AFRICA-CENTRIC SELF-SUPERVISED PRE-TRAINING FOR MULTILINGUAL SPEECH REPRESENTATION IN A SUB-SAHARAN CONTEXT

Antoine Caubrière & Elodie Gauthier, Orange Innovation, France
{antoine.caubriere,elodie.gauthier}@orange.com

ABSTRACT

We present the first self-supervised multilingual speech model trained exclusively on African speech. The model learned from nearly 60 000 hours of unlabeled speech segments in 21 languages and dialects spoken in sub-Saharan Africa. On the SSA subset of the FLEURS-102 dataset, our approach based on a HuBERT$_{base}$ (0.09B) architecture shows competitive results, for ASR downstream task, compared to the w2v-bert-51 (0.6B) pre-trained model proposed in the FLEURS benchmark, while being more efficient by using 7x less data and 6x less parameters. Furthermore, in the context of a LID downstream task, our approach outperforms FLEURS baselines accuracy by over 22%.

1 INTRODUCTION

Popular self-supervised learning (SSL) approaches have shown their potential to handle multilingual speech recognition (ASR) and are capable of achieving top performance (Conneau et al. (2021); Chung et al. (2021); Pratap et al. (2023)). They enable a model to be pre-trained on a vast amount of unlabeled data, producing richer audio representation for training downstream models, compared to standard features such as MFCCs or filterbanks. A pre-trained model can be used as a speech encoder with a fine-tuning or as a feature extractor by freezing its weights during the downstream task training. In any case, the performance of the downstream task models will be affected by the characteristics of the speech data used for pre-training (Zhao & Zhang, 2022).

Although Pires et al. (2019) already demonstrated, five years ago, that transfer learning from resource-rich to resource-poor languages is more effective when the languages share similar typological features and, later, Joshi et al. (2020), revealed that 48% of the typological features indexed in the World Atlas of Language Structures (WALS) classification project[1] do not appear in datasets, most of the multilingual pre-trained speech models publicly released today still are mainly learned from only very few languages, causing their over-representation at the cost of others (Valk & Alumäe (2021); Conneau et al. (2022); Babu et al. (2022); Zhang et al. (2023)). African languages, which have unique characteristics and are underresourced, are severely affected by this situation (Clements & Rialland (2007); Yadav & Sitaram (2022)).

Fortunately, African languages gain interest in the NLP community. Several studies have demonstrated the effectiveness of Africa-centric pre-trained models, showing superior performance compared to large multilingual pre-trained models that are primarily trained on English (Ogueji et al. (2021); Adelani et al. (2022); Dossou et al. (2022); Adebara et al. (2022)). In speech processing, several challenges and publications of new resources recently appeared (Sikasote & Anastasopoulos (2021); Boito et al. (2022); Olatunji et al. (2023); Wanjawa et al. (2023)). On the ASR downstream task, Ritchie et al. (2022) got better performance for several African languages when applying self-supervised techniques and multilingual modeling, compared to traditional approaches.

In line with these works, we tackle in this paper the under-representation of African languages by proposing a multilingual speech pre-trained model specifically made for performing downstream tasks in sub-Saharan Africa (SSA) languages, by only using spoken data from this region.

[1] https://wals.info/feature

2 DATASETS

Unlabeled The pre-trained dataset we created is composed of broadcast news recordings from diverse sources publicly available on the Web, across several countries, during May 2023. Sometimes, the same recording could be available in different languages spoken in the country. Data collected contained both studio recordings (controlled environment, prepared talks) and street interviews (noisy environment, spontaneous speech). Occasionally, jingles or songs appeared in the audio content. We therefore applied a voice activity detection (VAD) tool (Bredin, 2023) to get segments containing only speech. Finally, we gathered a dataset which comprises nearly 60 000 hours of speech segments and covers 21 languages and variants. For details, see appendix A.

Labeled Conneau et al. (2022) publicly released a parallel speech dataset in 102 languages and proposed it as benchmark. Data are divided in seven macro family, including a sub-Saharan Africa group. We therefore evaluate our approach on this SSA subset (FLEURS$_{SSA}$) which is composed of 20 languages, 5 of which are present in our pre-trained dataset.

3 EXPERIMENTS

Experiments were carried out using the well-known HuBERT approach (Hsu et al., 2021) with the base configuration (90M parameters). The pretraining task was achieved using the unlabeled data and the fairseq toolkit (Ott et al., 2019) through two successive iterations on 4 A100 40Gb GPUs. The first iteration was trained for 275k steps, using a K-means clustering computed on the MFCCs extracted from the training set as target labels. The second iteration was trained for 500k steps, and used embeddings from the 6th transformer layer using 600 hours of the training set. The ratio between languages has been preserved. The finally obtained pre-trained model is publicly available[2].

For downstream task training, we used the SpeechBrain toolkit (Ravanelli et al., 2021). The final pretrained model is considered as a speech encoder and is fully fine-tuned with two 1024 linear layers and a softmax output at the top. A first pool of speech recognition system (60k$_{(0.09B)}$) is obtained by a direct fine-tuning of the whole model on each language of the FLEURS dataset. A second pool (60k$_{FT-ALL(0.09B)}$) is then obtained by first jointly fine-tuning on all languages before fine-tuning again on each language.

Following the methodology of the FLEURS paper (Conneau et al., 2022) and to be consistent with their results, we did not rescore the hypothesis with a language model. Average character error rates (CERs) obtained on the 20 languages of the FLEURS$_{SSA}$ test set are given in table 1. The detailed scores per language are provided in appendix B.

	CER			WER	
	60k$_{(0.09B)}$	60k$_{FT-ALL(0.09B)}$	FLEURS$_{w2v-bert(0.6B)}$	60k	60k$_{FT-ALL}$
average	15.8	13.8	13.6	52.3	47.7

Table 1: Average results on SSA subpart of FLEURS-102 test set. *(detailed results in appendix)*

Results show that a model that is six times smaller and trained with seven times less data can achieve a performance level that is very close to the best baseline of FLEURS. This model is a step in the direction of more specific but cost-effective pre-trained approaches.

To ensure the quality of the speech representation, we fine-tuned our pretrained model using SpeechBrain for a language identification (LID) downstream task. We employed adaptive average pooling to produce output with shape *[Batch,1,20]* and we applied a softmax. We call this model 60K$_{LID}$. The model is trained for 15 epochs on the 20 languages of the FLEURS$_{SSA}$ subset.

We also propose a second scenario where we employed adaptive average pooling to produce output with shape *[Batch,1,768]*, with the addition of two linear layers to smoothly decrease the dimension from 768 to 256 then from 256 to 20. We call this model 60K$_{LID-smooth}$. It is trained under the same conditions as 60K$_{LID}$. Accuracy for both scenarios is presented in Table 2.

[2] https://huggingface.co/Orange/

	FLEURS$_{w2v-bert}$	FLEURS$_{mSLAM}$	60K$_{LID}$	60K$_{LID-smooth}$
FLEURS$_{SSA}$	59.1	62.2	84.9	90.4

Table 2: LID accuracy on SSA subset of FLEURS-102 test set.

Experiments have shown that our pre-trained model yields significantly improved results. This improvement can be attributed to the model's specialization in SSA languages. Specifically, we utilized only SSA speech data for pretraining and, during fine-tuning, the model was trained solely on the 20 SSA languages from the FLEURS dataset, rather than the full dataset of 102 languages.

The results obtained on both downstream tasks suggest that our models produce relevant multilingual speech representations within the specific context of SSA languages.

4 Conclusion

To the best of our knowledge, we present the first open source SSL model exclusively pre-trained on sub-Saharan African languages. By only focusing on African speech that contains specific features unobserved in other languages spoken in the world, we improved the robustness on the ASR downstream task for SSA languages. While we obtain similar results on the overall SSA subset than the best model presented in the FLEURS paper (w2v-BERT-51), yet our approach is more efficient by using much less data and a reduced number of parameters for pre-training. On a LID downstream task, results show that our specialized model trained on the SSA context performs better than the two FLEURS baselines, by obtaining more than 22% in absolute accuracy.

References

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. Serengeti: Massively multilingual language models for africa. *arXiv preprint arXiv:2212.10785*, 2022.

David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo Lerato Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Oluwaseun Adeyemi, Gilles Quentin Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu, and Dietrich Klakow. MasakhaNER 2.0: Africa-centric Transfer Learning for Named Entity Recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 4488–4508, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.298. URL https://aclanthology.org/2022.emnlp-main.298.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Proc. Interspeech 2022*, pp. 2278–2282, 2022. doi: 10.21437/Interspeech.2022-143.

Marcely Zanon Boito, Fethi Bougares, Florentin Barbier, Souhir Gahbiche, Loïc Barrault, Mickaël Rouvier, and Yannick Estève. Speech resources in the tamasheq language. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 2066–2071, 2022.

Hervé Bredin. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. Interspeech 2023*, 2023.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 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)*, pp. 244–250, 2021. doi: 10.1109/ASRU51503.2021.9688253.

G. N. Clements and Annie Rialland. *Africa as a phonological area*, pp. 36–85. Cambridge Approaches to Language Contact. Cambridge University Press, 2007.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In *Proc. Interspeech 2021*, pp. 2426–2430, 2021. doi: 10.21437/Interspeech.2021-329.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pp. 798–805, 2022. doi: 10.1109/SLT54892.2023.10023141.

Bonaventure FP Dossou, Atnafu Lambebo Tonja, Oreen Yousuf, Salomey Osei, Abigail Oppong, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi, and Chris Emezue. 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)*, pp. 52–64, 2022.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. doi: 10.1109/TASLP.2021.3122291.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*, 2020.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 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*, pp. 116–126, 2021.

Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure F. P. Dossou, Joanne Osuchukwu, Salomey Osei, Atnafu Lambebo Tonja, Naome Etori, and Clinton Mbataku. AfriSpeech-200: Pan-African Accented Speech Dataset for Clinical and General Domain ASR. *Transactions of the Association for Computational Linguistics*, 11:1669–1685, 12 2023. ISSN 2307-387X. doi: 10.1162/tacl_a_00627. URL https://doi.org/10.1162/tacl_a_00627.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. *CoRR*, abs/1904.01038, 2019. URL http://arxiv.org/abs/1904.01038.

Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4996–5001, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1493. URL https://aclanthology.org/P19-1493.

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. Scaling speech technology to 1,000+ languages, 2023.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. Speechbrain: A general-purpose speech toolkit, 2021.

Sandy Ritchie, You-Chi Cheng, Mingqing Chen, Rajiv Mathews, Daan van Esch, Bo Li, and Khe Chai Sim (eds.). *Large vocabulary speech recognition for languages of Africa: multilingual modeling and self-supervised learning*, 2022. URL https://arxiv.org/abs/2208.03067.

Claytone Sikasote and Antonios Anastasopoulos. Bembaspeech: A speech recognition corpus for the bemba language, 2021.

Jörgen Valk and Tanel Alumäe. Voxlingua107: a dataset for spoken language recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 652–658. IEEE, 2021.

Barack W. Wanjawa, Lilian D. A. Wanzare, Florence Indede, Owen Mconyango, Lawrence Muchemi, and Edward Ombui. Kenswquad—a question answering dataset for swahili low-resource language. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(4), apr 2023. ISSN 2375-4699. doi: 10.1145/3578553. URL https://doi.org/10.1145/3578553.

Hemant Yadav and Sunayana Sitaram. A survey of multilingual models for automatic speech recognition, 2022.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, Zhong Meng, Ke Hu, Andrew Rosenberg, Rohit Prabhavalkar, Daniel S. Park, Parisa Haghani, Jason Riesa, Ginger Perng, Hagen Soltau, Trevor Strohman, Bhuvana Ramabhadran, Tara Sainath, Pedro Moreno, Chung-Cheng Chiu, Johan Schalkwyk, Françoise Beaufays, and Yonghui Wu. Google usm: Scaling automatic speech recognition beyond 100 languages, 2023.

Jing Zhao and Wei-Qiang Zhang. 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, 2022. doi: 10.1109/JSTSP.2022.3184480.

A PRE-TRAINED DATASET DETAILED

In the following table 3, we present the languages distribution in the pre-training set.
We applied automatic segmentation on the raw recordings.
For the French language set, only African accented French was used.
"Unknown" row at the end of the table means speech recordings with language mixing.
No automatic LID has been applied to the segments.

Language	**ISO-3**	**Hours**
Bambara	bam	2 552
Dyula	dyu	14
French	fra	5 670
Fula	ful	702
Fulfulde	ffm	727
Fulfulde	fuh	446
Gulmancema	gux	13
Hausa	hau	9 211
Kinyarwanda	kin	8 046
Kituba	ktu	647
Lingala	lin	1 269
Luba-Lulua	lua	675
Mossi	mos	13
Maninkakan	mwk	791
Sango	sag	1 268
Songhai	son	780
Swahili	swc	706
Swahili	swh	13 926
Tamasheq	taq	1 212
Wolof	wol	64
Zarma	dje	567
Unknown	—	10 272
Total	—	*59 572*

Table 3: Languages distribution in the pre-training set.

B DETAILED RESULTS ON SSA SUBPART OF FLEURS-102

Results listed below are obtained when applying monolingual fine-tuning on each sub-Saharan African languages provided in the Test set of FLEURS benchmark.
Scores in bold show the best result depending on the approach. We show character error rate (CER) scores along with word error rates (WERs).

	CER		WER*	
Language	60k$_{(0.09B)}$	60k$_{FT-ALL(0.09B)}$	60k	60k$_{FT-ALL}$
Seen languages				
Fula	21.2	**17.8**	61.9	**56.4**
Hausa	10.5	**9.0**	32.5	**29.4**
Lingala	8.7	**6.9**	24.7	**20.9**
Swahili	7.1	**5.5**	23.8	**20.3**
Wolof	19.4	**17.0**	55.0	**50.7**
average	13.4	**11.2**	39.6	**35.5**
Unseen languages				
Afrikaans	23.3	**20.3**	68.4	**62.6**
Amharic	15.9	**14.9**	52.7	**49.0**
Ganda	11.5	**10.7**	52.8	**50.3**
Igbo	19.7	**17.2**	57.5	**52.9**
Kamba	16.1	**15.6**	53.9	**53.7**
Luo	9.9	**8.2**	38.9	**34.9**
Northen-Sotho	13.5	**11.7**	43.2	**38.9**
Nyanja	13.3	**10.9**	54.2	**48.3**
Oromo	22.8	**20.1**	78.1	**74.8**
Shona	11.6	**8.3**	50.2	**39.3**
Somali	21.6	**19.7**	64.9	**60.3**
Umbundu	21.7	**18.8**	61.7	**54.2**
Xhosa	11.9	**9.9**	51.6	**45.9**
Yoruba	24.3	**23.5**	67.5	**65.7**
Zulu	12.2	**9.6**	53.4	**44.9**
average	16.6	**14.6**	56.6	**51.7**
overall average	15.8	**13.8**	52.3	**47.7**

Table 4: Results obtained on the Test set of the 20 languages from the SSA subpart of FLEURS-102.