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

Représentation de la parole multilingue par apprentissage auto-supervisé dans un contexte subsaharien

Antoine Caubrière[1] Elodie Gauthier[2]
(1) Orange Innovation, 2 Av. de Belle Fontaine, 35510 Cesson-Sévigné
(2) Orange Innovation, 2 Av. Pierre Marzin, 22300 Lannion
antoine.caubriere@orange.com elodie.gauthier@orange.com

RÉSUMÉ

Les approches auto-supervisées ont conduit à des avancées majeures dans le domaine de l'apprentissage profond. Par l'exploitation d'une grande quantité de données non annotées, ces approches ont notamment permis des améliorations dans des contextes peu dotés. Toutefois, les langues africaines restent majoritairement sous-représentées dans les jeux de données de préentraînement publiquement distribués. Dans ces travaux, nous préentraînons des modèles de parole auto-supervisés multilingues à partir de langues subsahariennes exclusivement. Nous étudions la pertinence des représentations apprises sur la tâche de reconnaissance de parole, en utilisant le jeu d'évaluation FLEURS-102. Notre modèle HuBERT$_{base}$ obtient des résultats similaires face à l'approche multilingue w2v-bert de FLEURS, tout en étant plus efficient, avec 6 fois moins de paramètres et 7 fois moins de données. Nous présentons aussi un second modèle exploitant une sous-sélection équilibrée des données initiales, obtenant des performances compétitives avec près de 80 fois moins de données de préentraînement.

ABSTRACT

Multilingual speech representation by self-supervised learning for sub-Saharan languages.

Self-supervised approaches are now unmissable and represent a major advance in deep learning. While self-supervised approaches have shown strong gains in a low-resource setting by leveraging the large amount of unlabeled data available on the web, languages spoken in sub-Saharan Africa (SSA) are still underrepresented in the datasets used in publicly available pre-trained models. In this paper, we build a multilingual pre-trained SSL model that uses only speech data in local languages spoken in SSA. We conducted experiments for downstream speech recognition task on the SSA subset of the FLEURS-102 dataset. Experiments conducted on speech recognition shown that our model, based on the HuBERT$_{base}$ architecture, obtains competitive results on the FLEURS dataset compared to the multilingual pre-trained w2v-bert-51 model, while being more efficient by using 7x less data and 6x less parameters. We trained another model with 80x less data, by using an equilibrated data selection.

MOTS-CLÉS : Apprentissage auto-supervisé, Langues subsahariennes, Reconnaissance de la parole multilingue, HuBERT.

KEYWORDS: Self-supervised representation, African languages, Multilingual ASR, HuBERT.

Actes des 35èmes Journées d'Études sur la Parole, pages 163–172.

1 Introduction

Récemment, les approches auto-supervisées ont montré leur potentiel pour la mise en place de systèmes de reconnaissance de la parole performants (Chung *et al.*, 2021; Conneau *et al.*, 2021; Pratap *et al.*, 2023). Ce type d'approche permet l'exploitation d'une grande quantité de données non transcrites pour l'apprentissage d'une représentation dense de la parole. Elles sont plus riches que certaines caractéristiques classiques comme les MFCC ou les bancs de filtres. Un modèle pré-entraîné de façon auto-supervisé peut ensuite être utilisé soit comme un encodeur de parole, dont les paramètres seront adaptés, soit comme un extracteur de caractéristiques qui sera figé. Indépendamment de l'utilisation finale d'un modèle pré-entraîné, les données exploitées pour son apprentissage impacteront ses performances sur les tâches finales (Zhao & Zhang, 2022).

Les travaux de (Pires *et al.*, 2019) ont montré que le transfert d'apprentissage d'une langue bien dotée vers une langue sous-dotée est plus efficace lorsque les langues partagent des caractéristiques typologiques similaires. Toutefois, (Joshi *et al.*, 2020) fait état que 48% des caractéristiques typologiques répertoriées par le projet de classification WALS [1] (*World Atlas of Language Structures*) n'apparaissent pas dans les jeux de données. En complément, la plupart des modèles multilingues accessibles publiquement sont entraînés sur quelques langues seulement, ce qui entraîne leur sur-représentation au détriment d'autres langues (Valk & Alumäe, 2021; Babu *et al.*, 2022; Conneau *et al.*, 2023; Zhang *et al.*, 2023). Encore sous-dotées de nos jours, les langues africaines, de par leurs richesses et la présence de caractéristiques uniques, sont fortement impactées par cette situation (Clements & Rialland, 2007; Yadav & Sitaram, 2022).

Ces dernières années, l'intérêt des langues africaines est grandissant au sein de la communauté du traitement des langues. En surpassant des modèles multilingues pré-appris en majorité sur des données en anglais, plusieurs études ont montré l'intérêt d'un modèle pré-appris principalement sur les langues africaines (Ogueji *et al.*, 2021; Adelani *et al.*, 2022; Dossou *et al.*, 2022; Adebara *et al.*, 2022). En traitement de la parole, de nouveaux challenges et ressources sont publiés (Gutkin *et al.*, 2020; Sikasote & Anastasopoulos, 2021; Boito *et al.*, 2022; Olatunji *et al.*, 2023; Wanjawa *et al.*, 2023). Dans le cadre de la tâche de reconnaissance de la parole, les travaux de (Ritchie *et al.*, 2022) ont permis de meilleures performances en exploitant une approche multilingue auto-supervisée par rapport à une approche plus classique.

En phase avec tous ces travaux, nous proposons dans ce papier un modèle de représentation de la parole multilingue centré sur les langues africaines. Plus particulièrement, en utilisant des données exclusivement issues de la région subsaharienne, nous pré-entraînons un modèle fait pour être adapté à des tâches de traitement de la parole pour ces langues. Nous pré-entraînons également un second modèle pour lequel nous équilibrons les données du jeu d'apprentissage en fonction des langues et du genre des locuteurs. L'objectif est de produire un modèle non orienté vers un groupe de langues ou de locuteurs en raison d'une sur-représentation. Dans le cadre de nos expérimentations, nous nous attachons à résoudre la tâche de reconnaissance de la parole pour évaluer la pertinence des représentations fournies par nos modèles auto-supervisés.

Dans ce papier, nous présentons tout d'abord le jeu de données brutes – non transcrit –, que nous avons construit, avant d'apporter des détails sur l'architecture employée. Nous poursuivons ensuite par la description de nos expériences, ainsi que des résultats obtenus. Nous terminons en comparant nos systèmes avec le modèle de référence proposé par (Conneau *et al.*, 2023), avant de conclure.

1. https ://wals.info/feature

2 Jeux de données

2.1 Données non transcrites

Nous avons collecté, sur le web, des données issues de plusieurs sources émises dans des pays d'Afrique subsaharienne. Ces données correspondent à des journaux d'informations diffusés en ligne, qui traitent de divers sujets comme la politique, l'environnement, la santé, l'éducation, l'actualité régionale. Ces sujets sont abordés sous forme d'interviews, de débats et d'émissions. Ces journaux d'information s'engagent à favoriser la liberté d'expression, la mixité et le dialogue entre les cultures et à contribuer à l'égalité entre femmes et hommes. Dans ces travaux, néanmoins, ces données ne sont utilisées qu'à des fins de pré-apprentissage, afin de construire une représentation acoustique multilingue pertinente pour le traitement des langues parlées dans la zone. La teneur lexicale et sémantique des enregistrements audio n'est ainsi jamais exploitée. Les enregistrements récoltés permettent de couvrir un ensemble de 21 langues et variantes.

Parmi ces données, nous pouvons trouver des enregistrements en environnement contrôlé (studio), des interviews bruitées (extérieur), ainsi que des éléments non relatifs à la parole comme de la musique. Notre jeu de données mélange ainsi de la parole préparée, de la parole spontanée, ainsi que des segments de parole bruitée. Afin d'isoler les segments de parole, nous avons appliqué une détection d'activité vocale à l'aide de l'outil *pyannote* (Bredin, 2023). Nous effectuons aussi des pré-traitements permettant l'uniformisation des enregistrements (format, fréquence d'échantillonnage, ...).

L'ensemble de nos pré-traitements nous permet de construire un jeu de données de plus de 59 500h de parole exploitable pour un apprentissage auto-supervisé.

Nous donnons la répartition brute par langue dans la figure 1. Les langues sont identifiées par leur code de langue issu de la norme ISO 639-3 et les variantes sont considérées comme des langues distinctes. La catégorie "Unknown" correspond à des segments provenant d'enregistrements mixant plusieurs des 21 langues considérées, avec la présence d'alternance codique sur une partie des segments. Nous n'avons pas appliqué d'algorithme d'identification de langue. La langue française "FRA" correspond à des données accentuées produites par des locuteurs africains.

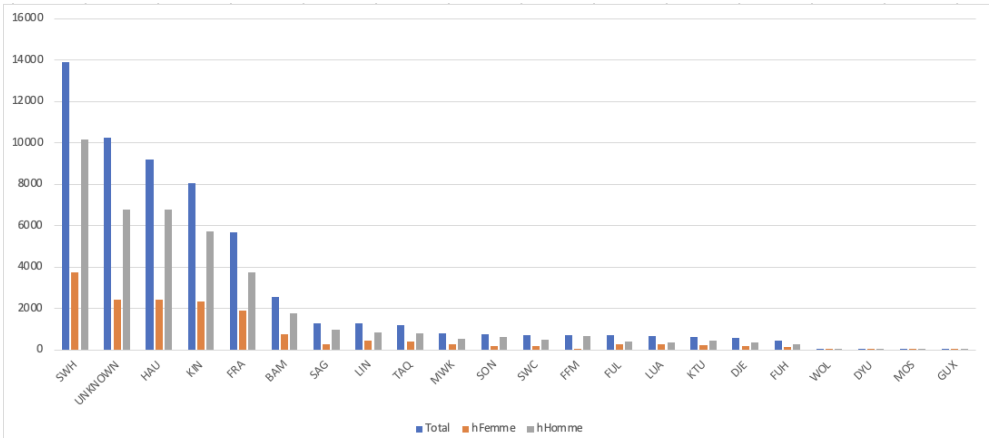

FIGURE 1 – Répartition brute du jeu d'apprentissage en heures par langues / variantes.

En complément des données brutes, nous produisons une sous-sélection de notre jeu de données la

plus équilibrée possible. Cette répartition équilibrée vise à produire un modèle auto-supervisé le moins orienté possible vers un groupement de langues sur-représentées dans les données d'apprentissage. Pour chaque langue, nous regroupons, dans la mesure du possible, 400 heures de segments de parole équilibrées selon les critères de langues et de genre des locuteurs. Nous appliquons un algorithme de détection du genre basé sur l'apprentissage fin d'un modèle XLSR-53 (Conneau *et al.*, 2021), sur librispeech (Panayotov *et al.*, 2015), pour produire l'annotation en genre des segments de notre corpus non supervisé [2]. Nous considérons les variantes d'une même langue comme appartenant à la langue et nous accumulons en quantité équivalente par variantes.

Cet équilibrage nous permet de construire un second jeu de données totalisant 5660 heures de parole. Nous donnons la répartition du jeu de données équilibré par langues dans la figure 2.

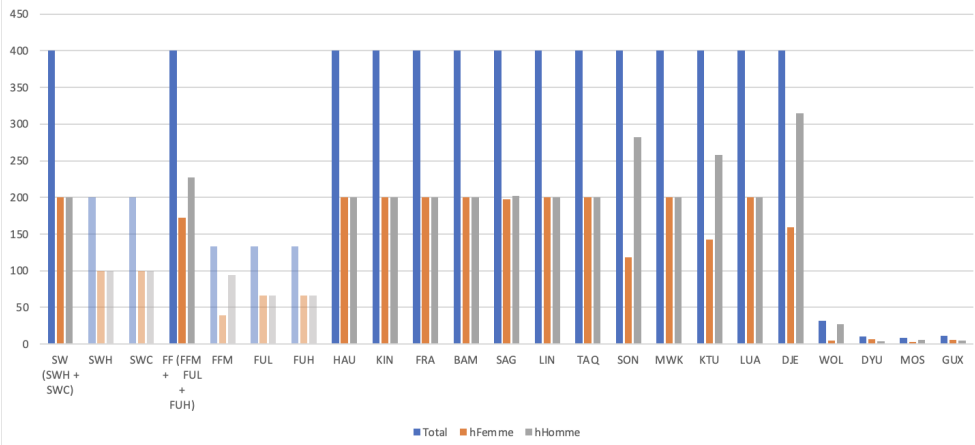

FIGURE 2 – Répartition équilibrée du jeu d'apprentissage en heures par langues.

2.2 Données transcrites

Dans le cadre de cette étude, nous exploitons le jeu d'évaluation FLEURS (Conneau *et al.*, 2023). Il s'agit d'un corpus au sein duquel 102 langues sont représentées et pour lesquelles environ 12h de parole transcrite est fournie (par langue). Nous nous concentrons plus particulièrement sur le sous-ensemble de données relatif à la région subsaharienne (FLEURS$_{SSA}$). Ces données sont composées de segments de parole couvrant 20 langues différentes, dont 5 sont communes aux données non transcrites. Dans ce papier, nous exploitons les transcriptions normalisées fournies par le corpus.

3 Système

3.1 Pré-apprentissage

Le système mis en place dans cette étude correspond à l'architecture "*base*" de l'approche HuBERT (Hsu *et al.*, 2021). Cette approche est auto-supervisée par l'annotation automatique en étiquette cible.

2. https ://huggingface.co/alefiury/wav2vec2-large-xlsr-53-gender-recognition-librispeech

Cette annotation est produite pour l'ensemble des segments de parole du corpus non transcrit, à l'aide de l'algorithme des K-moyennes.

L'architecture "*base*" correspond à un encodeur convolutif complété par 12 couches de type *transformers* produisant des plongements de taille 768. Elles possèdent 8 têtes d'attention et des linéaires internes de 3 072 unités. Cet ensemble de paramètres conduit à un modèle HuBERT d'environ 95 millions de paramètres.

L'apprentissage d'un système de ce type s'effectue en plusieurs étapes :

1. Apprendre un modèle non supervisé à partir des K-moyennes sur 100 classes, en exploitant les descripteurs acoustiques MFCC.
2. Créer des étiquettes auto-supervisées avec les K-moyennes sur l'ensemble d'entraînement et de validation.
3. Apprendre de zéro le modèle *transformer* avec les étiquettes auto-supervisées.
4. Extraire les représentations neuronales intermédiaires des *transformers* ("*base*" = couche 6).
5. Reproduire les étapes 1. à 3. avec ces représentations neuronales et 500 classes.

3.2 Apprentissage fin

Lors de l'apprentissage fin, nous considérons le modèle pré-entraîné comme un encodeur de la parole qui ne sera pas gelé. Nous exploitons ses représentations neuronales, de taille 768, issues de la dernière couche *transformer*. Nous ajoutons par-dessus deux couches linéaires de 1 024 unités, ainsi qu'une couche de sortie softmax dont la taille dépend des langues traitées par le système final.

4 Expériences

Nous effectuons l'ensemble du préentraînement de notre système à l'aide de l'outil Fairseq (Ott *et al.*, 2019). Nous avons parallélisé les calculs sur 4 GPU A40. Le pré-entraînement brut (sur environ 60k heures de données) a duré pendant un peu plus de 35 jours (soit plus de 3 360 heures de calcul GPU) tandis que le préentraînement sur données équilibrées (sur environ 5k heures de données) a duré un plus plus de 12 jours (soit plus de 1 150 heures de calcul GPU). Dans les deux cas, et notamment en raison de limite de mémoire, nous exploitons 600 heures du jeu d'entraînement pour effectuer l'apprentissage du modèle des K-moyennes basé sur les représentations internes des *transformers*. Nous sélectionnons ces 600 heures aléatoirement en conservant la contrainte de proportionnalité des langues et des genres représentés. Nous nommons les modèles pré-entraînés en fonction de la quantité d'heures non-supervisée utilisée. Nous apprenons ainsi le modèle 60k sur les quelques 59 500 heures récoltées et le modèle 5k sur la sous-sélection équilibrée de ces quelques 5 600 heures de parole. Les deux modèles pré-entraînés sont disponibles publiquement sur la plate-forme Hugging Face [3].

L'apprentissage fin n'est pas dépendant du modèle pré-entraîné utilisé. Pour chacune des deux expérimentations, nous utilisons l'outil SpeechBrain (Ravanelli *et al.*, 2021) afin d'entraîner les système de reconnaissance de la parole pour chacune des langues du corpus FLEURS$_{SSA}$. En moyenne, et en fonction de la langue, le temps d'affinement de l'entraînement prend 10 heures sur un seul GPU RTX 3090. L'ensemble des apprentissages supervisés par langues sont regroupés sous les systèmes *60k* et *5k*.

3. Les URLs seront données lors de la version finale de l'article.

En complément, nous réalisons un apprentissage conjoint sur l'ensemble des 20 langues, sur un seul GPU RTX 3090 pendant 48 heures. Ce premier apprentissage supervisé sert de base commune pour un transfert d'apprentissage vers un apprentissage fin sur chacune des langues du corpus FLEURS$_{SSA}$. Les systèmes supervisés obtenus à la suite de cet apprentissage conjoint sont regroupés sous les systèmes *60k*$_{joint}$ et *5k*$_{joint}$.

4.1 Résultats

Nous compilons dans la table 1, les résultats de nos expérimentations sur l'ensemble de test de FLEURS$_{SSA}$, exprimés en taux d'erreur sur les caractères (CER) et sur les mots (WER). Nous regroupons les langues entre celles vues lors du pré-apprentissage et celles non vues. Nous ajoutons aussi, pour chacun des groupes, la moyenne des scores, ainsi que la moyenne globale (toutes langues confondues). Afin de se comparer le plus justement possible, nous appliquons une méthodologie similaire aux travaux de (Conneau *et al.*, 2023). Ainsi, nous n'exploitons pas de modèles de langue pour réorganiser les hypothèses émises par le système.

	CER				WER			
	60k	5k	60k$_{joint}$	5k$_{joint}$	60k	5k	60k$_{joint}$	5k$_{joint}$
Langues vues								
Peul	21,2	21,2	17,8	**17,7**	61,9	60,6	56,4	**55,4**
Haoussa	10,5	11,2	**9,0**	10,1	32,5	35,6	**29,4**	33,8
Lingala	8,7	8,7	**6,9**	7,4	24,7	24,2	**20,9**	21,4
Swahili	7,1	8,6	**5,5**	6,5	23,8	28,8	**20,3**	24,4
Wolof	19,4	19,2	**17,0**	17,3	55,0	54,2	50,7	**50,1**
Moyenne	13,4	13,8	**11,2**	11,8	39,6	40,7	**35,5**	37,0
Langues non vues								
Afrikaans	23,3	23,8	20,3	**19,9**	68,4	68,3	62,6	**61,1**
Amharique	15,9	15,5	14,9	**14,3**	52,7	51,4	49,0	**47,6**
Luganda	11,5	11,7	**10,7**	11,1	52,8	53,3	**50,3**	52,0
Igbo	19,7	20,9	**17,2**	**17,2**	57,5	57,9	52,9	**52,4**
Kamba	16,1	16,3	**15,6**	15,9	53,9	**53,7**	**53,7**	54,3
Luo	9,9	10,2	**8,2**	8,4	38,9	38,5	34,9	**34,3**
Sotho du Nord	13,5	14,4	**11,7**	11,5	43,2	44,6	38,9	**38,6**
Chewa (Nyanja)	13,3	13,7	**10,9**	11,3	54,2	54,5	48,3	**48,1**
Oromo	22,8	22,9	**20,1**	21,2	78,1	77,4	74,8	**74,3**
Shona	11,6	11,2	**8,3**	8,7	50,2	48,2	**39,3**	39,7
Somali	21,6	21,9	**19,7**	20,0	64,9	64,5	**60,3**	60,6
Umbundu	21,7	21,7	**18,8**	20,7	61,7	60,8	**54,2**	57,0
Xhosa	11,9	12,4	**9,9**	10,1	51,6	52,3	**45,9**	47,1
Yoruba	24,3	25,0	**23,5**	23,8	67,5	68,0	**65,7**	66,6
Zoulou	12,2	12,4	**9,6**	10,0	53,4	53,0	**44,9**	46,1
Moyenne	16,6	16,9	**14,6**	14,9	56,6	56,4	**51,7**	52,0
Moyenne globale	15,8	16,1	**13,8**	14,1	52,3	52,5	**47,7**	48,2

TABLE 1 – Résultats obtenus sur les 20 langues subsahariennes de l'ensemble de test issu de FLEURS.

Les résultats présentés dans le tableau 1 montrent l'apport bénéfique de l'utilisation conjointe de données transcrites pour alimenter les modèles multilingues. Pour l'ensemble des langues, les performances sont bien meilleures après application du transfert d'apprentissage depuis les données conjointes. Pour les modèles 60k, nous notons respectivement une amélioration relative de 12,6% et de 8,8% sur les taux de CER et de WER moyens. Dans le cas des modèles 5k, il s'agit d'une amélioration relative de 12,4% (CER) et de 8,1% (WER).

Un second résultat intéressant de ces expérimentations concerne la comparaison entre les performances du modèle 60k$_{joint}$ et du modèle 5k$_{joint}$. 55 000 heures de données supplémentaires et plus de 22 jours de préentraînement GPU ont été consommés afin que le modèle 60k$_{joint}$ converge, pour finalement n'observer qu'un gain relatif de 1,8% (CER) et 1,0% (WER).Ce constat incite à s'interroger sur la quantité de données utiles ainsi que sur leur répartition lors du processus d'apprentissage, dans le cadre d'une approche multilingue. Ceci est d'autant plus intéressant au regard des questionnements sur la frugalité des approches, en termes de coût énergétique notamment.

Par comparaison des deux modèles "$_{joint}$", en ce qui concerne les langues vues lors du pré-apprentissage, nous remarquons une dégradation des performances sur les langues les plus représentées dans l'approche 60k (houaussa, lingala, swahili). En contrepartie, nous observons de meilleures performances sur les langues qui étaient sous-représentées dans l'approche 60k (peul et wolof). Ces résultats suggèrent que l'apprentissage du modèle 5k conduit à des représentations de la parole mieux réparties entre les langues.

4.2 Comparaison à FLEURS

En complément de nos résultats, nous confrontons ici nos systèmes aux résultats publiés par (Conneau *et al.*, 2023). Dans la mesure où les auteurs ne fournissent pas le détail des scores [4], nous effectuons la comparaison uniquement en termes de score moyen sur les caractères (CER), sur les 20 langues du sous-ensemble subsaharien proposé dans FLEURS$_{SSA}$. L'approche considérée par (Conneau *et al.*, 2023) est un w2v-bert de 600M de paramètres, pré-appris sur plus de 400 000 heures de parole.

	5k	60k	5k$_{joint}$	60k$_{joint}$	FLEURS$_{w2v-bert}$
Moyenne	16.1	15.8	14.1	13.8	13.6

TABLE 2 – Scores moyens sur l'ensemble test des données FLEURS$_{SSA}$.

Les résultats obtenus avec le système 60k$_{joint}$ montrent des performances très proches du modèle w2v-bert de FLEURS, tout en exploitant 7 fois moins de données et 6 fois moins de paramètres. Notre modèle 5k$_{joint}$, bien qu'obtenant des résultats inférieurs de 3,5%, se montre particulièrement compétitif eu égard au volume des données de préentrainement (80 fois moins de données que w2v-bert). L'utilisation de données exclusivement en langues locales, représentatives des parlers en Afrique subsaharienne, permet de réaliser un pré-apprentissage bien plus efficient. En ciblant des aspects particuliers – ici les particularités inhérentes au contexte africain –, ces résultats tendent à démontrer la pertinence d'une modélisation à l'échelle, face à l'apprentissage de modèles surdimensionnés.

4. Les résultats diffèrent entre la version déposée sur arXiv et la version IEEE du papier. Nous considérons ici la version officiellement validée et publiée par IEEE.

5 Conclusion

Dans ce papier, nous pré-entraînons le premier modèle multilingue auto-supervisé, librement partagé, appris exclusivement sur 21 langues et variantes africaines. Ces langues présentent des caractéristiques riches et non observées dans les autres langues du monde, notamment les langues occidentales. Nous confrontons notre modèle au jeu d'évaluation FLEURS qui propose un ensemble composé de langues parlées en Afrique subsaharienne. Face à leur modèle de référence, nous obtenons des résultats très similaires, tout en proposant une approche bien plus modérée en termes de coût d'apprentissage (près de 7 fois moins de données et 6 fois moins de paramètres). Dans ce même objectif de frugalité et d'efficience, nous proposons un second modèle (5k) entraîné à partir d'un sous-ensemble équilibré, en langue et en genre, de nos données. Le dimensionnement de ce modèle a permis de réduire de 2 200 heures le calcul GPU nécessaire, tout en contenant la perte de performance (dégradation moyenne de 0,3 points de CER face à notre modèle 60k, et 0,3 points face au modèle w2v-bert de FLEURS). Ce modèle se montre ainsi être un bon compromis entre matière d'efficience. Lors de travaux futurs, une analyse approfondie de ces systèmes sera menée, notamment au travers de l'étude de l'impact de l'équilibrage du jeu d'apprentissage sur la qualité des représentations dans un contexte multilingue, ainsi que de l'étude de l'impact du genre des locuteurs sur les performances des modèles de reconnaissance de la parole.

Références

ADEBARA I., ELMADANY A., ABDUL-MAGEED M. & INCIARTE A. A. (2022). Serengeti : Massively multilingual language models for Africa. *arXiv preprint arXiv :2212.10785*.

ADELANI D., NEUBIG G., RUDER S., RIJHWANI S., BEUKMAN M., PALEN-MICHEL C., LIGNOS C., ALABI J., MUHAMMAD S., NABENDE P., DIONE C. M. B., BUKULA A., MABUYA R., DOSSOU B. F. P., SIBANDA B., BUZAABA H., MUKIIBI J., KALIPE G., MBAYE D., TAYLOR A., KABORE F., EMEZUE C. C., AREMU A., OGAYO P., GITAU C., MUNKOH-BUABENG E., MEMDJOKAM KOAGNE V., TAPO A. A., MACUCWA T., MARIVATE V., ELVIS M. T., GWADABE T., ADEWUMI T., AHIA O., NAKATUMBA-NABENDE J., MOKONO N. L., EZEANI I., CHUKWUNEKE C., OLUWASEUN ADEYEMI M., HACHEME G. Q., ABDULMUMIN I., OGUNDEPO O., YOUSUF O., MOTEU T. & KLAKOW D. (2022). MasakhaNER 2.0 : Africa-centric Transfer Learning for Named Entity Recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 4488–4508, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics. DOI : 10.18653/v1/2022.emnlp-main.298.

BABU A., WANG C., TJANDRA A., LAKHOTIA K., XU Q., GOYAL N., SINGH K., VON PLATEN P., SARAF Y., PINO J., BAEVSKI A., CONNEAU A. & AULI M. (2022). XLS-R : Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Proc. Interspeech 2022*, p. 2278–2282. DOI : 10.21437/Interspeech.2022-143.

BENAMARA F., HATOUT N., MULLER P. & OZDOWSKA S., Éds. (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

BOITO M. Z., BOUGARES F., BARBIER F., GAHBICHE S., BARRAULT L., ROUVIER M. & ESTÈVE Y. (2022). Speech resources in the tamasheq language. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 2066–2071.

BREDIN H. (2023). pyannote.audio 2.1 speaker diarization pipeline : principle, benchmark, and recipe. In *Proc. Interspeech 2023*.

CHUNG Y.-A., ZHANG Y., HAN W., CHIU C.-C., QIN J., PANG R. & WU Y. (2021). w2v-BERT : Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, p. 244–250. DOI : 10.1109/ASRU51503.2021.9688253.

CLEMENTS G. N. & RIALLAND A. (2007). *Africa as a phonological area*, p. 36–85.

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.

CONNEAU A., MA M., KHANUJA S., ZHANG Y., AXELROD V., DALMIA S., RIESA J., RIVERA C. & BAPNA A. (2023). Fleurs : Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, p. 798–805 : IEEE.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

DOSSOU B. F., TONJA A. L., YOUSUF O., OSEI S., OPPONG A., SHODE I., AWOYOMI O. O. & EMEZUE C. (2022). Afrolm : A self-active learning-based multilingual pretrained language model for 23 African languages. In *Proceedings of The Third Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, p. 52–64.

GUTKIN A., DEMIRSAHIN I., KJARTANSSON O., RIVERA C. E. & TÚBÒSÚN K. (2020). Developing an open-source corpus of yoruba speech. In *Proc. of Interspeech 2020*, p. 404–408, October 25–29, Shanghai, China, 2020.

HSU W.-N., BOLTE B., TSAI Y.-H. H., LAKHOTIA K., SALAKHUTDINOV R. & MOHAMED A. (2021). HuBERT : Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **29**, 3451–3460. DOI : 10.1109/TASLP.2021.3122291.

JOSHI P., SANTY S., BUDHIRAJA A., BALI K. & CHOUDHURY M. (2020). The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv :2004.09095*.

LAIGNELET M. & RIOULT F. (2009). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In A. NAZARENKO & T. POIBEAU, Éds., *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

LANGLAIS P. & PATRY A. (2007). Enrichissement d'un lexique bilingue par analogie. In (Benamara *et al.*, 2007), p. 101–110.

OGUEJI K., ZHU Y. & LIN J. (2021). Small data ? no problem ! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, p. 116–126.

OLATUNJI T., AFONJA T., YADAVALLI A., EMEZUE C. C., SINGH S., DOSSOU B. F. P., OSUCHUKWU J., OSEI S., TONJA A. L., ETORI N. & MBATAKU C. (2023). AfriSpeech-200 : Pan-African Accented Speech Dataset for Clinical and General Domain ASR. *Transactions of the Association for Computational Linguistics*, **11**, 1669–1685. DOI : 10.1162/tacl_a_00627.

OTT M., EDUNOV S., BAEVSKI A., FAN A., GROSS S., NG N., GRANGIER D. & AULI M. (2019). fairseq : A fast, extensible toolkit for sequence modeling. *CoRR*, **abs/1904.01038**.

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.

PIRES T., SCHLINGER E. & GARRETTE D. (2019). How multilingual is multilingual BERT ? In A. KORHONEN, D. TRAUM & L. MÀRQUEZ, Éds., *Proceedings of the 57th Annual Meeting*

of the Association for Computational Linguistics, p. 4996–5001, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1493.

PRATAP V., TJANDRA A., SHI B., TOMASELLO P., BABU A., KUNDU S., ELKAHKY A., NI Z., VYAS A., FAZEL-ZARANDI M., BAEVSKI A., ADI Y., ZHANG X., HSU W.-N., CONNEAU A. & AULI M. (2023). Scaling speech technology to 1,000+ languages.

RAVANELLI M., PARCOLLET T., PLANTINGA P., ROUHE A., CORNELL S., LUGOSCH L., SUBAKAN C., DAWALATABAD N., HEBA A., ZHONG J., CHOU J.-C., YEH S.-L., FU S.-W., LIAO C.-F., RASTORGUEVA E., GRONDIN F., ARIS W., NA H., GAO Y., MORI R. D. & BENGIO Y. (2021). Speechbrain : A general-purpose speech toolkit.

RITCHIE S., CHENG Y.-C., CHEN M., MATHEWS R., VAN ESCH D., LI B. & SIM K. C., Éds. (2022). *Large vocabulary speech recognition for languages of Africa : multilingual modeling and self-supervised learning*.

SERETAN V. & WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. In (Benamara *et al.*, 2007), p. 401–410.

SIKASOTE C. & ANASTASOPOULOS A. (2021). Bembaspeech : A speech recognition corpus for the bemba language.

VALK J. & ALUMÄE T. (2021). Voxlingua107 : a dataset for spoken language recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, p. 652–658 : IEEE.

WANJAWA B. W., WANZARE L. D. A., INDEDE F., MCONYANGO O., MUCHEMI L. & OMBUI E. (2023). Kenswquad—a question answering dataset for swahili low-resource language. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, **22**(4). DOI : 10.1145/3578553.

YADAV H. & SITARAM S. (2022). A survey of multilingual models for automatic speech recognition.

ZHANG Y., HAN W., QIN J., WANG Y., BAPNA A., CHEN Z., CHEN N., LI B., AXELROD V., WANG G., MENG Z., HU K., ROSENBERG A., PRABHAVALKAR R., PARK D. S., HAGHANI P., RIESA J., PERNG G., SOLTAU H., STROHMAN T., RAMABHADRAN B., SAINATH T., MORENO P., CHIU C.-C., SCHALKWYK J., BEAUFAYS F. & WU Y. (2023). Google usm : Scaling automatic speech recognition beyond 100 languages.

ZHAO J. & ZHANG W.-Q. (2022). Improving automatic speech recognition performance for low-resource languages with self-supervised models. *IEEE Journal of Selected Topics in Signal Processing*, **16**(6), 1227–1241. DOI : 10.1109/JSTSP.2022.3184480.