreurs sur les mots (WER%) obtenus par les différents modèles sur les données du corpus ETAPE. En grisé, l'intervalle de confiance.

Résultats Les résultats, en taux d'erreurs sur les mots (WER) obtenus par les différents modèles sont présentés dans le tableau 4. Parmi les systèmes s'appuyant sur des modèles wav2vec 2.0, un seul obtient de moins bons résultats que le modèle encodeur-décodeur alimenté par des bancs de filtres Mél (54,36% de WER sur le test ETAPE) : il s'agit du modèle XLSR-53-*large* pré-appris sur 53 langues dont le français (61,03% de WER).

Si le modèle En-LS960-*large*, pré-appris sur de l'anglais, obtient de meilleurs résultats que le modèle encodeur-décodeur avec un WER de 44,82%, ses performances sont bien en-deçà des modèles pré-appris sur du français. Parmi ces modèles, celui qui permet d'obtenir les meilleurs résultats est le modèle Fr-3K-*large*, pré-appris sur environ 3 000 heures de données (26,14%). Il obtient de meilleurs résultats que le modèle Fr-7K-*large*, pré-appris sur environ 7 000 heures de données (27,25%). Une raison à cela pourrait venir de la nature des données de pré-apprentissage. Alors que les corpus de pré-apprentissage de ces deux modèles contiennent la même quantité – environ 1 600 heures – de parole journalistique, proche des données ETAPE d'un point de vue acoustique, les 4 000 heures supplémentaires ajoutées dans le corpus de pré-apprentissage du modèle Fr-7K-*large* sont des enregistrements provenant du Parlement européen (discours, réunions, événements) (Wang *et al.*, 2021). L'utilisation de données hétérogènes pour construire le corpus de pré-apprentissage est motivé par l'hypothèse que ce corpus permettra de construire un modèle wav2vec 2.0 plus robuste à la variabilité des conditions acoustiques. Cependant, cela nuit peut-être aux performances sur traitement de données spécialisées comme celles qui constituent le corpus ETAPE. Cette question n'est pas abordée dans cet article et sera l'objet d'études à venir. L'article compagnon de cet article soumis également aux JEP 2022 peut donner quelques premiers éléments de réponse.

Enfin, nous pouvons noter que le meilleur système de reconnaissance automatique de la parole participant à la campagne officielle ETAPE en 2012 avait obtenu un WER de 23,6% (Bougares *et al.*, 2013). Cependant, ce système s'appuyait sur un corpus d'apprentissage de 511 heures de parole transcrites (contre 22 heures ici), et un modèle de langage appris sur des millions de mots (nous n'avons pas utilisé de modèle de langage dans les expériences présentées ici). Dans l'article compagnon évoqué plus haut, nous présentons également des résultats obtenus par un système Kaldi HMM/TDNN-f dont le modèle acoustique TDNN-f a été appris dans les mêmes conditions que les modèles présentés ici, et qui utilise un modèle de langage comparable à celui du meilleur système de

la campagne ETAPE. Avec une paramétrisation acoustique classique de type MFCC haute résolution, ce système hybride à l'état de l'art obtient un WER de 31,93%. Par ailleurs, ce système hybride a pu montrer que les modèles 1K n'était pas performants pour l'ASR (34,91% WER), c'est pour cela qu'ils n'ont pas été testé avec l'approche de bout en bout.

Ces résultats confirment l'intérêt des modèles de type wav2vec 2.0, et montrent également l'importance de la nature des données utilisées pour leur pré-apprentissage. Il apparaît clairement que pour traiter la langue française, les modèles proposés ici, pré-appris sur des données de cette langue, sont plus appropriés que des modèles appris sur d'autres langues. Enfin, il est notable que les architectures *large*, comprenant environ 330 millions paramètres, donnent de meilleurs résultats que les architectures *base*, dont le nombre de paramètres est d'environ 90 millions.

6 Conclusion

Dans cet article, nous avons présenté comment sept modèles pour le français ont été pré-appris sur de grands corpus d'enregistrement de parole. Ces modèles sont mis à la disposition de la communauté afin qu'ils puissent servir à de futures recherches en traitement automatique de la parole. Les modèles sont facilement utilisables via `https://huggingface.co/LeBenchmark`. L'architecture de ces modèles, les méthodes d'apprentissage et les données utilisées sont également été décrites. Nous avons estimé l'empreinte carbone de l'apprentissage de notre modèle 7k large à 270 kg de CO2. Ce qui, en comparaison du modèle GPT-3 nécessitant 28000 jours de GPUs (Anthony *et al.*, 2020) (soit 10 tonnes de CO2 s'il avait été appris en France) reste raisonnable pour un modèle partagé à la communauté.

Afin d'illustrer l'impact de tels modèles, nous décrivons également des résultats expérimentaux de reconnaissance automatique de la parole. La cadre expérimentale suit un scénario dans lequel peu de données d'apprentissage sont disponibles (ici, 22h de parole transcrite pour la reconnaissance automatique de parole radio- ou télédiffusée).

Nos résultats confirment que dans ce contexte les modèles wav2vec 2.0 sont bien plus performants que les approches plus classiques qui s'appuient sur des paramètres acoustiques de type MFCC ou banc de filtres. Nous avons également pu vérifier que les modèles pré-appris sur le français sont plus adaptés à la reconnaissance automatique de la parole en français que les modèles pré-appris sur l'anglais ou des corpus multilingues. La capacité du modèle est également importante, les modèles *large* ayant obtenu les meilleures performances par rapport aux modèles *base*.

Remerciements

Ce travail a bénéficié du programme 'Grand Challenge Jean Zay' et a également été partiellement soutenu par MIAI@Grenoble-Alpes (ANR-19-P3IA-0003) et partiellement financé par la Commission européenne dans le cadre du projet SELMA sous le contrat numéro 957017.

Références

ANTHONY L. F. W., KANDING B. & SELVAN R. (2020). Carbontracker : Tracking and predicting the carbon footprint of training deep learning models. *arXiv preprint arXiv :2007.03051*.

ATILF (2020). TCOF : Traitement de corpus oraux en français. https ://hdl.handle.net/11403/tcof/v2.1, ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

BACHMAN P., HJELM R. D. & BUCHWALTER W. (2019). Learning representations by maximizing mutual information across views. In *NeurIPS*.

BAEVSKI A. *et al.* (2020a). wav2vec 2.0 : A framework for self-supervised learning of speech representations. *arXiv :2006.11477*.

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

BAEVSKI A., ZHOU Y., MOHAMED A. & AULI M. (2020b). wav2vec 2.0 : A framework for self-supervised learning of speech representations. In H. LAROCHELLE, M. RANZATO, R. HADSELL, M. BALCAN & H. LIN, Eds., *Advances in Neural Information Processing Systems 33 : Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

BOITO M. Z., HAVARD W., GARNERIN M., LE FERRAND É. & BESACIER L. (2020). MaSS : A large and clean multilingual corpus of sentence-aligned spoken utterances extracted from the Bible. In *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France : European Language Resources Association.

BOUGARES F., DELÉGLISE P., ESTÈVE Y. & ROUVIER M. (2013). Lium asr system for etape french evaluation campaign : experiments on system combination using open-source recognizers. In *International Conference on Text, Speech and Dialogue*, p. 319–326 : Springer.

BRANCA-ROSOFF S., FLEURY S., LEFEUVRE F. & PIRES M. (2012). Discours sur la ville. Présentation du Corpus de Français parlé Parisien des années 2000 (CFPP2000). http ://cfpp2000.univ-paris3.fr/CFPP2000.pdf.

BÄNZIGER T., MORTILLARO M. & SCHERER K. (2012). Introducing the Geneva Multimodal Expression Corpus for Experimental Research on Emotion Perception. *Emotion (Washington, D.C.)*, **12**(5), 1161–79.

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

CHUNG Y. & GLASS J. (2020a). Generative pre-training for speech with autoregressive predictive coding. In *ICASSP*.

CHUNG Y.-A. & GLASS J. (2020b). Improved speech representations with multi-target autoregressive predictive coding.

CONNEAU A., BAEVSKI A., COLLOBERT R., MOHAMED A. & AULI M. (2020). Unsupervised cross-lingual representation learning for speech recognition. *arXiv :2006.13979*.

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

ESHKOL-TARAVELLA I., BAUDE O., MAUREL D., HRIBA L., DUGUA C. & TELLIER I. (2011). Un grand corpus oral "disponible" : le corpus d'Orléans 1968-2012. *Ressources Linguistiques Libres - Traitement Automatique des Langues*, **53**(2), 17–46.

ESTÈVE Y., BAZILLON T., ANTOINE J.-Y., BÉCHET F. & FARINAS J. (2010). The EPAC Corpus : Manual and Automatic Annotations of Conversational Speech in French Broadcast News. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta : European Language Resources Association (ELRA).

FRANÇOISE G. (2017). Les parlers jeunes dans l'île-de-France multiculturelle. *Paris and Gap, Ophrys*.

GALIBERT O., LEIXA J., ADDA G., CHOUKRI K. & GRAVIER G. (2014). The ETAPE speech processing evaluation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, p. 3995–3999, Reykjavik, Iceland : European Language Resources Association (ELRA).

GOURNAY P., LAHAIE O. & LEFEBVRE R. (2018). A Canadian French emotional speech dataset. In *MMSys*.

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*.

HENDRYCKS D. & GIMPEL K. (2016). Gaussian error linear units (gelus). *arXiv preprint arXiv :1606.08415*.

JANG E., GU S. & POOLE B. (2017). Categorical reparameterization with gumbel-softmax. In *ICLR 2017*.

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

LE MOINE C. & OBIN N. (2020). Att-HACK : An Expressive Speech Database with Social Attitudes. In *Speech Prosody*.

LEFÈVRE F., MOSTEFA D., BESACIER L., ESTÈVE Y., QUIGNARD M., CAMELIN N., FAVRE B., JABAIAN B. & ROJAS-BARAHONA L. (2012). Robustesse et portabilités multilingue et multi-domaines des systèmes de compréhension de la parole : le projet PortMedia. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2012*, volume 1 :JEP, p. 779–786, Grenoble, France.

MPF (2019). Multicultural Paris French. Type : dataset, https ://hdl.handle.net/11403/mpf/v3.

OTT M., EDUNOV S., BAEVSKI A., FAN A., GROSS S., NG N., GRANGIER D. & AULI M. (2019). fairseq : A fast, extensible toolkit for sequence modeling. In *Proc. of NAACL-HLT : Demonstrations*.

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*.

PRATAP V., XU Q., SRIRAM A., SYNNAEVE G. & COLLOBERT R. (2020). Mls : A large-scale multilingual dataset for speech research. In *INTERSPEECH*, Shanghai, China.

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`.

RIVIERE M., JOULIN A., MAZARÉ P.-E. & DUPOUX E. (2020). Unsupervised pretraining transfers well across languages. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 7414–7418 : IEEE.

SCHNEIDER S., BAEVSKI A., COLLOBERT R. & AULI M. (2019). wav2vec : Unsupervised Pre-Training for Speech Recognition. In *Proc. Interspeech 2019*, p. 3465–3469.

SLR57 (2003). African accented french. Type : dataset, https ://www.openslr.org/57/.

TORREIRA F., ADDA-DECKER M. & ERNESTUS M. (2010). The Nijmegen Corpus of Casual French. *Speech Communication*, **52**(3), 201. Publisher : Elsevier : North-Holland.

WANG C., RIVIÈRE M., LEE A., WU A., TALNIKAR C., HAZIZA D., WILLIAMSON M., PINO J. & DUPOUX E. (2021). Voxpopuli : A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *arXiv :2101.00390*.