Audiocite.net: A Large Spoken Read Dataset in French

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

Abstract

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.

Keywords: Spoken Datasets, French Speech, Self Supervised Learning, Automatic Speech Processing

1. Introduction

The emergence of Self-Supervised Learning (SSL) in the field of speech processing has enabled to leverage large amount of unlabeled data to learn pre-trained models (aka foundation models). Many models for deep representation of acoustic signal has emerged whether using generative SSL (PASE+ (Ravanelli et al., 2020), Mockingjay (Liu et al., 2020)); a contrastive loss (CPC (Oord et al., 2019), Speech SimCLR (Jiang et al., 2021), Wav2Vec 2.0 (Baevski et al., 2020)); or prediction target (HuBERT (Hsu et al., 2021), wavLM (Chen et al., 2022), data2vec (Baevski et al., 2022)) (Abdel-Rahman et al., 2022). These models have pushed speech processing performance forward by using such deep representation models as encoders of speech processing tasks (i.e. downstream tasks). For instance, Wav2Vec 2.0 has achieved state-of-the-art results with a pre-trained model and subsequent fine-tuning with minimal labeled data for an ASR (Automatic Speech Recognition) task in a reading context in English.

Pre-trained models based on SSL are heavily relying on a wealth of training data. If several large datasets for English and other languages have been released, such large datasets for French are scarce. Up until recently, it was difficult to find publicly available large datasets of French speech (with the exception of the 1,700 automatically transcribed hours of EPAC by (Estève et al., 2010), (Estève et al., 2010)). Recently, large multilingual corpora that include French have been made available, such as MLS (1,096 h) (Pratap et al., 2020) , (Pratap et al., 2020), or untranscribed voxpopuli (+4,500 h) (Wang et al., 2021), (Wang et al., 2021). However, these datasets total an amount still far from what is available for English. Multilingualism has been emphasized as a way to deal with under-resourced languages but the study undertaken in the LeBenchmark project (Parcollet et al., 2024) aiming at creating pre-trained speech models for French as shown that models trained on the monolingual target data are far more effective than multilingual ones.

In this paper, we present Audiocite.net an untranscribed corpus of about 6 600 hours of recording of read speech in French and freely available for the community. We describe the data selection and acquisition (see sec. 2) as well as the main characteristics of the dataset release (sec. 3). We also show how this dataset has impacted the 14k model of LeBenchmark 2.0 project (Parcollet et al., 2024) for some speech processing tasks including automatic recognition of read speech (sec. 4).

2. Data selection and acquisition

Large speech corpora have often consisted in web scraping free content such as Librispeech extracted from the LibriVox project (Panayotov et al., 2015). However, for French authors, literature becomes available in the public domain only 70 years after their death making more modern books published after 1953 non-accessible. To overcome this limitation, the Common Voice initiative (Ardila et al., 2020) has been set up by the Mozilla foundation to capture read speech using sentences collected on the web. In four years, about 1,100 hours of disconnected spoken utterances has been collected. To gather a greater amount of continuous read speech from classic to modern literature freely accessible we decided to concentrate on the *audiocite.net* website.

Audiocité is a non-profit association that offers a

* Now working with IRIT, Univ. Toulouse 2 Jean Jaurès

LREC-COLING 2024, pages 1795–1800

collaborative platform where volunteers can share their readings. Before making their first contribution, volunteers must undergo a reading test to assess their pronunciation, reading pace, recording conditions, and the final format of the audio file. Post-processing guidance is provided as needed (e.g. to reduce breathing noise). The recordings include about 5,000 audiobooks of classic public domain literary works in French (Balzac, Hugo, Maupassant, Molière, London, Doyle…) and around 700 audiobooks of contemporary authors who have chosen to freely share their work (Brussolo, Huchon, Del, Martin, Fée …).

2.1. Audiobook Selection Criteria

For a corpus to be useful and serve reproducible research it should ideally be both accessible and free to use. In the website, all audiobooks were distributed under a Creative Common license since before uploading their work, people have to take into account the copyright which provides the author of the book the exclusive rights to use, copy, license, perform and modify the creative work. So, readers are allowed to read books from the public domain (i.e. with no copyright) or contemporary books whose authors gave permission to record their text and distribute it on *audiocite.net* under a specific license.

2.2. Download Process

The data was collected in two steps in November 2021. Authorization from the website administrator was kindly granted. As a first step, all the catalog of the website was extracted. Then all the metadata about each audiobook (original work, topic, authors, reader, license…) was collected and stored. In a second step, all the audio files were downloaded, which took about a week. The downloaded files were either in .mp3 format or in .zip archives. Files in zip archives were all extracted. Subsequently, audio recordings that did not meet the criteria of being more than 5 seconds or lack of licenses permitting using them were discarded. A few audiobooks with faulty URLs were also excluded.

2.3. Collected Data

In total, 6,682 hours of audiobooks read by 130 different speakers among which 70 men (62%), 51 women (34%), and 9 people whose gender could not be identified (4%) were collected. This corresponds to a total size of 340 Go of audio and metadata files. Table 1 gives the number of files and audio duration for each book category.

Audio files: It is worth mentioning that 388 audiobooks were directly hosted by *audiocite.net* while the 3,990 remaining were hosted on *archive.org* instances. Contrary to the guidelines, recordings may be either mono or stereo, and variations in bit rates or sampling rates can also be observed. Some recordings may contain multiple speakers (one after the other), background noise or music. Furthermore, not all recordings necessarily involve the reading of published books; some are articles or podcasts.

Metadata: Alongside the audio files, metadata information were downloaded such as the duration of each audio file, the speaker's ID, the title of the book read, the author of the book, the book category and the audio-related license. These pieces of information were provided by the speakers themselves on each audiobook page of *audiocite.net*.

3. Audiocite.net organization

Although one of the main uses considered for this dataset is self-supervised learning, we anticipate other usages as well. Indeed, even if it is not transcribed, the dataset could be used for topic modeling, signal reconstruction or speech synthesis. This is why we released it with official partitions and easy-to-query metadata files.

3.1. Speaker Gender estimation

To minimize biases in the partitions, we inferred the gender of the speaker based on the speaker's ID or by listening to the voice of the speaker. This information has been added to the metadata. However, we do not guarantee the information is reliable, nor do we guarantee that the method used is viable for deducing a person's gender as it isn't based on the speaker's self-identification. Gender metadata should be treated with caution.

3.2. Data partitions

The dataset was split into three partitions : a training (train) set comprising 80% of the total duration, a development (dev) set consisting of 10%, and a test set also containing 10%. Table 2 gives the duration statistics by gender for each subset.

The development and test sets were carefully curated not to include content that could be considered sensitive, specifically those falling under the *charmes* (erotic), *planete-actuelle* (geopolitics), and religion categories. Additionally, for these sets, equal representation of male and female speech was ensured and files with unknown speaker gender were not included in these partitions. Table 1 gives the number of files and duration for each book category in the train, dev and test sets.

Category	# Files				Duration (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
All	**28 485**	23 808	2 332	2345	**6 682:00:18**	5 290:32:24	690:56:22	700:31:31

Table 1: Statistics (durations and number of files) by book category with partitioning details (**all**, train / dev / test)

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

Table 2: Statistics of Audiocite.net corpus - Number of files, speakers and duration per gender (all, female, male and unknown) per partitions

3.3. Dataset organisation

The dataset is organized as follows: we share a README and a datasheet (inspired by *Datasheet for datasets*, (Gebru et al., 2021)) where extended statistics, as well as details on the composition, use and distribution of the corpus can be found, and three folders (`wavs/`, `scripts/` and `metadata/`).

In the `wavs/` folder, audiobooks files are arranged in folders according to the title of the book read, sorted by alphabetical order.

We also provide a `scripts/` folder with scripts for generating statistics on the corpus and the json files provided.

Regarding the `metadata/` folder, two types of metadata files are shared with the dataset: *.csv* and *.json*.

download.csv file: Each audiobook has an entry in the csv file. We also give details such as the speaker's ID, the title of the book read, the author of the book, the genre of the book, the license of the recording, the url address of the audiobook on *audiocite.net* and the path to the audio file in the *wavs* folder.

***.json files:** We provide four json files: *train.json*, *dev.json*, *test.json* and *all.json* which is a concatenation of the three previous ones. In those files, one entry corresponds to an audio file (one audiobook can contain many audio files). Those files contain the speaker ID, the duration of the audio file in seconds, the path to the file in the *wavs* folder and the gender of the speaker (F/M/U). A json entry takes the following form:

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

4. Impact of Audiocite.net on speech processing tasks

The dataset collected was shared with the LeBenchmark team to train their 14k models (Parcollet et al., 2024). In this section, we compared

the 14k model performance against the 7k one, which was not trained on the corpus. We report Word Error Rates (WER) for ASR systems on a dataset of the same type of speech and situation (audiobook), but also ASR and speaker verification results from Parcollet et al. (2024) to give broader results.

4.1. LeBenchmark models

7k-large model: This model was trained on 7,000 h of speech including 1,626 h of radio broadcast, 1,115 h of read speech, 127 h of spontaneous speech, 38 h of acted telephone dialogue and 29 h of acted emotional speech.
14k-large model: This model was trained on 14,000 h of speech: the 7k model data plus the full Audiocite.net data (all partitions), and 111 h of radio broadcasts speech.

4.2. ASR experiment

We based our ASR systems on Speechbrain's Common Voice ASR CTC recipe[1], that we adapted to compose a system with LeBenchmark (Wav2Vec2) encoder followed by a BiLSTM layer and a final DNN layer.

We used the official splits for the French part of the Multilingual Librispeech (MLS) corpus, allocating 1,076.58 h for training, 10.07 h for development and 10.07 h for testing.

Two scenarios were used in the experiment: (1) LeBenchmark encoder frozen (i.e. no fine-tuning) or (2) unfrozen (i.e. with fine-tuning alongside the ASR training).

For the ASR training, the learning rates were initialized at 0.001 for the Wav2Vec2 model when unfrozen and 0.1 for the BiLSTM+DNN part, and annealing was used with a 0.9 and 0.8 factor respectively. We used warmup of the pre-trained model which means it was not updated for 500 steps when fine-tuned. Train batch size was 8 and output layer dimension was 43 (number of characters in the training set).

Model	WER (%) ↓	
	Frozen	**F-T**
7k-large	31.71	9.56
14k-large	9.96	9.98

Table 3: ASR WER results (%) on test set of the French part of MLS corpus for Frozen and Fine-Tuned (F-T) LeBenchmark acoustic models

Table 3 summarizes the results of the experiment. Fine-tuning of the encoder part leads very quickly to similar good performances for 14k and 7k models, indicating that Audiocite.net did not make a big difference when fine-tuning is considered. However, for the Frozen experiment, there is a large superiority for the 14k model, indicating that Audiocite.net did play an important role in modeling read speech.

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

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

Table 4: ASR WER (%) results from Parcollet et al. (2024) on Common Voice 6.1 and ETAPE test sets, with Wav2Vec2.0 models further fine-tuned on labeled ASR data.

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

Table 5: Speaker verification task results from Parcollet et al. (2024) on Fabiole Corpus. EER: Equal Error Rate, minDCF: Minimum of the Detection Cost Function

4.3. LeBenchmark experiments

Among the different experiments performed by the LeBenchmark team we report in tables 4 and 5 the ASR and speaker verification tasks results from (Parcollet et al., 2024).

As can be seen on ASR experiments with Common Voice and ETAPE, Audiocite.net (14K-large) brings no improvement over the 7K model. It is even degraded on ETAPE which is composed of broadcast speech, a kind of speech very different from Audiocite.net. However, in a speaker verification task on the Fabiole corpus (Ajili et al., 2016), Audiocite.net (14K-large) brings an definite improvement over the 7K model. Its seems that by adding more speakers in the 14K, such modeling was improved.

5. Conclusion

In this paper we introduce the Audiocite.net corpus composed of more than 6,600 hours of recordings from 130 speakers and that can be found distributed on OpenSLR (www.openslr.org/139/) with the same license as the audiobook (i.e. Creative commons). All the recordings are distributed in their raw format as we downloaded them from *audiocite.net* (with background music, noise, unexpected speakers, mp3 format, mono or stereo). No pre-processing was applied to the files or automatic transcription was performed on them. However, we added gender information in a 'best effort'

manner, by guessing the gender from the name and checking the voice in case of uncertainty. This Audiocite.net was used to learn the 14k models of LeBenchmark project which demonstrates it superiority but also its limits in several speech processing downstream tasks.

6. Acknowledgements

This work was partially supported by MIAI@Grenoble-Alpes (ANR-19-P3IA-0003), E-SSL project (ANR-22-CE23-0013) and the Banque Publique d'Investissement (BPI) under grant agreement THERADIA. The authors gratefully acknowledge William Havard for the original idea, Marcely Zanon Boito and Fabien Ringeval for their help in the first attempt of this work.

7. Ethical Considerations

The books selected by the readers are either exclusively licensed under Creative Commons (CC) or obtained through written distribution agreements from the authors. After reading, the speakers chose a second Creative Commons license for the audio before publishing their recordings on the *audiocite.net* website. This second license carried accordingly equal or greater restrictions than the one assigned to the original book. By uploading their recordings to the platform, readers were aware that their material might be used for various purposes beyond the original intent within the limits of the license granted.

Concerning the audio's content, all kinds of assertions can be found in it and we do not want to encourage anyone to develop any position of any kind. It is also necessary to recall that the gender information was inferred from the names of the authors, with verification upon listening in case of ambiguity. We acknowledge this method is not accurate enough for speech processing, but it was intended only to lead to a fairer distribution of data in each partition. We also commit to deleting the recording, and its metadata, and updating the corpus if a speaker requests to remove their data from the dataset without explicit reason.

The *audiocite.net* website administrator has expressly granted us permission to use and distribute the audio in accordance with their terms of use.

8. References

8.1. Bibliographical References

Mohamed Abdel-Rahman, Lee Hung-yi, Borgholt Lasse, D. Havtorn Jakob, Edin Joakim, Igel Christian, Kirchhoff Katrin, Li Shang-Wen, Livescu Karen, Maaløe Lars, N. Sainath Tara, and Watanabe Shinji. 2022. Self-supervised speech representation learning: A review. *IEEE JSTSP Special Issue on Self-Supervised Learning for Speech and Audio Processing*.

Moez Ajili, Jean-François Bonastre, Juliette Kahn, Solange Rossato, and Guillaume Bernard. 2016. Fabiole, a speech database for forensic speaker comparison. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 726–733.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 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.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *proceedings of NeurIPS*, Vancouver, Canada.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing*, 16:1–14.

Yannick Estève, Thierry Bazillon, Jean-Yves Antoine, Frédéric Béchet, and Jérôme Farinas. 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).

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM*, 64(12):86–92.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 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.

Dongwei Jiang, Wubo Li, Miao Cao, Wei Zou, and Xiangang Li. 2021. Speech SimCLR: Combining Contrastive and Reconstruction Objective for Self-Supervised Speech Representation Learning. In *proceedings of Interspeech*.

Andy Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee. 2020. Mockingjay: Unsupervised Speech Representation Learning with Deep Bidirectional Transformer Encoders. In *proceedings of ICASSP*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation Learning with Contrastive Predictive Coding. ArXiv:1807.03748.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Titouan Parcollet, Ha Nguyen, Solène Evain, Marcely Zanon Boito, Adrien Pupier, Salima Mdhaffar, Hang Le, Sina Alisamir, Natalia Tomashenko, Marco Dinarelli, Shucong Zhang, Alexandre Allauzen, Maximin Coavoux, Yannick Estève, Mickael Rouvier, Jerôme Goulian, Benjamin Lecouteux, François Portet, Solange Rossato, Fabien Ringeval, Didier Schwab, and Laurent Besacier. 2024. Lebenchmark 2.0: A standardized, replicable and enhanced framework for self-supervised representations of french speech. *Computer Speech & Language*, 86:101622.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A large-scale multilingual dataset for speech research. In *INTERSPEECH*, Shanghai, China.

Mirco Ravanelli, Jianyuan Zhong, Santiago Pascual, Pawel Swietojanski, João Monteiro Filho, Jan Trmal, and Y. Bengio. 2020. Multi-Task Self-Supervised Learning for Robust Speech Recognition. In *proceedings of ICASSP*.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 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.

8.2. Language Resource References

Yannick Estève and others. 2010. *EPAC Corpus: orthographic transcriptions*. distributed via ELRA: ELRA-S0305, ISLRN 483-703-007-740-8.

Vineel Pratap and Qiantong Xu and Anuroop Sriram and Gabriel Synnaeve and Ronan Collobert. 2020. *Multilingual LibriSpeech (MLS)*. distributed by OpenSLR. PID https://www.openslr.org/94/.

Changhan Wang and others. 2021. *VoxPopuli*. distributed by Facebook. PID https://github.com/facebookresearch/voxpopuli.