Technologies de la parole et données de terrain : le cas du créole haïtien

William N. Havard, Renauld Govain, Daphne Gonçalves Teixeira, Benjamin Lecouteux, Emmanuel Schang

Technologies de la parole et données de terrain : le cas du créole haïtien

William N. Havard[1,2], Renauld Govain[3], Daphne Gonçalves Teixeira[1], 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

RÉSUMÉ

Nous utilisons des données de terrain en créole haïtien, récoltées il y a 40 ans sur cassettes puis numérisées, pour entraîner un modèle natif d'apprentissage auto-supervisé (SSL) de la parole (WAV2VEC2) en haïtien. Nous utilisons une approche de pré-entraînement continu (CPT) sur des modèles SSL pré-entraînés de deux langues étrangères : la langue lexificatrice – le français – et une langue non apparentée – l'anglais. Nous comparons les performances de ces trois modèles SSL, et de deux autres modèles SSL étrangers directement affinés, sur une tâche de reconnaissance de la parole. Nos résultats montrent que le modèle le plus performant est celui qui a été entraîné en utilisant une approche CPT sur la langue lexificatrice, suivi par le modèle natif. Nous concluons que l'approche de "mobilisation des archives" préconisée par (Bird, 2020) est une voie prometteuse pour concevoir des technologies vocales pour de nouvelles langues.

ABSTRACT

Speech Technologies with Fieldwork Recordings : the case of Haitian Creole

We use fieldwork recordings in Haitian Creole, collected 40 years ago on cassettes and then digitised, to train a native self-supervised learning (SSL) model of speech (WAV2VEC2) in Haitian. We use a continuous pre-training (CPT) approach on pre-trained SSL models of two foreign languages : the lexifier language – French – and an unrelated language – English. We compare the performance of these three SSL models, and of two other directly finetuned foreign SSL models, on a speech recognition task. Our results show that the best-performing model is the one trained using a CPT approach on the lexifier language, followed by the native model. We conclude that the "mobilise the archive" approach advocated by (Bird, 2020) is a promising avenue for designing speech technologies for new languages.

MOTS-CLÉS : créole haïtien, enregistrement de terrain, modèles auto-supervisés, reconnaissance de la parole.

KEYWORDS: Haitian Creole, fieldwork recordings, self-supervised model, speech recognition.

1 Introduction

La plupart des langues peu dotées ne le sont souvent que du point de vue des informaticiens [1] : elles disposent souvent de nombreuses ressources collectées au fil des ans par des linguistes, des missionnaires religieux et, plus généralement, par la communauté des locuteurs elle-même (Bird, 2020). Les données ne sont souvent pas facilement accessibles (p. ex. sous un format numérique), mais elles existent néanmoins. La question à laquelle nous tentons de répondre dans cet article est la suivante : jusqu'où pouvons-nous aller avec les modèles de traitement de la parole état-de-l'art en utilisant *uniquement* des données de terrain *déjà existantes* ?

Par "données de terrain", nous entendons des données qui n'ont pas été collectées à l'origine pour servir de données d'entraînement pour des applications informatiques (p. ex. la reconnaissance automatique de la parole, RAP), mais qui ont été collectées à des fins linguistiques (p. ex. l'étude des variations dialectales). Dans cet article, nous nous concentrons sur des données orales en créole haïtien (*kreyòl ayisyen*), constituées d'entretiens enregistrés entre des linguistes et leurs collaborateurs. Le créole haïtien est un créole à base lexicale française (le français est sa langue lexificatrice, c'est à dire, la langue lui a apporté la plupart de son vocabulaire, voir Hazael-Massieux 2012), parlé par 13M de locuteurs (Simons & Fennig, 2023) à Haïti et par la diaspora haïtienne, principalement aux États-Unis d'Amérique.

La majorité des données que nous utilisons dans cet article (voir la section 2) a été collectée il y a 40 ans avec des magnétophones pour étudier les variations dialectales en haïtien, en mettant l'accent sur les variations lexicales. Contrairement aux livres audio couramment utilisés pour entraîner les modèles neuronaux (p. ex. Librispeech, Panayotov *et al.* 2015) qui jouissent d'une haute qualité d'enregistrement, les données que nous utilisons sont particulièrement bruitées : réverbération, echo, bruits ambiants (p. ex. poules, coqs, poussins, voitures, passants, etc.). Pourtant, ce type de données représente la majorité des données disponibles pour la plupart des langues du monde. La collecte et la transcription des données étant un processus coûteux, [2] ne pourrions-nous pas utiliser — comme le préconise (Bird, 2020) dans l'approche consistant à "mobiliser les archives" (*mobilise the archive*) — des données de terrain déjà existantes (et potentiellement anciennes) et les ré-utiliser pour des applications informatiques ?

Questions de recherche. Plus précisément, les questions que nous abordons dans cet article sont les suivantes : **(a)** Des données de terrain, bien que bruités (mais écologiques) seraient-elles utilisables pour entraîner des modèles d'apprentissage auto-supervisé (SSL) de la parole (p. ex. WAV2VEC2, Baevski *et al.* 2020) ? **(b)** Doit-on entraîner ces modèles à partir de zéro ou doit-on utiliser des approches de pré-entraînement continu (*continuous pre-training*, CPT, Nowakowski *et al.*, 2023; Gururangan *et al.*, 2020) ? **(c)** Quelle quantité de données d'entraînement est nécessaire pour affiner (*finetune*) les modèles sur une tâche de RAP ? Enfin, **(d)** est-il possible d'entraîner de tels modèles avec un budget limité ? (c'est-à-dire en utilisant un seul GPU et non 64 comme c'est le cas pour Baevski *et al.* 2020).

En outre, comme nous travaillons dans le contexte des langues créoles, nous visons également à explorer l'influence de la langue lexificatrice (comme un cas clair de langues apparentées) et explorons

1. Voir §§ 2 et 2.1 de (Bird, 2020) sur la notion de "*zero resource*" et la vision centrée "données" du traitement automatique des langues et de la parole.

2. Himmelmann (2018) rapporte que la transcription de 1 minute de parole peut prendre de 10 à 150 minutes, selon la langue, les connaissances du linguiste et le niveau de transcription (phonétique, phonologique, orthographique) et d'annotation annexe (morphologique, syntaxique, etc.).

(e) si l'approche CPT doit être effectuée sur des modèles SSL de la langue lexificatrice (p. ex. le français dans le cas du créole haïtien), ou si des modèles entraînés sur une langue non apparentée (p. ex. l'anglais dans le cas du créole haïtien) fonctionnent également ?

Travaux connexes. Le domaine du traitement de la parole pour les langues créoles par le biais de modèles neuronaux est relativement nouveau. Les seuls travaux de traitement de la parole pour ces langues sont ceux de (Breiter, 2014) pour le créole haïtien, ceux de (Macaire *et al.*, 2022) pour les créoles guadeloupéen et mauricien, et de (Gooda Sahib-Kaudeer *et al.*, 2019) pour le créole mauricien (avec un accent mis sur le domaine médical). Ainsi, le traitement de la parole pour les langues créoles — fussent-elles à base lexicale française, anglaise, portuguaises, etc. — reste largement inexploré.

Sans rapport direct avec le traitement de la parole pour les langues créoles — mais en rapport direct avec notre contexte méthodologique — Nowakowski *et al.* (2023) a exploré des approches de pré-entraînement continu, suivies d'une tâche d'affinage pour la reconnaissance vocale en ainu (langue native du nord du Japon) en utilisant d'anciennes données de terrain. Cependant, contrairement à l'objectif que nous nous fixons, ils n'entraînent pas leurs modèles avec un budget limité car (i) ils utilisent 4 GPU, (ii) utilisent le modèle XLSR-53 (Conneau *et al.*, 2021) qui est basé sur WAV2VEC2-LARGE et pré-entraîné sur 56k heures de données, et (iii) font un affinage multilingue par lequel le modèle de RAP n'est pas seulement entraîné sur la langue cible (ainu), mais sur plusieurs langues à la fois (anglais, japonais, en plus de l'ainu). Nous visons une approche plus stricte qui n'utilise que des données de terrain à toutes les étapes.

2 Données

ALH Nous avons utilisé le *Atlas Linguistique d'Haïti* (Fattier, 1998), constitué d'un ensemble de 499 enregistrements audio en créole haïtien collectés à Haïti entre 1978 et 1987 dans le but de créer un atlas linguistique. Les enregistrements ont été réalisés à l'origine sur des cassettes audio avec des magnétophones, puis numérisés dans les années 2010 par la Bibliothèque nationale de France. Chaque enregistrement dure en moyenne 45 minutes et consiste en un entretien dirigé entre un ou plusieurs enquêteurs qui demandent des mots ou des phrases à leurs collaborateurs Haïtiens. Ces enregistrements ont été numérisés et mis à disposition sur la plateforme COCOON (FLA and Fattier, 2015). [3] Bien que les enregistrements soient associés à des cahiers de terrain comportant des transcriptions manuscrites partielles (p. ex. transcription phonétique à l'échelle du mot), celles-ci n'ont pas été numérisées (ni alignées avec les enregistrements). Ainsi, ce corpus est entièrement constitué de parole brute.

Nous avons divisé l'ensemble de données (356,3 heures) en trois parties (train/val/test). Les données ont été réparties de manière à ce que l'ensemble de validation contienne au moins 5 heures de données et un minimum de 2 locuteurs inconnus, et l'ensemble de test au moins 5 heures de données et un minimum de 3 locuteurs inconnus. Nous avons obtenu la répartition suivante, qui répondait à nos contraintes : train = 345,6 heures ; val = 5,3 heures, 5 locuteurs inconnus ; et test = 5,4 heures, 8 locuteurs inconnus. [4]

CNCH Le *Corpus du créole haïtien du Nord* (*Corpus of Northern Haitian Creole*, Valdman *et al.*,

3. https://cocoon.huma-num.fr/exist/crdo/meta/cocoon-8ea988d2-bf16-303d-81a0-0c55cc0
4. L'ensemble de test n'a pas été utilisé dans les expériences présentées dans ce document, mais le sera dans les premiers travaux futurs énumérés dans la section 5.

Modèle	Langue SSL	WER ↓	**UER ↓**	Entraînement	Décodage	Classement
SSL-ETRANGER+CPT+FT	FR	**36.8**	21.6	320mn	4-gram	**1**
SSL-NATIF+Ø+FT	HAT	37.4	**21.5**	360mn (max)	3-gram	5
SSL-ETRANGER+CPT+FT	EN	37.5	22.4	320mn	4-gram	6
SSL-ETRANGER+Ø+FT	FR	42.5	24.5	360mn (max)	3-gram	27
SSL-ETRANGER+Ø+FT	EN	50.4	29.0	320mn	3-gram	49

Modèle	Langue SSL	WER ↓	**UER ↓**	Entraînement	Décodage	Classement
SSL-ETRANGER+CPT+FT	FR	**38.2**	**17.1**	320mn	Viterbi	**1**
SSL-NATIF+Ø+FT	HAT	39.8	17.8	360mn (max)	Viterbi	3
SSL-ETRANGER+CPT+FT	EN	40.3	18.6	360mn (max)	Viterbi	6
SSL-ETRANGER+Ø+FT	FR	46.2	21.7	360mn (max)	Viterbi	12
SSL-ETRANGER+Ø+FT	EN	57.1	26.6	360mn (max)	Viterbi	38

TABLE 1 – Architecture qui donnent les meilleures performances en termes de WER (en haut) et de UER (en bas) pour chaque type de modèle affiné. *Classement* montre le rang des modèles de 1 (meilleur) à 200 (pire) lorsque le WER/UER est utilisé comme clé de tri.

2015) [5] comprend 10 entretiens enregistrés, menés au Cap-Haïtien (Nord d'Haïti) pour étudier les variations dialectales par rapport au haïtien standard. Ce corpus a été entièrement transcrit par le linguiste l'ayant récolté. Cependant, nous tenons à mentionner que les transcriptions utilisées sont non-standards et impressionistes, dans le sens où des variations orthographiques déviant de la norme sont utilisées pour retranscrire plus fidèlement la prononciation du locuteur : "*Powoprens*"/"*Potoprens*", Port-au-Prince ; "*eskeu*"/"*eske*", est-ce que ; "*deu*"/"*de*", deux ; etc.). Ces variations pourront donc influer (de manière non favorable) sur le taux d'erreur mot (WER) et caractère (UER).

Nous avons divisé l'ensemble des données (9 heures) en trois parties (train/val/test). Les données ont été réparties de manière à ce que l'ensemble de validation contienne au moins 1 heure de données et un minimum d'un locuteur inconnu, et l'ensemble test au moins 1 heure de données et un minimum d'un locuteur inconnu. Nous avons obtenu la répartition suivante, qui répondait à nos contraintes : train = 6,9 heures ; val = 1,1 heure, 1 locuteur inconnu ; test = 1,0 heure, 2 locuteurs inconnus.

Autre ensemble de données Nous tenons à souligner l'existence d'autres ensembles de données présentant de la parole en créole haïtien, que nous avons volontairement exclus car ils ne consistent pas en des données de terrain : l'ensemble de données Haïti-CMU librement accessible [6] qui contient de la parole lue ($\sim$ 20 heures), principalement des sections de la Bible, qui ne reflètent pas l'utilisation quotidienne de la langue ; et l'ensemble de données propriétaire Babel-IARPA comprenant 203 heures et étant uniquement consistué de "parole conversationelle téléphonique scénarisée" (Andrus *et al.*, 2017).

3 Expériences

Compte tenu de notre contrainte de budget limité, nous nous concentrons uniquement sur l'architecture WAV2VEC2-BASE, excluant ainsi l'entraînement d'un modèle basé sur WAV2VEC2-LARGE, ainsi

5. https://archive.org/details/interview-8-ujf-107-a-ujm-107-a
6. http://www.speech.cs.cmu.edu/haitian/

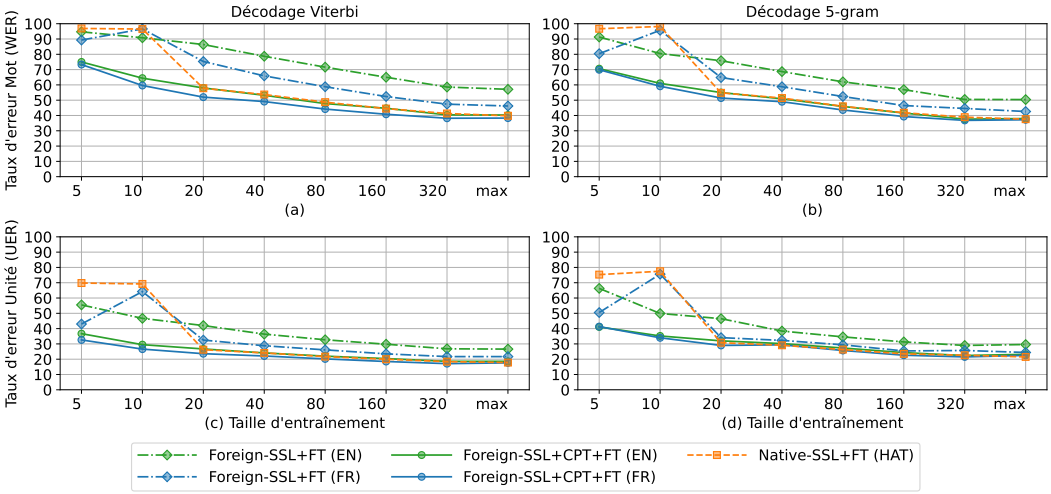

FIGURE 1 – (a, b) Taux d'erreur mot (WER) et (c, d) taux d'erreur unité (UER, au niveau des caractères) des modèles affinés sur une tâche de RAP avec décodage Viterbi (à gauche) et avec LM à 5-gram (à droite) en fonction de la quantité de données CNCH utilisées pour l'entraînement (en minutes, de 5 minutes à *max*, où $max = 6,9$ heures, ~ 360 minutes).

que l'affinage d'un modèle multilingue tel que XLSR-53 qui est basé sur l'architecture WAV2VEC2-LARGE.

Pré-entraînement SSL *natif* et *étranger*. Nous utilisons le corpus ALH pour entraîner nos modèles SSL. Un modèle de détection de l'activité vocale (Pyannote, Bredin *et al.*, 2020) a été utilisé pour isoler les sections correspondant à de la parole des bruits environnants, ce qui a permis d'obtenir 229h de sections parlées. Les segments résultants, plutôt courts ($\sim$ 2.3s), ont été fusionnés jusqu'à ce que les segments concaténés atteignent 19s en moyenne ($\sim$ 2.4s $\pm$ 5.8s). Les modèles WAV2VEC2 ont été entraînés sur un seul GPU [7] en accumulant le gradient pour 16 passes. Les modèles ont été entraînés jusqu'à convergence, définie soit comme le point d'intersection des courbes d'entraînement et de validation, soit comme le point où celles-ci sont restées stables pendant 10 000 passes.

Nous avons entraîné trois modèles. Le premier, entraîné à partir de zéro, que nous appelons désormais SSL-NATIF+Ø, puisque celui-ci n'a vu que du haïtien et ne se base pas sur un modèle existant dans le cadre d'un approche de pré-entrainement continu (+Ø). Les deux autres modèles ont été entrainés à partir des modèles existants dans le cadre d'une approche de pré-entraînement continu : l'un basé sur un modèle français (WAV2VEC2-FR-7K-BASE, pré-entraîné sur 7k heures en français, Parcollet *et al.* 2023), et l'autre à partir d'un modèle anglais (WAV2VEC2-BASE, Baevski *et al.* 2020) pré-entraîné sur Librispeech 960 (Panayotov *et al.*, 2015). Ces modèles sont appelés SSL-ETRANGER+CPT EN ou SSL-ETRANGER+CPT FR puisqu'ils ont été pré-entraînés sur une langue étrangère auparavant (soit de l'anglais, soit du français) et ont bénéficié d'une approche de pré-entrainement continu (+CPT) pendant laquelle ils ont été entraînés à modéliser de la parole en haïtien.

Affinage sur un tâche de RAP Nous avons affiné (+FT) les modèles pré-entraînés sur le corpus CNCH. 3 modèles de RAP ont été affinés à partir de modèles ayant vu du haitïen au pré-entraînement :

7. 32Gb Nvidia Tesla V100 ou 45Gb Nvidia A40 selon la disponibilité.

le modèle SSL-NATIF+Ø (appelé SSL-NATIF+Ø+FT après affinage), et les deux modèles SSL-ETRANGER+CPT basés sur de l'anglais ou du français (appelés SSL-ETRANGER+CPT+FT EN ou FR). En plus de ceux-ci, afin de comprendre la pertinence (ou non) d'un pré-entraînement continu sur des données de terrain, nous avons également affiné directement les modèles SSL-ETRANGER sans utiliser une approche CPT : SSL-ETRANGER+Ø+FT (EN ou FR). Ces modèles nous permettront ainsi de voir si le pré-entraînement sur des données de terrain permet de mieux transférer sur d'autres données de terrain dans une tâche de RAP ou non.

Afin de comprendre l'impact de la taille de l'entraînement sur les performances finales des modèles, nous utilisons différentes tailles d'entraînement : max (6.9 heures), 320, 160, 80, 40, 20, 10, et 5 minutes. Chaque taille d'entraînement inclut les tailles précédentes (par exemple, max $\supset$ ... $\supset$ 10 $\supset$ 5). Chaque modèle est affiné pour 20k passes. [8] Pour éviter le sur-entraînement, les paramètres ont été gelés pendant les 10k premières passes. Le meilleur modèle est sélectionné sur la base du WER le plus bas sur l'ensemble de validation. Le texte a été mis en minuscules et les diacritiques ont été supprimés (en raison d'une utilisation variable). Nous entraînons également des modèles de langue (LM) de 2 à 5 grammes sur les transcriptions pour chaque taille d'entraînement à l'aide de KenLM (Heafield, 2011), ce qui donne 32 LM différents (4 taille de n–gram $\times$ 8 taille de corpus d'affinage).

4 Résultats & Discussion

Nous avons utilisé l'outil SCTK [9] pour calculer le taux d'erreur mot (WER) et le taux d'erreur d'unité (UER, au niveau du caractère). Nous avons utilisé un décodage Viterbi standard ainsi qu'un réordonnancement a posteriori (*rescoring*) avec des LM de 2 à 5 grammes. Cela a permis d'obtenir 5 modèles $\times$ 8 tailles d'entraînement $\times$ (1 Viterbi + 4 ngram) décodages = 200 stratégies de décodage. Pour plus de clarté, seuls les rescores Viterbi et LM 5-grammes sont présentés dans la Fig. 1, et la meilleure configuration pour chacun des 5 types de modèles est présentée dans le Tab. 1.

Nos résultats montrent **(d)** qu'il est possible d'entraîner des modèles compétitifs avec un budget limité en utilisant un seul GPU et que **(a)** l'utilisation de données de terrain pour entraîner des modèles SSL de la parole est efficace. Bien que ces données soient intrinsèquement bruitées — par opposition aux livres audio ou aux discours radiodiffusés couramment utilisés pour entraîner les modèles SSL — le modèle haïtien SSL-NATIF+Ø que nous avons entraîné est resté très compétitif par rapport à d'autres approches. Ceci est particulièrement intéressant dans le cas des langues à faibles ressources, telles que la plupart des créoles à base française parlés dans les Caraïbes (haïtien, guadeloupéen, saint-lucien, etc.) ou en Amérique du Sud (guyanais). Cela signifie qu'il n'est pas nécessaire de collecter de nouvelles données, mais que les anciennes données enregistrées sur bande magnétique, une fois numérisées, peuvent être réutilisées à cette fin. Cela permettrait à de nombreuses langues du monde de disposer de modèles de traitement de la parole à la pointe de la technologie.

Quant à savoir **(b)** si nous devrions affiner les modèles SSL qui ont été pré-entraînés à partir de zéro ou les modèles pré-entraînés en utilisant une approche CPT, nos résultats montrent que les modèles entraînés dans une approche CPT montrent un léger avantage sur les modèles natifs entraînés à partir de zéro (-1.6 WER, et -0.7 UER, décodage de Viterbi, en utilisant l'UER le plus bas comme clef de tri). Cependant, nos résultats montrent que **(e)** cet avantage n'est vrai que lorsque le modèle utilisé

8. Compte tenu du peu de données dont nous disposons, les modèles convergent rapidement, restent stables et n'évoluent pas après 20k étapes, d'où cette valeur.

9. https://github.com/usnistgov/SCTK

pour le pre-entraînement continue est *celui de la langue lexificatrice* (ici, le français). Cet avantage semble disparaître lorsque ce n'est pas le cas, car le modèle affiné à partir d'une autre langue (ici, l'anglais) a généralement de moins bonnes performances qu'un modèle affiné à partir de la langue lexificatrice ($+2.1$ WER, $+1.5$ UER, *id.*) ou à partir de la langue cible ($+0.5$ WER, $+0.8$ UER, *id.*). Cependant, l'élément déterminant est l'utilisation de l'approche de pré-entraînement continue. Les modèles RAP directement affinés à partir des modèles SSL-ETRANGER+Ø+FT qui n'ont pas vu d'haïtien dans une approche CPT sont loin derrière ($+8$ WER, $+4.6$ UER pour les modèles basés sur le français, *id.*) ou très loin derrière ($+18.9$ WER, $+9.5$ UER pour les modèles basés sur l'anglais, *id.*) du meilleur modèle.

En ce qui concerne **(c)** la quantité de données nécessaires pour affiner les modèles SSL sur une tâche RAP, nos résultats montrent une différence marquée entre trois groupes de modèles : (i) SSL-ETRANGER+CPT+FT très robuste à une quantité réduite de données d'entraînement, (ii) SSL-ETRANGER+Ø+FT peu robuste à une quantité réduite de données, et (iii) SSL-NATIF+Ø montrant des résultats intermédiaires. L'utilisation de 20 minutes de données comble l'écart entre (i) et (iii) alors que les modèles du groupe (ii) ont nécessité environ 4 fois cette quantité de données (80 minutes) pour atteindre des performances similaires. Nous supposons que les modèles du groupe (i) bénéficient du fait d'avoir vu plus de parole, car ils ont été pré-entraînés dans leur langue respective (français ou anglais), ont vu des données haïtiennes dans la phase CPT, et ont été affinés, ce qui pourrait expliquer pourquoi ils sont plus robustes que les autres modèles. Enfin, nous avons observé des résultats mitigés avec l'utilisation des LM pour le décodage. Alors qu'ils n'améliorent pas de manière significative (ni ne nuisent) aux modèles SSL-NATIF+Ø+FT ou SSL-ETRANGER+CPT+FT, ils améliorent de manière significative les scores WER du SSL-ETRANGER+Ø+FT (Fig. 1a et 1b) : par exemple, -10 WER avec un LM 5-gram pour un modèle EN WAV2VEC2 affiné avec 40 minutes de données. Par conséquent, lorsqu'aucune donnée audio pour faire du pré-entraînement continu n'est disponible, l'utilisation d'un LM est indispensable. Cependant, il semble que l'utilisation des LM, tout en améliorant les scores WER, se fait au détriment de UER plus élevés (Fig. 1c et 1d) ; ce qui indique que, bien qu'il y ait plus de mots transcrits avec précision, les autres sont moins bien transcrits, ce qui se traduit par des UER plus élevés.

5 Limitations and Travaux Futurs

Dans cet article, nous nous sommes concentrés sur l'exploration de la validité de l'utilisation des données de terrain pour pré-entraîner des modèles auto-supervisés. Nous avons affiné ces modèles sur un tâche de RAP (évaluation intrinsèque), mais nous avons laissé de côté l'étude des modèles et des représentations pré-entraînés eux-mêmes (évaluation intrinsèque). Dans nos travaux futurs, nous souhaitons utiliser une tâche ABX (Schatz *et al.*, 2013) pour comparer les représentations latentes et leur transfert au niveau des phonèmes. Cela nous aiderait à mieux comprendre les performances de nos modèles. Les données que nous utilisons pour le pré-entraînement continu ont été collectées il y a 40 ans, et la langue entre cette époque et aujourd'hui a changé (p. ex. mots tombés en désuétude, évolution de la phonologie, etc.). La question de la mesure de ce phénomène et de son impact reste donc ouverte. Enfin, nos résultats montrent que 350 heures d'enregistrements sur le terrain sont suffisantes pour pré-entraîner un modèle SSL natif et obtenir des résultats compétitifs lorsqu'ils sont affinés sur une tâche de RAP. Cependant, un tel trésor avec autant d'heures d'enregistrement n'existe pas pour toutes les langues : la question de la quantité minimale de données de terrain à utiliser reste ouverte.

6 Conclusion

Nous avons utilisé des données de terrain en haïtien, enregistrées sur bandes magnétiques il y a 40 ans, puis numérisées, pour entraîner un modèle SSL natif. Nous avons également utilisé une approche CPT sur des modèles SSL pré-entraînés de la langue lexificatrice (le français) et d'une langue non apparentée (l'anglais), que nous avons affinés sur un autre ensemble de données de terrain dans le cadre d'une tâche de RAP. Nous avons obtenu des résultats compétitifs et montré que le meilleur modèle est le modèle pré-entraîné de la langue lexificatrice avec CPT sur des enregistrements de terrain haïtiens, suivi par le modèle SSL natif. Par conséquent, lorsqu'aucun modèle de la langue lexificatrice n'est disponible, il est toujours utile d'entraîner un modèle natif à l'aide de données de terrain. Ceci est d'autant plus important qu'un modèle natif peut être une source de fierté pour la communauté des locuteurs, contrairement à un modèle dérivé de la langue lexificatrice, généralement celle de l'ancienne puissance colonisatrice. Par conséquent, l'approche consistant à mobiliser les données d'archive, comme préconisée par (Bird, 2020), est une voie prometteuse.

Références

ANDRUS T., BILLS A., CONNERS T., CRABB E. S., DUBINSKI E., FISCUS J. G., GILLIES B., HARPER M., HAZEN T. J., HEFRIGHT B., JARRETT A., LE H., RAY J., RYTTING A., SHEN W., SILBER R. & TZOUKERMANN E. (2017). Iarpa babel haitian creole language pack iarpa-babel201b-v0.2b. DOI : 10.35111/ENHB-6110.

BAEVSKI A., ZHOU H., MOHAMED A. & AULI M. (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.

BIRD S. (2020). Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, p. 3504–3519, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.313.

BREDIN H., YIN R., CORIA J. M., GELLY G., KORSHUNOV P., LAVECHIN M., FUSTES D., TITEUX H., BOUAZIZ W. & GILL M.-P. (2020). pyannote.audio : neural building blocks for speaker diarization. In *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain.

BREITER W. (2014). Rapid bootstrapping of haitian creole large vocabulary continuous speech recognition.

CONNEAU A., BAEVSKI A., COLLOBERT R., MOHAMED A. & AULI M. (2021). Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In *Proc. Interspeech 2021*, p. 2426–2430. DOI : 10.21437/Interspeech.2021-329.

FATTIER D. (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.

FLA, FACULTÉ DE LINGUISTIQUE APPLIQUÉE DE L'UNIVERSITÉ D'ETAT D'HAÏTI (ANCIENNEMENT CENTRE DE LINGUISTIQUE APPLIQUÉE (CLA)) & FATTIER, DOMINIQUE (2015). Atlas linguistique d'Haïti. DOI : 10.34847/COCOON.8EA988D2-BF16-303D-81A0-0C55CC035240.

GOODA SAHIB-KAUDEER N., GOBIN-RAHIMBUX B., BAHSU B. S. & MAGHOO M. F. A. (2019). Automatic speech recognition for kreol morisien : A case study for the health domain. In A. A.

SALAH, A. KARPOV & R. POTAPOVA, Éds., *Speech and Computer*, p. 414–422, Cham : Springer International Publishing.

GURURANGAN S., MARASOVIĆ A., SWAYAMDIPTA S., LO K., BELTAGY I., DOWNEY D. & SMITH N. A. (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*, p. 8342–8360, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.740.

HAZAEL-MASSIEUX M.-C. (2012). *Les Créoles à base française*. Gap, France : Editions Ophrys.

HEAFIELD K. (2011). KenLM : Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, p. 187–197, Edinburgh, Scotland : Association for Computational Linguistics.

HIMMELMANN N. P. (2018). *Meeting the transcription challenge*. University of Hawai'i Press.

MACAIRE C., SCHWAB D., LECOUTEUX B. & SCHANG E. (2022). Automatic speech recognition and query by example for creole languages documentation. In *Findings of the Association for Computational Linguistics : ACL 2022*, p. 2512–2520, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.findings-acl.197.

NOWAKOWSKI K., PTASZYNSKI M., MURASAKI K. & NIEUWAŻNY J. (2023). Adaptation of a multilingual speech representation model for a new, underresourced language via multilingual fine-tuning and continued pretraining. *Science Talks*, **8**, 100249. DOI : https ://doi.org/10.1016/j.sctalk.2023.100249.

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)*, p. 5206–5210. DOI : 10.1109/ICASSP.2015.7178964.

PARCOLLET T., NGUYEN H., EVAIN S., BOITO M. Z., PUPIER A., MDHAFFAR S., LE H., ALISAMIR S., TOMASHENKO N., DINARELLI M., ZHANG S., ALLAUZEN A., COAVOUX M., ESTEVE Y., ROUVIER M., GOULIAN J., LECOUTEUX B., PORTET F., ROSSATO S., RINGEVAL F., SCHWAB D. & BESACIER L. (2023). Lebenchmark 2.0 : a standardized, replicable and enhanced framework for self-supervised representations of french speech.

SCHATZ T., PEDDINTI V., BACH F., JANSEN A., HERMANSKY H. & DUPOUX E. (2013). Evaluating speech features with the minimal-pair ABX task : analysis of the classical MFC/PLP pipeline. In *Proc. Interspeech 2013*, p. 1781–1785. DOI : 10.21437/Interspeech.2013-441.

SIMONS G. F. & FENNIG C. D., Éds. (2023). *Ethnologue : Languages of the world*. Summer Institute of Linguistics, Academic Pub.

VALDMAN A., VILLENEUVE A.-J. & SIEGEL J. F. (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. DOI : 10.1075/jpcl.30.1.01val.