Speech Technologies with Fieldwork Recordings : the Case of Haitian Creole

William N Havard, Renauld Govain, Benjamin Lecouteux, Emmanuel Schang

Speech Technologies with Fieldwork Recordings: the Case of Haitian Creole

William N. Havard[1,2], **Renauld Govain**[3], **Benjamin Lecouteux**[2], **Emmanuel Schang**[1]

[1] LLL, Université d'Orléans, CNRS, 45000 Orléans, France
[2] LIG, Université Grenoble Alpes, CNRS, Grenoble INP, 38000 Grenoble, France
[3] LangSé, Université d'État d'Haïti, Port-au-Prince, Haïti

william.havard@univ-orleans.fr

Abstract

We use 40-year-old digitalised tape-recorded fieldwork data in Haitian Creole to train a *native* self-supervised learning (SSL) model of speech representation (WAV2VEC 2). We also use a continued pre-training approach on pre-trained SSL models of two *foreign* languages: the lexifier language – French – and an unrelated language – English. We compare the performances of these three SSL models, and of two other foreign SSL models directly fine-tuned, on an ASR task, where all five models are fine-tuned on transcribed fieldwork recordings in Haitian Creole. Our results show the best-performing model is the one trained using a continued pre-training approach on the lexifier language, followed by the native model. We conclude that the 'mobilising the archive'-approach advocated by (Bird, 2020) is a promising way forward to design speech technologies for new languages.

1 Introduction

Most of the so-called low-resourced languages are often low-resourced from the perspective of computer scientists only: they often have many resources that were collected over the years by linguists, missionaries, and generally by the community of speakers itself (Bird, 2020). The data is often not readily accessible (i.e. in a digitalised format), but existent nonetheless. The question we aim to answer in this paper is the following: how far can we go with state-of-the-art speech processing models using *only* fieldwork data? By 'fieldwork data', we mean data that was *not* originally collected to serve as training data for computational applications (e.g. Automatic Speech Recognition, ASR), but was collected for linguistic purposes (e.g. to study dialectal variation). In this paper, we focus on spoken data in Haitian Creole, consisting of recorded interviews between linguists and their collaborators. Haitian Creole is a French-based Creole (i.e. French is called its lexifier language, the language that gave Haitian Creole most of its vocabulary (Hazael-Massieux, 2012)), spoken by 13M speakers (Simons and Fennig, 2023) in Haiti and by the Haitian diaspora.

Most of the data we use in this paper (see Section 2) was collected 40 years ago with tape recorders to study dialectal variation in Haitian, with a focus on lexical variations. Contrary to the clean audiobooks commonly used to train neural models (e.g. Librispeech, (Panayotov et al., 2015)), the data we used is inherently noisy: reverberated, echo-y, full of environmental noise (e.g. chickens, cars, passers-by, etc.). Yet, this type of data represents the majority of the data available for most of the world languages. As collecting and transcribing data is a costly process (Himmelmann, 2018), is it possible to make use — as advocated by (Bird, 2020) in the 'mobilising the archive'-approach — of already existing (and potentially old) fieldwork data and re-purpose them for computational applications?

1.1 Related Works

The field of speech processing for Creole languages is relatively sparse, except for the work of (Breiter, 2014) for Haitian Creole, that of (Macaire et al., 2022; Le Ferrand et al., 2023; Le Ferrand and Prud'hommeaux, 2024) for Guadeloupean and Mauritian Creole, and (Gooda Sahib-Kaudeer et al., 2019) for Mauritian Creole (with a focus on the medical domain). Hence, speech processing for Creole languages — whatever the lexifier language, be it French, English, Portuguese, etc. — remains largely unexplored.

Unrelated to Creole languages — but related to our experimental settings — (Nowakowski et al., 2023) explored continuous pre-training (CPT) approaches, followed by an ASR fine-tuning task for Ainu speech recognition using old fieldwork data. In short, CPT is a form of transfer learning which

consists in using large quantities of unlabelled data (i.e. raw speech) to continue to pre-train models that were already pre-trained on another language. However (Nowakowski et al., 2023) do not train their models 'on a budget' as (i) they use 4 GPUs and (ii) use the XLSR-53 model (Conneau et al., 2021) which is based on WAV2VEC 2-LARGE and pre-trained on 56k hours of data, and (iii) use multilingual fine-tuning by which the ASR model is not only trained on the target language (Ainu), but on several languages at once (English, Japanese, alongside Ainu). We aim for a stricter approach that only uses fieldwork data at all steps.

1.2 Research Questions

In this work, **we *only* assume the existence of (potentially old) fieldwork data** to train the models, which corresponds to several real-world use cases: that of field linguists documenting a language and that have gathered a certain amount of both untranscribed and transcribed recordings (our case), or that of a community of speakers that uses archival material to build models for their language.

More precisely, the questions we tackle in this paper are the following: **(a)** Would noisy, but ecologically valid, fieldwork data be usable to train self-supervised learning (SSL) models of speech (e.g. WAV2VEC 2, (Baevski et al., 2020))? **(b)** Should said models be trained from scratch or should continued pre-training (CPT) (Gururangan et al., 2020; Nowakowski et al., 2023) approaches be used? **(c)** How much training data is necessary to fine-tune the models on an ASR task? And finally, **(d)** given our use-cases, is it possible to train such models 'on a budget'? (i.e. only 1 GPU, as having more – e.g. 64 as (Baevski et al., 2020) – is generally impossible for laypeople).

Additionally, as we work in the context of Creole languages, we also aim to explore the influence of the lexifier language (as a clear case of related languages) and **(e)** whether CPT be done on SSL models of the lexifier language (e.g. French in the case of Haitian Creole), or do models trained on an unrelated language (e.g. say English in the case of Haitian Creole) also work?

2 Data

ALH. We used the *Atlas Linguistique d'Haïti* (Fattier, 1998), which consists of a collection of 499 audio recordings in Haitian Creole (Kreyòl ayisyen) collected in Haiti between 1978 and 1987 for the purpose of creating a linguistic atlas. The recordings were originally done on audio cassettes with tape recorders which were then digitalised by the French National Library (*Bibliothèque Nationale de France*, BNF) in 2010. Each recording is on average 45 minutes long and features one or several interviewers eliciting words or phrases from their native collaborators. This data has been made publically available by the BNF and is accessible on the COCOON [1] platform. Although the recordings are associated with field notebooks containing partial handwritten transcriptions (phonetic transcription at word level), these have not been digitised (nor aligned with the recordings). As a result, this corpus consists entirely of raw speech. We partitioned the data set (356.3 hours) into 3 splits (train/val/test). The data was partitioned so that the validation set would contain at least 5 hours of data and a minimum of 2 unseen speakers, and the test set at least 5 hours of data and a minimum of 3 unseen speakers. We reached the following distribution which fulfilled our constraints: train = 345.6 hours; val = 5.3 hours, 5 unseen speakers; and test = 5.4 hours, 8 unseen speakers, the latter being left for future work.

CNCH. The *Corpus of Northern Haitian Creole*[2] (Valdman et al., 2015) consists of 10 recorded interviews, conducted in Cap-Haïtien (Northern Haiti) to study dialectal variation with regard to standard Haitian. This corpus was entirely transcribed by the linguist who collected it. However, the transcriptions used are non-standard and impressionistic, in the sense that spelling variations deviating from the norm are used to transcribe the speaker's pronunciation more faithfully: "*Powoprens*"/"*Potoprens*", Port-au-Prince, the capital city of Haiti; "*eskeu*"/"*eske*", question words; "*deu*"/"*de*", two; etc.). We partitioned the data set (9.0 hours) into 3 splits (train/val/test). The data was partitioned so that the val set would contain at least 1 hour of data and a minimum of 1 unseen speaker, and the test set at least 1 hour of data and a minimum of 1 unseen speaker. We reached the following distribution which fulfilled our constraints: train = 6.9 hours; val = 1.1 hours, 1 unseen speaker; test = 1.0 hours, 2 unseen speakers.

Other data sets. The two previous data sets are the ***only*** publicly available data sets of fieldwork recordings in Haitian Creole. We however wish to

[1] https://cocoon.huma-num.fr/
[2] https://archive.org/details/interview-8-ujf-107-a-ujm-107-a

acknowledge the existence of other data sets featuring speech in Haitian Creole, which we purposefully excluded during the training phase as they do not consist of fieldwork data: the freely accessible Haiti-CMU data set[3] which features read speech (∼ 20 hours), mainly from sections of the Bible, which do not reflect everyday language use; and the proprietary IARPA-Babel data set consisting of "203 hours of Haitian Creole conversational and scripted telephone speech" (Andrus et al., 2017). We use both data sets to test our models on out-of-domain data and compare them with Facebook's MMS model (Pratap et al., 2023). For Haiti-CMU, we generated a test set that consists of 2 hours of data by randomly sampling recordings; and for IARPA-Babel we used the development set as a test set (as it is commonly done with IARPA-Babel data sets, as the evaluation set was kept private), which consists of 20 hours of data.

3 Experimental Settings

Given our low-budget setting, we focus on the WAV2VEC 2-BASE architecture, thus excluding fine-tuning a multilingual model such as XLSR-53 which is based on the LARGE architecture.

3.1 Native and Foreign-SSL Pre-Training

We use the ALH corpus to pre-train our SSL models. A voice activity detection model (Pyannote (Bredin et al., 2020)) was used to isolate sections corresponding to speech from surrounding noises, resulting in 229h of spoken sections. The resulting segments were rather short (∼ 2.3s) and unsuited to pre-train SSL models as-is. Thus, we merged them until the resulting concatenated segments reached 19s on average (19.4s ±5.8). The WAV2VEC 2 models were trained on a single GPU[4] using gradient accumulation for 16 steps (to simulate 16 GPUs) with 16-bits floats and a maximum batch size of 5.2 minutes. All the models were implemented using fairseq's standard WAV2VEC 2-BASE implementation and training pipeline (Ott et al., 2019). Three models were trained:

- One model pre-trained from scratch (i.e. not based on any existing pre-trained model):
  - NATIVE -HAT-SSL+Ø : this model was pre-trained on the ALH data and has never been exposed to any other language other than Haitian (HAT) throughout pre-training;
- Two models pre-trained using a continued pre-training approach:
  - FOREIGN -FRA-SSL+C PT: the base model was pre-trained on a French (FRA) (wav2vec2-FR-7K-base, pre-trained on 7k hours in French (Parcollet et al., 2023)) and was continued pre-trained on the ALH data;
  - FOREIGN -ENG-SSL+C PT: the base model was pre-trained on English (ENG) (wav2vec2-base pre-trained on Librispeech 960 (Baevski et al., 2020)) and was continued pre-trained on the ALH data.

3.2 ASR fine-tuning

We fine-tuned our pre-trained and continued pre-trained models on the CNCH data set. 5 models were fine-tuned:

- Three models from models that had seen Haitian speech in a (continued) pre-training phase:
  - NATIVE -HAT-SSL+Ø +FT: where NATIVE -HAT-SSL+Ø was fine-tuned after the pre-training phase on ALH;
  - FOREIGN -FRA-SSL+C PT+FT: where FOREIGN -FRA-SSL+C PT was fine-tuned after the continued pre-training phase on ALH;
  - FOREIGN -ENG-SSL+C PT+FT: where FOREIGN -ENG-SSL+C PT was fine-tuned after the continued pre-training phase on ALH;
- Two models from models that hadn't seen any Haitian speech before being fine-tuned:
  - FOREIGN -FRA-SSL+Ø +FT: where the French (wav2vec2-FR-7K-base) was directly fine-tuned.
  - FOREIGN -ENG-SSL+Ø +FT: where the English (wav2vec2-base) was directly fine-tuned.

In order to understand the impact of the training size on the final performance of the models, we use different train sizes: max (360minutes), 320

[3] http://www.speech.cs.cmu.edu/haitian/

[4] 32Gb Nvidia Tesla V100 or 45Gb Nvidia A40 depending on availability.

Table 1: Configurations that yield the best performances in terms of WER (left) and CER (right) for each type of fine-tuned model. *Rank* shows the models' rank(from 1/best to 200/worst) when WER/CER is used as sorting key.

Model Type	**WER** ↓	CER ↓	Train Size	Decoding	Rank
FOREIGN -FRA-SSL+C PT+FT	**36.8**	21.6	320	4-gram	**1**
NATIVE -HAT-SSL+Ø +F T	37.4	**21.5**	360 (max)	3-gram	5
FOREIGN -ENG-SSL+C PT+FT	37.5	22.4	320	4-gram	6
FOREIGN -FRA-SSL+Ø +F T	42.5	24.5	360 (max)	3-gram	27
FOREIGN -ENG-SSL+Ø +F T	50.4	29.0	320	3-gram	49

Model Type	WER ↓	**CER** ↓	Train Size	Decoding	Rank
FOREIGN -FRA-SSL+C PT+FT	**38.2**	**17.1**	320	Viterbi	**1**
NATIVE -HAT-SSL+Ø +F T	39.8	17.8	360 (max)	Viterbi	3
FOREIGN -ENG-SSL+C PT+FT	40.3	18.6	360 (max)	Viterbi	6
FOREIGN -FRA-SSL+Ø +F T	46.2	21.7	360 (max)	Viterbi	12
FOREIGN -ENG-SSL+Ø +F T	57.1	26.6	360 (max)	Viterbi	38

Table 2: Comparison between Facebook's MMS (Pratap et al., 2023), our best-performing model (FOREIGN -FRA-SSL+Ø +F T) and a native model (NATIVE -HAT-SSL+Ø +F T). For a fair comparison between models, only Viterbi decoding was used. Note that MMS was pre-trained on the IARPA-Babel data.

Corpus	Model	CER↓
CNCH	FOREIGN -FRA-SSL+C PT+FT	**17.1**
	NATIVE -HAT-SSL+Ø +F T	17.8
	MMS (Pratap et al., 2023)	28.4
Haiti-CMU	FOREIGN -FRA-SSL+C PT+FT	09.5
	NATIVE -HAT-SSL+Ø +F T	11.6
	MMS (Pratap et al., 2023)	**07.9**
IARPA-Babel	FOREIGN -FRA-SSL+C PT+FT	36.6
	NATIVE -HAT-SSL+Ø +F T	38.5
	MMS (Pratap et al., 2023)	**34.6**

160, 80, 40, 20, 10, and 5 minutes. Each train size including the previous sizes (e.g. max ⊇ ... ⊇ 10 ⊇ 5). Each model is fine-tuned for 20k steps[5] with a CTC loss and the best model is selected on the lowest WER on the validation set. To prevent overfitting, the parameters were frozen for the first 10k steps. The text was lower-cased and diacritics removed (due to inconsistent use).

Finally, we also train 2-to-5-gram LMs using KenLM (Heafield, 2011), with default Kneser-Ney discounting parameters. LMs were trained on the transcriptions of the CNCH data set only (hence, preserving our 'fieldwork data'-only setting). We trained a separate LM for each size of the training data set (e.g. a LM trained on train-10 only uses the text corresponding to the transcription of 10 minutes of speech), resulting in 32 different LMs (4 n-gram sizes × 8 train sizes) that will be used to compare raw (i.e. Viterbi) decoding and LM-rescored decodings.

4 Results & Discussion

Results. We used the SCTK toolkit [6] to compute standard Word Error Rate (WER) and Character Error Rate (CER). Standard Viterbi decoding, and LM rescoring with 2-to-5-gram LMs was used. This resulted in 5 fine-tuned models × 8 training sizes × (1 Viterbi + 4 ngram) decoding = 200 decoding strategies. A general overview of our results is shown in Fig.1 (for clarity, only Viterbi, and 5-gram LM rescoring is shown) and the best configuration for each of the 5 model types is shown in Tab. 1. A performance comparison between Facebook's MMS and our models is shown in Tab. 2.

Using Fieldwork Data. Turning back to our original research questions, our results show that **(d)** it is possible to train competitive models on a budget using a single GPU and that **(a)** using fieldwork data to train SSL models of speech is effective. Despite such data being inherently noisy — as opposed to audiobooks or broadcast speech commonly used to train SSL models — the NATIVE -HAT-SSL+Ø Haitian model we trained remained very competitive compared to other approaches. This is particularly interesting in the case of low-resourced languages, such as are most of the French-based Creole languages spoken in the Caribbean (Haitian, Guadeloupean, Saint Lucian, etc.) or in South America (Guianan). This means that *no new data needs to be collected*, but that old tape-recorded fieldwork data, once digitalised, can be repurposed for this matter. This opens an avenue for many languages of the world to have cutting-edge speech processing models at their disposal.

Train From Scratch or Use CPT. Now, turning to whether we should fine-tune SSL models that have been pre-trained from scratch or models pre-trained using a CPT approach **(b)**, our results show that the CPT models show a slight advantage over native models trained from scratch (−1.6 WER points, and −0.7 CER points, Viterbi decoding, using lowest CER as sorting key). However, our result show that **(e)** this advantage is only true when the model used for continued finetuning is *that of the lexifier language* (here, French). This advantage seems to disappear when it is not the case, as

[5] Given how little data we have, the models quickly converge and remain stable and do not evolve after 20k steps, hence this cutoff value.

[6] https://github.com/usnistgov/SCTK

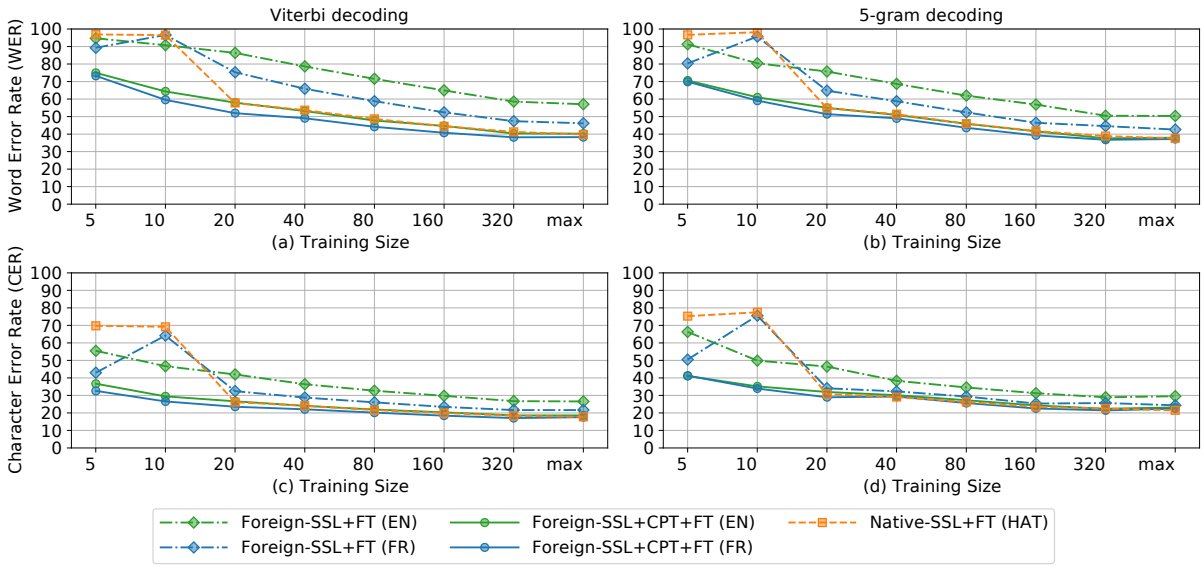

Figure 1: (a, b) Word Error Rate (WER) and (c, d) Character Error Rate (CER, at the character level) of fine-tuned models on an ASR task with Viterbi decoding (left) and with 5-gram LM (right) as a function of the amount of CNCH data used for training (in minutes, from 5 to *max*, where *max* = 6.9 hours, ∼ 360minutes).

the model fine-tuned from another language (here, English) has generally worse performances than either a model fine-tuned from the lexifier language (+2.1 WER, +1.5 CER, *id.*) or from the native language (+0.5 WER, +0.8 CER, *id.*). However, what seems most critical is the CPT approach. The ASR models directly fine-tuned from SSL models (FOREIGN -SSL+Ø +F T) that have *not* seen any Haitian speech in a CPT setting lag far behind (+8 WER, +4.6 CER for the French-based models, *id.*) or very far behind the best model (+18.9 WER, +9.5 CER for the English-based models, *id.*).

Amount of Fine-tuning Data. Turning to **(c)** and the amount of data necessary to fine-tune SSL models on an ASR task, our results show a marked difference between three groups of models: (i) FOREIGN -SSL+C PT+FT very robust to a reduced amount of training data, (ii) F OREIGN -SSL+Ø +FT not very robust to a reduced amount of data, and (iii) N ATIVE -HAT-SSL+Ø showing in between results. Using 20 minutes of data closes the gap between (i) and (iii) while models in group (ii) required approximately 4 times this amount of data (80 minutes) to reach similar performances. We hypothesise that models in group (i) benefit from having seen more speech altogether, as they were pre-trained in their respective language (French or English), have seen Haitian data in the CPT phase, and were further fine-tuned, which could explain why they are more robust.

Viterbi or LM Decoding. Finally, we observed mixed results with the use of LMs for decoding. While they do not significantly improve (nor hurt) the NATIVE -HAT-SSL+Ø +F T or FOREIGN -SSL+C PT+FT models, they significantly improved the WER scores of the F OREIGN -SSL+Ø +F T (Fig. 1a and 1b): e.g. −10 WER with a 5-gram LM for FOREIGN -ENG-SSL+Ø +F T model fine-tuned with 40 minutes of data. Hence, when no pre-training data is available and that foreign models can only be directly fine-tuned, using LM-rescoring is indispensable. However, it seems that using LMs, while improving WER scores, comes at the expense of higher CERs (Fig. 1c and 1d); which hints at the fact that while there are more words accurately transcribed, the others are less well transcribed resulting in a higher CERs.

Comparison with MMS. Tab. 2 shows a comparison of the performances between our models and Facebook's MMS (Pratap et al., 2023) model with the Haitian adapter. To ensure a fair comparison, only Viterbi decoding was used. MMS obtains better scores (−1.6CER for Haiti-CMU, and −2 CER for IARPA-Babel) compared to our best-performing model FOREIGN -FRA-SSL+C PT+FT (though, the comparison is not entirely fair, as MMS was pre-trained on the IARPA-Babel data). However, both FOREIGN -FRA-SSL+C PT+FT and NATIVE -HAT-SSL+Ø +F T obtain better CERs than MMS on fieldwork data (−11.3and −10.6 respectively). This shows that our models are very competitive compared to MMS, particularly given

the fact that MMS was pre-trained on 491k hours of data and fine-tuned 44.7k hours of labelled data (including roughly 20 hours of Haitian). In contrast, our models are pre-trained on 340 hours of data and fined-tune on less than 6 hours of data. It also shows that using fieldwork recordings does not hinder zero-shot adaptation to out-of-domain (i.e. non-fieldwork) data, contrary to MMS which performs much worse on out-of-domain fieldwork data.

5 Limitations and Future Work

In this paper, we focused on exploring the validity of using fieldwork data to pre-train self-supervised models to ultimately fine-tune ASR models from them (*ex*trinsic evaluation), but have left aside the study of the pre-trained models and representations themselves (*in*trinsic evaluation). In future works, we wish to use an ABX task (Schatz et al., 2013) to compare the latent representations and their transfer at the phoneme level. This would help us gain more insight into the performances of our models. The data we use for continued pre-trained was collected 40 years ago, and the language between that time and now has evolved (e.g. its phonology, etc.). Hence, the question of the impact of the diachronic shift and how to measure it is open. Finally, our results show that 350 hours of fieldwork recordings is enough to pre-train a native SSL model and obtain competitive results when fine-tuned on an ASR task. Yet, such a treasure trove with as many recording hours might not exist for all languages: the question of the minimal amount of fieldwork data to use is open.

6 Conclusion

In this work, we used 40-years old digitalised tape-recorded fieldwork data in Haitian to train SSL models. We trained a native SSL model, and also used a CPT approach on pre-trained SSL models of the lexifier language (French) and of an unrelated language (English), which we fine-tuned on another data set of fieldwork recordings on an ASR task. We obtained competitive results and showed that the best model is the pre-trained model of the lexifier language with CPT on Haitian fieldwork recordings, followed by the native SSL model, obtaining close results. Hence, when no model of the lexifier language is available, it is still worth training a native model with fieldwork data. Being able to train a native model is all the most important, as a native model might be a matter of self-pride to the speaker community, as opposed to a model derived from the lexifier language, generally that of the former colonising power.

Contrary to the work of (Nowakowski et al., 2023), ours is the first that demonstrates the feasibility of training SSL models using ***only*** fieldwork recordings, and their usability on downstream tasks, such as ASR. This methodology opens an avenue for many languages of the world to have cutting-edge speech-processing models at their disposal, by digitalising recordings collected decades ago. Hence, the 'mobilising the archive'-approach advocated by (Bird, 2020) constitutes a promising way forward.

The best-performing foreign & native models will be made public, along with the scripts used to format the data.

References

Tony Andrus, Aric Bills, Thomas Conners, Erin Smith Crabb, Eyal Dubinski, Jonathan G. Fiscus, Breanna Gillies, Mary Harper, T. J. Hazen, Brook Hefright, Amy Jarrett, Hanh Le, Jessica Ray, Anton Rytting, Wade Shen, Ronnie Silber, and Evelyne Tzoukermann. 2017. Iarpa babel haitian creole language pack iarpa-babel201b-v0.2b.

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 the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2020. pyannote.audio: neural building blocks for speaker diarization. In *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain.

Wojtek Breiter. 2014. Rapid bootstrapping of haitian creole large vocabulary continuous speech recognition.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In *Proc. Interspeech 2021*, pages 2426–2430.

Dominique Fattier. 1998. *Contribution à l'étude de la genèse d'un créole: l'Atlas linguistique d'Haïti, cartes et commentaires, 6 vol.* Bibliographical record, Presses Universitaires du Septentrion, Villeneuve d'Ascq. Ph.D. Dissertation, Université de Provence.

Nuzhah Gooda Sahib-Kaudeer, Baby Gobin-Rahimbux, Bibi Saamiyah Bahsu, and Maryam Farheen Aasiyah Maghoo. 2019. Automatic speech recognition for kreol morisien: A case study for the health domain. In *Speech and Computer*, pages 414–422, Cham. Springer International Publishing.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 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*, pages 8342–8360, Online. Association for Computational Linguistics.

Marie-Christine Hazael-Massieux. 2012. *Les Créoles à base française*. Editions Ophrys, Gap, France.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Nikolaus P. Himmelmann. 2018. *Meeting the transcription challenge*. University of Hawai'i Press.

Eric Le Ferrand, Claudel Pierre-Louis, Ruoran Dong, Benjamin Lecouteux, Daphné Gonçalves-Teixeira, William N Havard, and Emmanuel Schang. 2023. Outiller la documentation des langues créoles. In *LIFT 2023 : journées scientifiques du GdR Linguistique Informatique, Formelle et de Terrain*, Vandoeuvre-Lès-Nancy, France.

Éric Le Ferrand and Emily Prud'hommeaux. 2024. Automatic transcription of grammaticality judgements for language documentation. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 33–38, St. Julians, Malta. Association for Computational Linguistics.

Cécile Macaire, Didier Schwab, Benjamin Lecouteux, and Emmanuel Schang. 2022. Automatic speech recognition and query by example for creole languages documentation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2512–2520, Dublin, Ireland. Association for Computational Linguistics.

Karol Nowakowski, Michal Ptaszynski, Kyoko Murasaki, and Jagna Nieuwa żny. 2023. Adaptation of a multilingual speech representation model for a new, underresourced language via multilingual fine-tuning and continued pretraining. *Science Talks*, 8:100249.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

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)*, pages 5206–5210.

Titouan Parcollet, Ha Nguyen, Solene Evain, Marcely Zanon Boito, Adrien Pupier, Salima Mdhaffar, Hang Le, Sina Alisamir, Natalia Tomashenko, Marco Dinarelli, Shucong Zhang, Alexandre Allauzen, Maximin Coavoux, Yannick Esteve, Mickael Rouvier, Jerome Goulian, Benjamin Lecouteux, Francois Portet, Solange Rossato, Fabien Ringeval, Didier Schwab, and Laurent Besacier. 2023. Lebenchmark 2.0: a standardized, replicable and enhanced framework for self-supervised representations of french speech. *Preprint*, arXiv:2309.05472.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling speech technology to 1,000+ languages. *arXiv*.

Thomas Schatz, Vijayaditya Peddinti, Francis Bach, Aren Jansen, Hynek Hermansky, and Emmanuel Dupoux. 2013. Evaluating speech features with the minimal-pair ABX task: analysis of the classical MFC/PLP pipeline. In *Proc. Interspeech 2013*, pages 1781–1785.

Gary F Simons and Charles D Fennig, editors. 2023. *Ethnologue: Languages of the world*. Summer Institute of Linguistics, Academic Pub.

Albert Valdman, Anne-José Villeneuve, and Jason F. Siegel. 2015. On the influence of the standard norm of haitian creole on the cap haïtien dialect: Evidence from sociolinguistic variation in the third person singular pronoun. *Journal of Pidgin and Creole Languages*, 30(1):1–43.