Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Audiocite.net: Un grand corpus d'enregistrements vocaux de lecture en Français

Soline Felice* Solène Evain Solange Rossato François Portet
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
solene.evain, solange.rossato, francois.portet@univ-grenoble-alpes.fr,
soline.felice@univ-tlse2.fr

RÉSUMÉ

L'arrivée de l'apprentissage auto-supervisé dans le domaine du traitement automatique de la parole a permis l'utilisation de grands corpus non étiquetés pour obtenir des modèles pré-appris utilisés comme encodeurs des signaux de parole pour de nombreuses tâches. Toutefois, l'application de ces méthodes de SSL sur des langues telles que le français s'est montrée difficile due à la quantité limitée de corpus de parole du français publiquement accessible. C'est dans cet objectif que nous présentons le corpus Audiocite.net comprenant 6682 heures d'enregistrements de lecture par 130 locuteurs et locutrices. Ce corpus est construit à partir de livres audio provenant du site *audiocite.net*. En plus de décrire le processus de création et les statistiques obtenues, nous montrons également l'impact de ce corpus sur les modèles du projet LeBenchmark dans leurs versions 14k pour des tâches de traitement automatique de la parole.

ABSTRACT

Audiocite.net : A Large Spoken Read Dataset in French

The advent of self-supervised learning (SSL) in speech processing has allowed the use of large unlabeled datasets to learn pre-trained models, serving as powerful encoders for various downstream tasks. However, the application of these SSL methods to languages such as French has proved difficult due to the scarcity of large French speech datasets. To advance the emergence of pre-trained models for French speech, we present the Audiocite.net corpus composed of 6 682 hours of recordings from 130 readers. This corpus is built from audiobooks from the *audiocite.net* website. In addition to describing the creation process and final statistics, we also show how this dataset impacted the models of LeBenchmark project in its 14k version for speech processing downstream tasks.

MOTS-CLÉS : Ensembles de données vocales, apprentissage auto-supervisé, traitement automatique de la parole.

KEYWORDS: Spoken Datasets, French Speech, Self Supervised Learning, Automatic Speech Processing.

1 Introduction

L'arrivée de l'apprentissage auto-supervisé (*Self-Supervised Learning* – SSL) dans le domaine du traitement automatique de la parole a permis l'utilisation de grands corpus non étiquetés pour obtenir

∗. maintenant à l'IRIT, Univ. Toulouse 2 Jean Jaurès.

des modèles pré-appris.

De nombreux modèles profonds pré-appris modélisant le signal acoustique de la parole ont émergé utilisant l'apprentissage auto-supervisé génératif (PASE+ (Ravanelli *et al.*, 2020), Mockingjay (Liu *et al.*, 2020)) ; une fonction de coût contrastive (CPC (Oord *et al.*, 2019), Speech SimCLR (Jiang *et al.*, 2021), Wav2Vec 2.0 (Baevski *et al.*, 2020)) ; ou prédictive (HuBERT (Hsu *et al.*, 2021), wavLM (Chen *et al.*, 2022), data2vec (Baevski *et al.*, 2022)) (Abdel-Rahman *et al.*, 2022). Ces modèles ont fait avancer les performances du traitement de la parole en les adaptant et utilisant comme encodeurs pour les tâches de traitement automatique de la parole (*downstream tasks*). Par exemple, Wav2Vec 2.0 a pu atteindre des résultats à l'état de l'art avec un modèle pré-appris puis ajusté avec un minimum de données étiquetées pour une tâche de reconnaissance automatique de la parole (RAP) dans un contexte de lecture en anglais.

Les modèles pré-appris par SSL dépendent fortement de la disponibilité d'une grande quantité de données d'apprentissage. Bien que plusieurs grands ensembles de données pour l'anglais et multilingues ont été publiés, des ensembles de données aussi volumineux pour le français sont rares. Jusqu'à récemment, il était difficile de trouver de grands ensembles de données de parole française disponibles publiquement (à l'exception des 1,700 heures de parole transcrites automatiquement d'EPAC par (Estève *et al.*, 2010)). Récemment, de grands corpus multilingues incluant le français ont été rendus disponibles, tels que MLS (1,096 h) (Pratap *et al.*, 2020), ou VoxPopuli (non transcrit +4,500 h) (Wang *et al.*, 2021). Cependant, ces ensembles de données représentent toujours une quantité bien inférieure à ce qui est disponible pour l'anglais. Le multilinguisme a été souligné comme un moyen de traiter les langues sous-dotées, mais l'étude menée dans le projet LeBenchmark (Parcollet *et al.*, 2024) visant à créer des modèles de parole pré-appris pour le français a montré que les modèles entraînés sur des données cibles monolingues sont bien plus efficaces que ceux multilingues.

Dans cet article, nous présentons Audiocite.net, un corpus non transcrit d'environ 6 600 heures d'enregistrements de parole lue en français, disponible librement pour la communauté. Nous décrivons la sélection et l'acquisition des données (voir sec. 2) ainsi que les principales caractéristiques de la publication du jeu de données (sec. 3). Nous montrons également comment ce jeu de données a impacté le modèle de 14k du projet LeBenchmark 2.0 (Parcollet *et al.*, 2024) pour certaines tâches de traitement de la parole, notamment la reconnaissance automatique de la parole lue (sec. 4).

Cet article est l'adaptation vers le français d'un article publié à LREC/COLING 2024 (Felice *et al.*, 2024).

2 Sélection et acquisition des données

Les grands corpus de parole proviennent souvent de l'extraction de contenus libres sur le web, comme Librispeech, extrait du projet LibriVox (Panayotov *et al.*, 2015). Cependant, les oeuvres publiées en France ne deviennent disponibles dans le domaine public que 70 ans après le décès de leurs auteurs et autrices, rendant les livres plus modernes publiés après 1953 inaccessibles. Pour surmonter cette limitation, l'initiative Common Voice (Ardila *et al.*, 2020) a été mise en place par la fondation Mozilla pour capturer la parole lue en utilisant des phrases collectées sur le web. En quatre ans, environ 1,100 heures de segments de phrases ont été collectées en Français. Pour rassembler un plus grand volume de parole lue continue, allant de la littérature classique à la moderne et librement accessible, nous avons décidé de nous concentrer sur le site *audiocite.net*.

Catégorie	# Fichiers				Durée (hh :mm :ss)			
	All	Train	Dev	Test	All	Train	Dev	Test
animaux	**160**	108	31	21	**26 :49 :04**	16 :12 :46	05 :33 :54	05 :02 :23
juniors	**35**	18	7	10	**04 :56 :39**	01 :45 :14	00 :41 :59	02 :29 :24
charme	**166**	166	0	0	**33 :02 :05**	33 :02 :05	00 :00 :00	00 :00 :00
contes	**2430**	1711	415	304	**490 :40 :12**	325 :44 :09	94 :56 :54	69 :59 :08
cuisine	**39**	35	3	1	**02 :44 :47**	02 :19 :18	00 :13 :37	00 :11 :51
documents	**1494**	1191	181	122	**367 :17 :28**	265 :35 :35	58 :22 :12	43 :19 :41
histoire	**1341**	1167	99	75	**397 :33 :01**	333 :51 :58	33 :19 :12	30 :21 :50
nouvelles	**2772**	1721	534	517	**721 :09 :55**	375 :11 :21	167 :15 :11	178 :43 :21
philosophies	**1052**	773	53	226	**181 :04 :51**	117 :50 :07	17 :02 :07	46 :12 :36
planete-actuelle	**145**	145	0	0	**18 :27 :20**	18 :27 :20	00 :00 :00	00 :00 :00
poesies	**2274**	1956	160	158	**116 :09 :42**	84 :37 :27	14 :41 :07	16 :51 :07
religions	**777**	777	0	0	**213 :21 :47**	213 :21 :47	00 :00 :0	00 :00 :0
romans	**14664**	13088	713	863	**3 943 :54 :09**	3 377 :46 :53	267 :58 :14	298 :09 :01
science-fiction	**478**	349	117	12	**122 :03 :39**	87 :48 :10	28 :16 :00	05 :59 :28
theatre	**658**	603	19	36	**42 :45 :29**	36 :58 :06	02 :35 :49	03 :11 :34
Tout	**28 485**	23 808	2 332	2345	**6 682 :00 :18**	5 290 :32 :24	690 :56 :22	700 :31 :31

TABLE 1 – Statistiques (durées et nombre de fichiers) par catégorie de livres avec les détails des partitions (**all**, train / dev / test)

Audiocité est une association à but non lucratif qui met à disposition une plate-forme où les bénévoles peuvent partager leurs lectures. Avant leur première contribution, les bénévoles doivent passer un test de lecture afin d'évaluer leur prononciation, leur rythme de lecture, les conditions d'enregistrement et le format final du fichier audio. Des conseils de post-traitement sont fournis selon les besoins (par exemple, pour réduire les bruits de respiration). Les enregistrements comprennent environ 5 000 livres audio d'œuvres littéraires classiques issues du domaine public en français (Balzac, Hugo, Maupassant, Molière…) et environ 700 livres audio d'auteurs et autrices de l'époque contemporaine qui ont choisi de partager librement leurs oeuvres (Brussolo, Huchon, Del, Martin, Fée …).

2.1 Critères de sélection des livres audio

Pour qu'un corpus soit utile et serve la recherche reproductible, il devrait idéalement être à la fois accessible et gratuit. Sur le site web, tous les livres audio sont distribués sous une licence Creative Commons, car, avant de déposer leur enregistrement, les personnes doivent prendre en compte les droits d'auteur qui confèrent à l'auteur ou autrice du livre les droits exclusifs d'utiliser, de copier, de licencier, d'exécuter et de modifier l'œuvre. Ainsi, les lecteurs et lectrices sont autorisés à lire des livres du domaine public (c'est-à-dire sans droits d'auteur) ou des livres contemporains dont les auteurs et autrices ont donné l'autorisation d'enregistrer leur texte et de le distribuer sur *audiocite.net* sous une licence spécifique.

2.2 Processus de téléchargement

Les données ont été collectées en deux étapes en novembre 2021. L'administrateur du site web a aimablement donné son aval. Dans un premier temps, tout le catalogue du site a été extrait afin de collecter toutes les métadonnées concernant chaque livre audio (œuvre originale, sujet, auteurs/autrices, lecteurs/lectrices, licence. . .). Dans un second temps, tous les fichiers audio ont été téléchargés ce qui a pris environ une semaine. Les fichiers téléchargés étaient soit au format MP3, soit dans des archives zip. Par la suite, les enregistrements audio qui ne répondaient pas aux critères d'une durée supérieure à 5 secondes ou qui étaient fournis sans licence permettant leur utilisation ont été retirés. Quelques livres audio avec des URL défectueuses ont également été exclus.

2.3 Données recueillies

Au total, 6,682 heures de livres audio lus par 130 locuteurs et locutrices, dont 70 hommes (62%), 51 femmes (34%) et 9 personnes dont le genre n'a pu être identifié (4%) ont été collectées. Cela correspond à une taille totale de 340 Go de fichiers audio et de métadonnées. Le tableau 1 indique le nombre de fichiers et la durée des fichiers audio pour chaque catégorie de livre.

Fichiers audio : Il convient de mentionner que sur les 4,378 livres audio, 388 étaient directement hébergés par *audiocite.net*, tandis que les 3,990 restants étaient hébergés sur des instances de *archive.org*. Contrairement aux lignes directrices données, les enregistrements peuvent être mono ou stéréo, et des variations dans les débits binaires ou les taux d'échantillonnage peuvent également être constatés. Certains enregistrements peuvent contenir une succession de plusieurs locuteurs et locutrices et des bruits de fond ou de la musique. En outre, tous les enregistrements n'impliquent pas nécessairement la lecture de livres publiés ; certains sont des articles ou des podcasts.

Métadonnées : Parallèlement aux fichiers audio, des métadonnées ont été téléchargées, telles que la durée de chaque fichier audio, l'identifiant du locuteur ou de la locutrice, le titre du livre lu, l'auteur ou l'autrice du livre, la catégorie du livre et la licence liée à l'audio. Ces informations ont été fournies par les locuteurs et locutrices eux-mêmes sur la page web de chaque livre audio de *audiocite.net*.

3 Organisation du corpus Audiocite.net

Bien que l'un des principaux usages envisagés pour ce jeu de données soit l'apprentissage auto-supervisé, nous anticipons également d'autres utilisations. En effet, même s'il n'est pas transcrit, le jeu de données pourrait être utilisé pour la modélisation thématique, la reconstruction de signaux ou la synthèse vocale. C'est pourquoi nous l'avons publié avec des partitions officielles et des fichiers de métadonnées faciles à interroger.

3.1 Estimation du genre des locuteurs et locutrices

Pour minimiser les biais dans les partitions, nous avons déduit le genre des locuteurs et locutrices en fonction de leurs identifiants ou en écoutant leurs voix. Cette information a été ajoutée aux métadonnées. Cependant, nous ne garantissons pas que l'information soit fiable ni que la méthode

utilisée soit viable pour déduire le genre d'une personne puisqu'elle n'est pas basée sur l'auto-identification de celle-ci. Les métadonnées concernant le genre doivent être traitées avec prudence.

3.2 Partitions des données

Le jeu de données a été divisé en trois partitions : une partition d'apprentissage (train) comprenant 80% des enregistrements, une partition de développement (dev) en comprenant 10%, et une partition de test (test) comprenant les 10% restants. Le tableau 2 fournit les statistiques de durée par genre pour chaque sous-ensemble.

# Personnes	Durée Totale	Durée Moyenne	# Fichiers
TRAIN			
74 T	5290 :32 :24	00 :13 :19	23808
35 F	1577 :23 :53	00 :16 :54	5600
30 H	3431 :01 :21	00 :11 :52	17329
9 I	282 :07 :09	00 :19 :15	879
DEV			
78 T	690 :56 :22	00 :17 :46	2332
44 F	344 :14 :51	00 :15 :40	1317
34 H	346 :41 :31	00 :20 :29	1015
TEST			
61 T	700 :31 :31	00 :17 :55	2345
38 F	350 :39 :38	00 :15 :39	1344
23 H	349 :51 :53	00 :20 :58	1001
ALL			
130 T	6682 :00 :18	00 :14 :04	28485
70 F	2272 :18 :23	00 :16 :30	8261
51 H	4127 :34 :45	00 :12 :48	19345
9 I	282 :07 :09	00 :19 :15	879

TABLE 2 – Statistiques du corpus Audiocite.net - Nombre de fichiers, de personnes (locuteurs/locutrices) et durée par genre (tous, femme, homme et inconnu) par partitions

Les partitions de développement et de test ont été conçues pour ne pas inclure de contenu potentiellement sensible, spécifiquement ceux relevant des catégories *charmes* (érotique), *planete-actuelle* (géopolitique) et *religion*. De plus, pour ces partitions, une représentation égale de la parole masculine et féminine a été assurée et les fichiers dont le genre de la personne était inconnu n'ont pas été inclus. Le tableau 1 indique le nombre de fichiers et la durée pour chaque catégorie de livre pour les partitions d'apprentissage, de développement et de test.

3.3 Organisation de l'ensemble de données

Le jeu de données est organisé comme suit : nous partageons un fichier README et une fiche technique (inspirée par *Datasheet for datasets*, (Gebru *et al.*, 2021)) où l'on peut trouver des statistiques détaillées, ainsi que des précisions sur la composition, l'utilisation et la distribution du corpus, et trois dossiers (`wavs/`, `scripts/` et `metadata/`). Dans le dossier `wavs/`, les fichiers de livres audio sont rangés dans des dossiers selon le titre du livre lu, triés par ordre alphabétique. Nous fournissons également un dossier `scripts/` avec des scripts pour générer des statistiques sur le corpus et les fichiers json fournis. Concernant le dossier `metadata/`, deux types de fichiers de métadonnées sont partagés avec le jeu de données : *.csv* et *.json*.

Fichier download.csv : Chaque livre audio possède une entrée dans le fichier csv. Nous fournissons également des informations tels que l'identifiant du locuteur ou de la locutrice, le titre du livre lu, l'auteur ou l'autrice du livre, le genre du livre, la licence de l'enregistrement, l'adresse URL du livre audio sur *audiocite.net* et le chemin vers le fichier audio dans le dossier *wavs*.

Fichiers json : Nous fournissons quatre fichiers json : *train.json*, *dev.json*, *test.json* et *all.json*, ce dernier étant une concaténation des trois précédents. Dans ces fichiers, une entrée correspond à un fichier audio (un livre audio peut contenir plusieurs fichiers audio). Ces fichiers contiennent l'identifiant du locuteur ou de la locutrice, la durée du fichier audio en secondes, le chemin d'accès vers le fichier dans le dossier *wavs* et le genre du locuteur ou de la locutrice (F/M/U). Une entrée json prend la forme suivante :


```
"Raiponce.mp3": {
 "path": "../wavs/Raiponce/Raiponce.mp3",
 "trans": "",
 "duration": 471.552,
 "spk_id": "Demelza",
 "spk_gender": "F"
},
```


4 Impact de Audiocite.net sur les tâches de traitement automatique de la parole

Le jeu de données collecté a été partagé avec l'équipe LeBenchmark pour entraîner leurs modèles 14k (Parcollet *et al.*, 2024). Dans cette section, nous avons comparé la performance du modèle 14k à celle du modèle 7k, qui n'a pas été entraîné sur le corpus. Nous rapportons les taux d'erreur de mots (WER) pour les systèmes de reconnaissance automatique de la parole (RAP) sur un jeu de données du même type de parole et de situation (livre audio), mais aussi les résultats de reconnaissance automatique de la parole et de vérification de locuteur issus de Parcollet *et al.* (2024).

4.1 Modèles LeBenchmark

Modèle 7k-large : Ce modèle a été entraîné sur 7,000 h de parole, incluant 1,626 h de radio, 1,115 h de parole lue, 127 h de parole spontanée, 38 h de dialogue téléphonique joué et 29 h de parole émotionnelle jouée.

Modèle 14k-large : Ce modèle a été entraîné sur 14,000 h de parole contenant les données utilisées pour le modèle 7k plus l'intégralité des données de Audiocite.net (toutes les partitions), ainsi que 111 h de parole issues de diffusions radiophoniques.

4.2 Expérimentations de Reconnaissance Automatique de la Parole (RAP)

En utilisant la recette CTC (*Connectionist Temporal Classification*) de SpeechBrain pour Common Voice [1], nous avons composé un système de RAP avec LeBenchmark (Wav2Vec2) en encodeur suivi d'une couche de BiLSTM et d'une dernière couche de DNN. Nous avons utilisé les partitions

1. https://github.com/speechbrain/speechbrain/tree/develop/recipes/CommonVoice/ASR/CTC

officielles de la partie française du corpus *Multilingual Librispeech* (MLS), allouant 1,076.58 h pour l'apprentissage, 10.07 h pour le développement et 10.07 h pour les tests. Deux scénarios ont été utilisés dans l'expérience : (1) avec l'encodeur LeBenchmark figé (c'est-à-dire sans ajustement) ou (2) avec ajustement de l'encodeur (c'est-à-dire avec ajusté en même temps que le modèle de RAP). Pour l'apprentissage de la RAP, les taux d'apprentissage ont été initialisés à 0,001 pour le modèle Wav2Vec2 et à 0,1 pour la partie BiLSTM+DNN, et un recuit a été utilisé avec des facteurs de 0,9 et 0,8 respectivement. Dans le cas de l'ajustement complet, le gradient n'est propagé dans l'encodeur LeBenchmark qu'après 500 étapes. La taille du lot d'apprentissage était de 8 et la dimension de la couche de sortie était de 43 (nombre de caractères dans l'ensemble d'apprentissage).

	WER (%) ↓	
Encodeur	**Figé**	**Ajusté**
7K-*large*	31.71	9.56
14K-*large*	9.96	9.98

TABLE 3 – Résultats de RAP (WER %) sur l'ensemble de test de la partie française du corpus MLS pour les systèmes de RAP avec encodeur LeBenchmark figé ou ajusté

Le tableau 3 résume les résultats de l'expérience. L'ajustement de la partie encodeur conduit très rapidement à des performances similaires pour les modèles 14k et 7k, indiquant que Audiocite.net n'a pas beaucoup d'impact lorsque l'encodeur est ajusté. Cependant, pour l'expérience avec l'encodeur figé, il y a une nette supériorité du modèle 14k indiquant que Audiocite.net a joué un rôle important dans la modélisation de la parole lue.

4.3 Expérimentations LeBenchmark

Parmi les différentes expériences réalisées par l'équipe LeBenchmark, nous rapportons dans les tableaux 4 et 5 les résultats des tâches de reconnaissance automatique de la parole (RAP) et de vérification du locuteur issues de (Parcollet *et al.*, 2024).

Comme on peut le voir dans les expériences de reconnaissance automatique de la parole (RAP) avec Common Voice et ETAPE, Audiocite.net (14K-large) n'apporte aucune amélioration par rapport au modèle 7K. Il est même dégradé sur ETAPE qui est composé de discours radiophoniques, un type de parole très différent de Audiocite.net. Cependant, dans une tâche de vérification de locuteur sur le corpus Fabiole (Ajili *et al.*, 2016) constitué de discours d'émissions de radio et de télévision, Audiocite.net (14K-large) apporte une nette amélioration par rapport au modèle 7K. Il semble que l'ajout de locuteurs et de locutrices dans le 14K ait amélioré ce type de modélisation.

	WER (%) ↓	
Encodeur	**Common Voice**	**ETAPE**
7K-*large*	**9.39**	**23.46**
14K-*large*	9.83	26.03

TABLE 4 – Résultats de RAP (WER %) de Parcollet *et al.* (2024) sur les partitions de test de Common Voice 6.1 et ETAPE, avec des modèles Wav2Vec2.0 ajustés sur des données de RAP étiquetées

Encodeur	EER	minDCF^{-10} ↓	minDCF^{-100} ↓
7K-*large*	5.228	0.3833	0.5754
14K-*large*	**3.535**	**0.2965**	**0.4801**

TABLE 5 – Résultats de la tâche de vérification du locuteur de Parcollet *et al.* (2024) sur le corpus Fabiole. EER : *Equal Error Rate*, minDCF : *Minimum of Detection Cost Function*

5 Conclusion

Dans cet article, nous présentons le corpus Audiocite.net composé de plus de 6,600 heures d'enregistrements provenant de 130 locuteurs et locutrices, disponible sur OpenSLR (www.openslr.org/139/) avec la même licence que les livres audio (c'est-à-dire Creative Commons). Tous les enregistrements sont distribués dans leur format brut tels que nous les avons téléchargés depuis *audiocite.net* (avec des musiques de fond, des bruits, des participations inattendues, au format MP3, en mono ou stéréo). Aucun prétraitement n'a été appliqué aux fichiers, ni aucune transcription automatique effectuée sur ceux-ci. Cependant, nous avons ajouté des informations sur le genre en l'inférant à partir du nom et en vérifiant la voix en cas d'incertitude. Le corpus Audiocite.net a servi à l'apprentissage des modèles de 14k du projet LeBenchmark, révélant à la fois des performances élevées et certaines limites dans plusieurs tâches de traitement de la parole.

6 Remerciements

Ce travail a été soutenu en partie par MIAI@Grenoble-Alpes (ANR-19-P3IA-0003), le projet E-SSL (ANR-22-CE23-0013) et la Banque Publique d'Investissement (BPI) dans le cadre de la convention de subvention THERADIA. Les auteurs tiennent à remercier chaleureusement William Havard pour l'idée originale, Marcely Zanon Boito et Fabien Ringeval pour leur aide lors de la première ébauche de ce travail.

7 Considérations éthiques

Les livres sélectionnés par les lecteurs et lectrices sont soit exclusivement sous licence Creative Commons (CC), soit obtenus par des accords de distribution écrits avec les auteurs ou autrices. Une fois la lecture terminée, les locuteurs et locutrices choisissent une seconde licence Creative Commons pour l'audio avant de publier leur enregistrement sur le site *audiocite.net*. Cette seconde licence comporte des restrictions égales ou supérieures à celles assignées au livre original. En publiant leurs enregistrements sur la plateforme, les lecteurs et lectrices étaient au courant que leur matériel pourrait être utilisé à diverses fins au-delà de l'intention originale, dans les limites de la licence attribuée. L'administrateur du site *audiocite.net* nous a explicitement donné l'autorisation d'utiliser et de distribuer les audios conformément à leurs conditions d'utilisation.

Concernant le contenu des audios, toutes sortes d'affirmations peuvent y être trouvées et nous ne souhaitons encourager personne à développer une position quelconque. Nous nous engageons à supprimer l'enregistrement, ses métadonnées et à mettre à jour le corpus à la demande de toute personne contributrice désirant retirer ses données du corpus sans raison explicite.

Références

ABDEL-RAHMAN M., HUNG-YI L., LASSE B., JAKOB D. H., JOAKIM E., CHRISTIAN I., KATRIN K., SHANG-WEN L., KAREN L., LARS M., TARA N. S. & SHINJI W. (2022). Self-supervised speech representation learning : A review. *IEEE JSTSP Special Issue on Self-Supervised Learning for Speech and Audio Processing*.

AJILI M., BONASTRE J.-F., KAHN J., ROSSATO S. & BERNARD G. (2016). Fabiole, a speech database for forensic speaker comparison. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 726–733.

ARDILA R., BRANSON M., DAVIS K., KOHLER M., MEYER J., HENRETTY M., MORAIS R., SAUNDERS L., TYERS F. & WEBER G. (2020). Common Voice : A Massively-Multilingual Speech Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France.

BAEVSKI A., HSU W.-N., XU Q., BABU A., GU J. & AULI M. (2022). data2vec : A General Framework for Self-supervised Learning in Speech, Vision and Language. In *Proceedings of the 39 th International Conference on MachineLearning*, Baltimore, Maryland, USA.

BAEVSKI A., ZHOU H., MOHAMED A. & AULI M. (2020). wav2vec 2.0 : A Framework for Self-Supervised Learning of Speech Representations. In *proceedings of NeurIPS*, Vancouver, Canada.

CHEN S., WANG C., CHEN Z., WU Y., LIU S., CHEN Z., LI J., KANDA N., YOSHIOKA T., XIAO X., WU J., ZHOU L., REN S., QIAN Y., QIAN Y., ZENG M., YU X. & WEI F. (2022). WavLM : Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing*, **16**, 1–14.

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).

FELICE S., EVAIN S., ROSSATO S. & PORTET F. (2024). Audiocite.net : A large spoken read dataset in french. In *The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

GEBRU T., MORGENSTERN J., VECCHIONE B., VAUGHAN J. W., WALLACH H., III H. D. & CRAWFORD K. (2021). Datasheets for datasets. *Commun. ACM*, **64**(12), 86–92. DOI : 10.1145/3458723.

HSU W.-N., BOLTE B., TSAI Y.-H. H., LAKHOTIA K., SALAKHUTDINOV R. & MOHAMED A. (2021). HuBERT : Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, **29**, 3451–3460.

JIANG D., LI W., CAO M., ZOU W. & LI X. (2021). Speech SimCLR : Combining Contrastive and Reconstruction Objective for Self-Supervised Speech Representation Learning. In *proceedings of Interspeech*.

LIU A., YANG S.-W., CHI P.-H., HSU P.-C. & LEE H.-Y. (2020). Mockingjay : Unsupervised Speech Representation Learning with Deep Bidirectional Transformer Encoders. In *proceedings of ICASSP*.

OORD A. V. D., LI Y. & VINYALS O. (2019). Representation Learning with Contrastive Predictive Coding. arXiv :1807.03748.

PANAYOTOV V., CHEN G., POVEY D. & KHUDANPUR S. (2015). Librispeech : An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. DOI : 10.1109/ICASSP.2015.7178964.

PARCOLLET T., NGUYEN H., EVAIN S., ZANON BOITO M., PUPIER A., MDHAFFAR S., LE H., ALISAMIR S., TOMASHENKO N., DINARELLI M., ZHANG S., ALLAUZEN A., COAVOUX M., ESTÈVE Y., ROUVIER M., GOULIAN J., LECOUTEUX B., PORTET F., ROSSATO S., RINGEVAL F., SCHWAB D. & BESACIER L. (2024). Lebenchmark 2.0 : A standardized, replicable and enhanced framework for self-supervised representations of french speech. *Computer Speech & Language*, **86**, 101622. DOI : https ://doi.org/10.1016/j.csl.2024.101622.

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

RAVANELLI M., ZHONG J., PASCUAL S., SWIETOJANSKI P., MONTEIRO FILHO J., TRMAL J. & BENGIO Y. (2020). Multi-Task Self-Supervised Learning for Robust Speech Recognition. In *proceedings of ICASSP*.

WANG C., RIVIERE 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. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, Online.