Application of Speech Processes for the Documentation of Kréyòl Gwadloupéyen

Éric Le Ferrand[1,2], **Fabiola Henri**[3], **Benjamin Lecouteux**[2], **Emmanuel Schang**[1]
[1]LLL, Université d'Orléans, [2]LIG, Université Grenoble Alpes, France
[3]University at Buffalo, USA

Abstract

In recent times, there has been a growing number of research studies focused on addressing the challenges posed by low-resource languages and the transcription bottleneck phenomenon. This phenomenon has driven the development of speech recognition methods to transcribe regional and Indigenous languages automatically. Although there is much talk about bridging the gap between speech technologies and field linguistics, there is a lack of documented efficient communication between NLP experts and documentary linguists. The models created for low-resource languages often remain within the confines of computer science departments, while documentary linguistics remain attached to traditional transcription workflows. This paper presents the early stage of a collaboration between NLP experts and field linguists, resulting in the successful transcription of Kréyòl Gwadloupéyen using speech recognition technology.

1 Introduction

The fields of descriptive and documentary linguistics concentrate on gathering information and describing language phenomena. This work is typically performed on small, Indigenous, and regional languages that have a limited number of speakers. The linguist's process typically involves recording raw speech, either spontaneous or elicited, transcribing the recordings, translating them, and conducting an analysis. In this pipeline, the transcription becomes the data, but transcribing raw speech is a time-consuming task and is often seen as a bottleneck when a large amount of speech is collected but only a small portion is used.

Speech technologies have been viewed as a solution to this bottleneck issue by automatically annotating raw speech collections. Regular automatic speech recognition (ASR) has proven to be challenging due to the lack of data available in most languages to train robust models. However, alternative methods, such as spoken term detection, phone recognition, and the use of universal models, offer new possibilities for collaboration between field linguists and NLP experts.

We present here an application of speech processing on raw field linguistics recordings in Kréyòl Gwadloupéyen. Our objective has two parts: firstly, to exhibit the capability of a wav2vec and CTC-based system for our target language, and secondly, to illustrate how the transcription output can be valuable and utilised by field linguists.

2 Background

2.1 Fieldwork technologies

In the past decade, there have been ongoing discussions about developing technology for the purpose of linguistic fieldwork (Gessler, 2022; Gauthier, 2018; Moeller, 2014). The main argument has been to adapt emerging technologies such as smartphones for fieldwork. The recent improvement of speech recognition for low-resource languages has also been seen as a way to mitigate the transcription bottleneck (Himmelmann, 1998) automatically transcribing large amount of untranscribed speech data (e.g. Foley et al., 2018; Shi et al., 2021; Adams et al., 2021).

Looking at the role of technologies in the current linguistics fieldwork workflow, only a few tools are still widely used (e.g. Boersma and Weenink, 1996; Wittenburg et al., 2006). The other projects involving tools design often end up discontinued (Bird et al., 2014; Gauthier et al., 2016) or stayed at the prototype stage (Lane et al., 2021; Le Ferrand et al., 2022; Bettinson and Bird, 2017). Leveraging speech technologies for scaling up language documentation has had limited impact as well, probably because of lack of data available for low-resource languages to build robust models (Gupta and Boulianne, 2020a,b).

Proceedings of the Second Workshop on NLP Applications to Field Linguistics, pages 17–22
May 6, 2023 ©2023 Association for Computational Linguistics

The recent expansion of speech recognition models based on wav2vec2.0 (Conneau et al., 2021) combined with CTC algorithms (e.g. Macaire et al., 2022) open new opportunities for low-resource languages. Such an architecture is not restricted by a language model and can produce tokens out of vocabulary.

2.2 Kréyòl gwadloupéyen

Kréyòl gwadloupéyen is spoken on Guadeloupe Island and in mainland France by approximately 800 000 speakers. Kréyòl gwadloupéyen was born in the colonial context from the contact between French settlers and African slaves in the French West Indies (see (Prudent, 1999), (Chaudenson, 2004) among others). It has historically been stigmatised and viewed as a "lesser" form of language compared to French, the language of the colonisers. In terms of language use, Kréyòl gwadloupéyen is the primary language of daily communication for a large part of the population of Guadeloupe, particularly in informal settings. French, on the other hand, is used in formal and official contexts, such as in schools, government institutions, and the media. In this context of diglossia (Jeannot-Fourcaud and Jno-Baptiste, 2008), code-mixing is frequent, which is an obvious challenge for ASR systems. In short, creole languages share most of their lexicon with the dominant language (the lexifier language), while their grammar is significantly different from the grammar of the lexifier. The origins of the grammatical differences might be a matter of debate (see (Mufwene, 1997; Velupillai, 2015) among others). To give only one example of the distance and similarities of French and Kréyòl gwadloupéyen, see (1):

(1) a. Jan pa sav palé kréyol
Jean NEG know speak creole
'Jean doesn't speak creole'

b. Jean ne sait pas parler créole
Jean NEG know NEG speak creole
'Jean doesn't speak creole'

The NSF-IRES 1952568: Experimental linguistics in the Caribbean seeks to provide students with an international experience conducting linguistic research on low-resource and under-described creole languages like Kréyòl gwadloupéyen. During this 5-7 weeks program, fellows investigate a linguistic phenomenon in Gwadloupéyen on the basis of raw data (spontaneous speech or directed interviews) they collect to contribute to the description and documentation of the language. As previously noted, one of biggest challenges for field linguists and even more so, for the NSF-IRES fellows, remains time invested with transcriptions. Often, these recordings are unexploited for lack of time, adding to the issue of under-description. Only 60min of the approximately 10 hours of recordings collected in 2022 was transcribed, and this only after the program had ended. Notwithstanding code switching/mixing, the fellows' unfamiliarity with the language's phonology made the transcription exercise arduous and lengthier.

3 Automations

3.1 Data

The ASR experiments are based on the work of Macaire et al. (2022), who used a 60-minute-long speech corpus of spontaneous speech in Kréyòl gwadloupéyen for training.

The testing data consist of several hours of raw, unsegmented, and untranscribed speech recorded during a 2022 fieldwork. The speech is spontaneous and sparse across the recording, with overlapping speech, laughters, silences, and random noises spread across the collection. The speech segments are also not necessarily in Kréyòl, and even if the limit between French and Kréyòl gwadloupéyen is not clear, some segments are clearly in French and even English. One 1-hour-long recording was selected, which, after some verification, contains a majority of segments in Kréyòl.

3.2 Preprocessing

Speech processing systems generally expect short utterances of clear speech, so the type of data described previously is not usable as is and needs to be preprocessed. Following the ideas of the sparse transcription model (Bird, 2020), we used *auditok*[1], a Voice Activity Detection tool, to filter out non-speech segments. This tool works in an unsupervised fashion, with detection based on the energy of the audio signal. Although more accurate VAD tools are available, auditok provides a good baseline for this preliminary study.

3.3 ASR and Self-supervised Learning

Self-supervised learning (SSL) is the task of learning powerful representations from huge unlabeled

[1]https://auditok.readthedocs.io/en/latest/

data to recognise and understand patterns from a less common problem. These models allows to improve performance on downstream tasks for ASR in low-resource contexts (Baevski et al., 2019; Kawakami et al., 2020). These work are based on the Wav2Vec2.0 (Baevski et al., 2020) model. It builds context representations from continuous speech representations and dependencies are obtained by the self-attention mechanism across the entire sequence of latent representations end-to-end. In (Conneau et al., 2021), multilingual pre-training of Wav2Vec2.0 model on 53 languages with more than 56k hours of unlabeled speech data (*XLSR-53*) has shown to construct better speech representations for cross-lingual transfer. It is in this context that we consider fine-tuning this model on creole languages. In (Evain et al., 2021), several Wav2Vec2.0 models (*LeBenchmark*) specific to French language were pretrained. We propose to fine-tune these models on creole languages. Results are generated with a Connectionist Temporal Classification (CTC) beam search decoder (Graves et al., 2006). CTC is an algorithm that assign a probability for any Y given an X. In our case X represent the acoustic features generated by *LeBenchmark* and Y the items in the orthographic transcription. The combination of *LeBenchmark* and CTC allowed us to produce an orthographic transcription of every speech segment provided by the VAD algorithm.

3.4 Evaluation

A gold standard has been created by the second author using the transcription automatically generated. We computed a Character Error Rate (CER) and a Word Error Rate (WER) on a set of 549 utterances. WER and CER calculates the percentage of items (words or character) that are incorrectly recognised in relation to the total number of items in a reference transcript. We obtained a CER of 0.45 and a WER of 0.728. We present in figure 1 the distribution of the WER and CER per utterances. To improve the visibility of the figure, We removed 5 examples that were too high. Although the overall results may be deemed suboptimal, the boxplot analysis reveals that a considerable proportion of utterances exhibit a WER of less than 50%. This suggests that a significant number of the generated utterances remain usable for downstream applications.

While evaluating a speech recognition system, its usability is often only based on the WER and CER . The results obtained are not groundbreaking but our collaboration between NLP scientists and linguists could help us understand how the system created is useful, how it can be exploited and how it can be improved.

Code-mixing: An under-resourced language is generally in contact with a widely spoken language. In our case, because French is the official language of Guadeloupe island and because some of the linguists involved in the data collection were English speakers, Gwadloupéyen, French and English were intertwined in the recordings. Non-Gwadloupéyen segments were then transcribed with the Gwadloupéyen norms. It seems unlikely to automatically differentiate French and Gwadloupéyen segments due to their lexical similarity. However, recent language diarisation tool could help us to filter out English segments (e.g. Liu et al., 2021).

Voice Activity Detection: VAD was highly accurate and saved time by filtering out non-speech segments. A few inaccuracies have however been mentioned specifically for segments starting with non-voiced consonants. the algorithm also tended to over-segment some segments that belonged together.

Automatic transcription: The quality of the transcriptions generated was not uniform across the recording (cf. Figure 1). While some transcriptions were not exploitable at all, others happen to be helpful support for transcription. On one hand, some of the utterance had a WER closed to 0 which allowed us to just copy paste the generated transcription to the gold standard with minor corrections. On the other, for utterances with more errors, the transcription could help to more clearly identify what is said.

Transcription errors: Besides the errors due to code mixing, most of the errors of the systems were due to oversegmentation of tokens. However, this type of errors could be mitigated by plugging a language model at the end of the CTC system. Another error noticed was the difficulty of the system to correctly identify the nasals which are usually recognised as orals (cf. Table 1).

4 Conclusion

We have detailed the first stage of a joint effort between field linguists and NLP experts to aid in transcribing Kréyòl Gwadloupéyen field linguistic data. Our approach involved using a voice ac-

comments	gold standard	automatic generation
the final nasal is recognised as two orals	zot matinike gwadloupeyen	zoln patinike gwadloup ee
the sentence was French	deux saison	deu sezon
segmentation error	zo kay an grante	jo kay angrandte
segmentation errors and nasal confusion	matinik e gwadloupeyen	martini ke gwadelou pe ent
segmentation error	se limajiner a sa	se limaj jener a sa
segmentation and transcription errors	byen pale de bonda nou kay soukre bonda	mye fame de gonda nou ka ai soucebo

Table 1: Examples of transcriptions

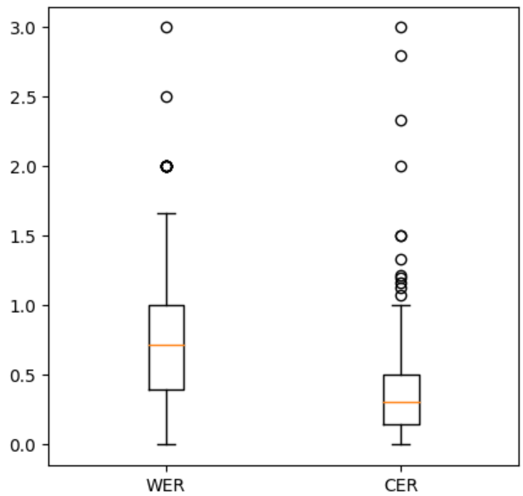

Figure 1: WER and CER distributions

tivity detection system combined with a wav2vec and CTC-based speech recognition model to transcribe raw recordings. The automatically generated transcription was then utilised to establish a gold standard.

Our initial work has prompted us to consider possibilities beyond conventional metrics such as WER and CER and to explore how even a transcription with a high error rate can still be useful. These early results have led us to question the relevance of standard metrics for evaluating a transcription system that can output words out of vocabulary. While a naive approach would be to assume that an automatically generated transcription is simply a starting point for post-editing and corrections (Bird, 2020, p.2), we have found that it can offer support for creating a gold standard and help transcribers better identify the content of a recording, especially when they are not confident in the target language. Moreover, the errors made by the system have increased our understanding of the requirements for a speech recognition system, potentially leading to improved recording quality in the future.

Moving forward, we will look to improve the output of the system. This will involve utilising an overlapping speech detector to eliminate noisy utterances, employing a language model to prevent token hyper-segmentation, and gradually improving the quality of the training data to enhance the transcription.

References

Oliver Adams, Benjamin Galliot, Guillaume Wisniewski, Nicholas Lambourne, Ben Foley, Rahasya Sanders-Dwyer, Janet Wiles, Alexis Michaud, Séverine Guillaume, Laurent Besacier, et al. 2021. User-friendly automatic transcription of low-resource languages: Plugging ESPnet into ELPIS. In *ComputEL-4: Fourth Workshop on the Use of Computational Methods in the Study of Endangered Languages*.

Alexei Baevski, Michael Auli, and Abdelrahman Mohamed. 2019. Effectiveness of self-supervised pre-training for speech recognition. *arXiv preprint arXiv:1911.03912*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Mat Bettinson and Steven Bird. 2017. Developing a suite of mobile applications for collaborative language documentation. In *2nd Workshop on Computational Methods for Endangered Languages*, pages 156–164.

Steven Bird. 2020. Sparse transcription. *Computational Linguistics*, 46(4):713–744.

Steven Bird, Florian Hanke, Oliver Adams, and Haejoong Lee. 2014. Aikuma: A mobile app for collaborative language documentation. In *Proceedings of the Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 1–5. ACL.

Paul Boersma and David Weenink. 1996. Praat, a system for doing phonetics by computer, version 3.4. *Institute of Phonetic sciences of the University of Amsterdam, Report*, 132:182.

Robert Chaudenson. 2004. La créolisation: théorie, applications, implications. *La créolisation*, pages 1–480.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. *Proceedings of Interspeech 2021*.

Solène Evain, Ha Nguyen, Hang Le, Marcely Zanon Boito, Salima Mdhaffar, Sina Alisamir, Ziyi Tong, Natalia Tomashenko, Marco Dinarelli, Titouan Parcollet, Alexandre Allauzen, Yannick Estève, Benjamin Lecouteux, François Portet, Solange Rossato, Fabien Ringeval, Didier Schwab, and Laurent Besacier. 2021. LeBenchmark: A Reproducible Framework for Assessing Self-Supervised Representation Learning from Speech. In *Proc. Interspeech 2021*, pages 1439–1443.

Ben Foley, Joshua T Arnold, Rolando Coto-Solano, Gautier Durantin, T Mark Ellison, Daan van Esch, Scott Heath, Frantisek Kratochvil, Zara Maxwell-Smith, David Nash, et al. 2018. Building speech recognition systems for language documentation: The CoEDL Endangered Language Pipeline and Inference System (ELPIS). In *Proceedings of The 6th International Workshop on Spoken Language Technologies for Under-Resourced Languages*, pages 205–209.

Elodie Gauthier. 2018. *Collecter Transcrire Analyser: quand la machine assiste le linguiste dans son travail de terrain*. Ph.D. thesis, Université Grenoble Alpes.

Elodie Gauthier, David Blachon, Laurent Besacier, Guy-Noël Kouarata, Martine Adda-Decker, Annie Rialland, Gilles Adda, and Grégoire Bachman. 2016. LIG-Aikuma: A mobile app to collect parallel speech for under-resourced language studies. *Interspeech 2016*, pages 381–382.

Luke Gessler. 2022. Closing the NLP gap documentary linguistics and NLP need a shared software infrastructure. In *Proceedings of the 5th Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 119–126.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Vishwa Gupta and Gilles Boulianne. 2020a. Automatic transcription challenges for Inuktitut, a low-resource polysynthetic language. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 2521–27.

Vishwa Gupta and Gilles Boulianne. 2020b. Speech transcription challenges for resource constrained Indigenous language Cree. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-Resourced Languages and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 362–367.

Nikolaus P. Himmelmann. 1998. *Documentary and descriptive linguistics*, volume 36. de Gruyter.

Béatrice Jeannot-Fourcaud and Paulette Durizot Jno-Baptiste. 2008. L'enseignement du français en contexte diglossique Guadeloupéen: état des lieux et propositions. *Former les enseignants du XXIème siècle dans toute la francophonie*, pages 61–73.

Kazuya Kawakami, Luyu Wang, Chris Dyer, Phil Blunsom, and Aaron van den Oord. 2020. Learning robust and multilingual speech representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1182–1192, Online. Association for Computational Linguistics.

William Lane, Mat Bettinson, and Steven Bird. 2021. A computational model for interactive transcription. In *Proceedings of the 2nd Workshop on Data Science with Human in the Loop: Language Advances*, pages 105–111.

Éric Le Ferrand, Steven Bird, and Laurent Besacier. 2022. Fashioning local designs from generic speech technologies in an Australian Aboriginal community. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4274–4285.

Hexin Liu, Leibny Paola García Perera, Xinyi Zhang, Justin Dauwels, Andy W.H. Khong, Sanjeev Khudanpur, and Suzy J. Styles. 2021. End-to-End Language Diarization for Bilingual Code-Switching Speech. In *Proc. Interspeech 2021*, pages 1489–1493.

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.

Sarah Ruth Moeller. 2014. SayMore, a tool for language documentation productivity. *Language Documentation and Conservation*, 8:66–74.

Salikoko S Mufwene. 1997. Jargons, pidgins, creoles, and koines: What are they? *CREOLE LANGUAGE LIBRARY*, 19:35–70.

Lambert-Félix Prudent. 1999. Des baragouins à la langue antillaise. *Des Baragouins à la langue Antillaise*, pages 1–214.

Jiatong Shi, Jonathan D Amith, Rey Castillo García, Esteban Guadalupe Sierra, Kevin Duh, and Shinji Watanabe. 2021. Leveraging end-to-end asr for endangered language documentation: An empirical study on Yolóxochitl Mixtec. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1134–1145.

Viveka Velupillai. 2015. Pidgins, creoles and mixed languages. *Pidgins, Creoles and Mixed Languages*, pages 1–626.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *5th International Conference on Language Resources and Evaluation*, pages 1556–15.